Knowledge-Based Systems 229 (2021) 107316

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A multimodal hierarchical approach to speech emotion recognition from audio and text

Prabhav Singh[1], Ridam Srivastava[1], K.P.S. Rana *, Vineet Kumar

Instrumentation and Control Engineering Department, Netaji Subhas University of Technology, Sector-3, Dwarka, New Delhi 110078, India

ARTICLE INFO

Article history:
Received 27 December 2020
Received in revised form 15 July 2021
Accepted 16 July 2021
Available online 22 July 2021

Keywords:
Speech emotion recognition
Hierarchical approach
Multimodal
Deep learning
Lexical features

ABSTRACT

Speech emotion recognition (SER) plays a crucial role in improving the quality of man–machine interfaces in various fields like distance learning, medical science, virtual assistants, and automated customer services. A deep learning-based hierarchical approach is proposed for both unimodal and multimodal SER systems in this work. Of these, the audio-based unimodal system proposes using a combination of 33 features, which include prosody, spectral, and voice quality-based audio features. Further, for the multimodal system, both the above-mentioned audio features and additional textual features are used. Embeddings from Language Models v2 (ELMo v2) is implemented to extract word and character embeddings which helped to capture the context-dependent aspects of emotion in text. The proposed models' performances are evaluated on two audio-only unimodal datasets – SAVEE and RAVDESS, and one audio-text multimodal dataset – IEMOCAP. The proposed hierarchical models offered SER accuracies of 81.2%, 81.7%, and 74.5% on the RAVDESS, SAVEE, and IEMOCAP datasets, respectively. Further, these results are also benchmarked against recently reported techniques, and the reported performances are found to be superior. Therefore, based on the presented investigations, it is concluded that the application of a deep learning-based network in a hierarchical manner significantly improves SER over generic unimodal and multimodal systems.

1. Introduction

A large variety of animals convey their emotions through the articulation of sound and movement. However, this process is much more complex in humans, who tend to convey their emotions through the articulation of both sound and language. Emotion is described as a strong feeling derived from one's circumstances, mood, or relationship with others and forms an essential part of interpersonal communication. Further, human communication is an intangible part of everyday life, and its inherent complexity and required effort go unnoticed. It involves several collaborative processes, such as transforming ideas into words, extracting meaning from words, and relying on context and emotion to narrow down the most appropriate interpretation of an ongoing dialog.

A system that recognizes human emotion is a valuable asset in many fields such as automatic customer service, distance education [1], and personal AI-based assistants. Additionally, these systems have been claimed to assist the treatment of neurological disorders [2]. The applicability of speech emotion recognition (SER) in business intelligence is another contributing factor. It can boost customer relationship management, improve recommendation systems' capabilities, prevent troll filtering, and detect spam in online social interactions [3]. SER also finds its application in intelligent dialog systems and in helping people with physical disabilities. For instance, Ashok and John [4] developed a facial expression recognition (FER) system for visually impaired people which identifies emotions from facial expressions and conveys them to the user to facilitate effective social interaction. Further, SER systems can also be used to improve the safety and comfort of a driver by enhancing attentiveness [5]. Therefore, it becomes essential to research how emotion modulates and improves the verbal and nonverbal aspects of human communication to develop interfaces that are more in line with the need of users [6].

Due to the subjectivity of emotion, which humans interpret and display in various ways, SER remains challenging even after decades of research. This is proven by the fact that most

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: prabhavs.ic18@nsut.ac.in (P. Singh), ridams.ic18@nsut.ac.in (R. Srivastava), kpsrana@nsut.ac.in (K.P.S. Rana), vineet.kumar@nsut.ac.in (V. Kumar).

[1] Equal contribution of the authors.

https://doi.org/10.1016/j.knosys.2021.107316
0950-7051/

SER datasets have a human recognition accuracy of only 60%–70%. Other obstacles include the shortage of datasets with natural recordings. Most publicly available datasets are composed of acted (simulated) and elicited (induced) audio samples, which result in models having an excellent dataset-centric recognition rate but a very low real-life recognition accuracy [7]. Additionally, factors such as gender, number of speakers, and noise further increase the complexity. Finally, selection of the most suitable features for SER has also been a point of contention [7], with most of the existing research concentrated on the use of low-level descriptors (LLD's) like energy, zero-crossing rate, Mel-frequency cepstral coefficients (MFCC's), and their higher-order deltas. While LLDs are closely related to the raw signal and help extract instantaneous audio-based speech features, they do not provide any global information about the utterance [8].

In contrast, humans generally perceive emotion from speech expressed over a while. Therefore, to improve attention on a more global scope, high-level statistical functions (HSFs) are used on these features [8,9]. Recently, researchers have utilized automatic feature extraction methods such as 2D/1D convolution networks, long-short term memory cells (LSTMs) [10,11], and prosody-based audio features like pitch, tone, and duration of silence. Hence, to leverage the benefits of both local and global level features, this work incorporates LLDs as well as HSFs of spectral and prosody-based features for SER.

Another aspect of human communication is that humans do not rely on verbal indicators alone to display emotion. Facial expressions, verbal context, and body language contribute more to emotion than speech alone. For example, low-level audio-based features like energy and pitch might help differentiate between "Angry" and "Sad" emotions but may fail for "Sad" and "Fear." On the other hand, supporting modalities such as facial expression and textual context might be more helpful in determining the correct emotion, which makes it crucial to include multiple modalities in an SER system. Facial expressions [12], text transcription [13], and bio-signals like electrodermal activity and electrocardiogram [14] have been used as additional modalities leading to enhanced performance. Since humans can produce entirely new combinations of words with different emotional contexts, analyzing the text transcription of audio samples can further increase the recognition rate of emotions. Further, with the advent of sophisticated automatic speaker recognition (ASR) techniques, it has become much easier and faster to obtain textual transcriptions of audio samples. However, the application of textual features as a supporting modality is relatively less explored.

Drawing motivations from the above background, the proposed method uses embeddings generated from textual transcriptions as a supporting modality for SER. As shown in Fig. 1, the suggested framework focuses on utilizing embeddings from language models v2 (ELMo v2) [15] to generate a high-dimensional vector representation of words. These are used as a supporting modality along with audio and prosody-based features to develop a hierarchical deep learning-based SER system. The hierarchical structure of the proposed model is based on the binary decision tree approach introduced by Lee et al. [16] to develop a framework that uses deep learning-based binary classification units arranged hierarchically. There are two significant benefits of using a hierarchical approach to this task [16–19]: Firstly, as hypothesized by Lazarus [20] in his Appraisal Theory of Emotion, an individual's emotional response to a situation depends on how it unfolds. The emotional response is thus a multi-stage process rather than a single-stage one. Therefore, for systems to accurately recognize emotion, it became necessary to adopt an approach similar to this multi-stage process. Secondly, adopting a hierarchical approach makes it easier to account for imbalance in data by taking care of ambiguous and data-light classes towards the end of the tree while working on data-heavy classes at the top. This reduces error-propagation in the tree. This paper thus proposes an approach that filters out one or more emotion(s) at each level using deep neural networks. The performance of the proposed model is first evaluated on audio-only datasets and then scaled to multimodal systems. Thus, the significant contributions of this paper are:

(1) A deep learning-based hierarchical approach to SER is proposed and successfully tested on unimodal and multimodal datasets using audio and textual features.

(2) A combination of prosody, voice quality (VQ) based, and spectral features are used in both local and global level audio descriptors to capture the way humans perceive emotions.

(3) ELMo v2 is successfully introduced for extracting lexical features.

The rest of the paper is organized as follows. Section 2 provides a brief survey of the related work. Three significant aspects of the proposed SER systems, namely, feature selection, unimodal systems, and multimodal systems, are reviewed, leading to relevant conclusions. Section 3 presents the used datasets, highlighting their composition and specific features. Section 4 describes the developed methodology and architecture of the proposed model, while Section 5 presents the results and provides a comparative study with other relevant works. Finally, Section 6 concludes the work with possible future proposals.

2. Related works

This section is further organized into three subsections. In Section 2.1, different features and feature extraction methods for SER have been briefly reviewed. Following this, recent works related to audio-based frameworks have been reviewed in Section 2.2. Finally, Section 2.3 presents the works on multimodal and multitask emotion recognition systems. Following this, inferences and research gaps are drawn and presented.

2.1. Feature selection and feature extraction methods from speech

The frequently used features for emotion classification tasks can be classified into three major categories: Spectral, Prosody, and VQ Features [7]. Spectral features such as MFCCs are frequently applied for audio classification and speech recognition tasks [21,22]. MFCCs are based on the frequency domain and use the Mel-scale, consummate to the human-ear scale, making them much more suited to SER [23]. As claimed by Ghazale and Hansen, MFCCs outperformed other spectral linear features like linear predictive coding in studies involving emotion recognition [24]. Their work proposed two new feature extraction methods: a modified Mel-frequency scale, and an exponential–logarithmic scale, both of which were reported to outperform MFCCs in stressed speech recognition. Other recently used spectral features include Gamma tone frequency cepstral coefficients (GFCCs). In 2018, Liu claimed GFCCs to have a higher recognition rate than MFCCs in SER tasks [25]. More recently, Nagarajan et al. compared the use of human-factor cepstral coefficients (HFCCs) with GFCCs and MFCCs [26]. They reported superior performance of the former when used with a support vector machine (SVM) on the EMO-DB and SAVEE datasets.

In 2012, Lin et al. found prosody-based features to extract distinctive properties of emotion [27] efficiently. Further, they suggested that prosodic features such as energy can be a helpful determiner in analyzing high v/s low arousal emotions like anger and sadness. Furthermore, Busso et al. [6] demonstrated that the pitch contour/fundamental frequency is specifically important among all prosodic features since it is affected by emotional

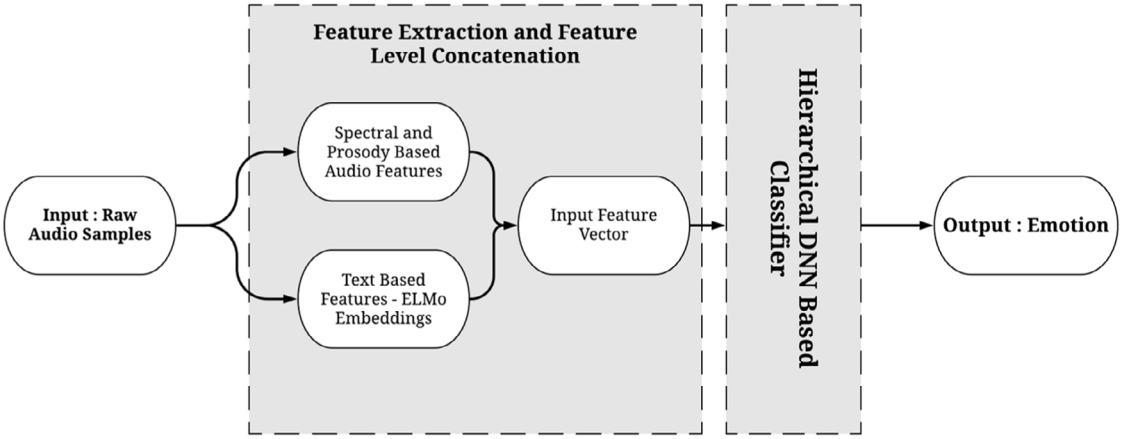

Fig. 1. The framework of the proposed SER system.

modulation directly. They further reported gross pitch contour statistics like mean and maximum to be more emotionally prominent than other features describing the pitch shape. Based on this, Schuller et al. compared gross statistics of energy and pitch to their instantaneous values using hidden Markov models (HMMs) and claimed an improvement of 7% over human-judge recognition accuracy using global features [28]. Exploring a different aspect, Rao et al. [29] studied the effect of feature extraction from words and syllables located in different parts of the sentence. Compared to other positions, they stated that words found in the final position of sentences and syllables in the final position of the words exhibited more emotionally discriminative details. Another important outcome of their study was the enhanced performance due to the combined local and global prosodic features.

Jitter, shimmer, harmonics to noise ratio (HNR) and other VQ features have also been successfully leveraged for SER. In their study on emotion recognition in human–computer interaction, Cowie et al. [30] claimed a strong correlation between emotion and voice quality. In support of this, VQ-based features were claimed to be more suited for recognizing emotions such as sadness, anxiety, and anger, as these are more colored with non-modal voice qualities [31]. When prosody-based features were used in amalgamation with VQ-based features, an increase of 7.8% in recognition was reported in [31]. Along similar lines, in [32], Zhang explored a combination of prosody and VQ-based features using SVM on the Chinese natural emotional speech corpus and claimed a 10% increase over systems using only prosody-based features. Moreover, as claimed by Jacob in [33], the use of minimal VQ features such as jitter and shimmer can help develop faster systems. This study reported a 64.8% accuracy on English SER and 83.3% recognition accuracy on Hindi SER.

Several techniques have also been explored for the automatic processing of these features. For many audio processing tasks, including music onset detection, speech enhancement, and ASR, convolutional neural networks (CNNs) have been extensively studied to limit the number of parameters and memory requirements [34]. Recurrent Neural Networks (RNNs), on the other hand, enable parameters to be shared across time and have proven to be superior to conventional HMM-based models in a variety of speech and audio processing tasks [3]. Further, Sequence-to-Sequence models, Generative Models, and Reinforcement learning-based frameworks have attained much interest in recent years [34]. Table 1 summarizes the frequently used audio-based features along with relevant reported applications. Moreover, the audio-based features used in this work are also indicated.

Table 1
Feature selection for emotion classification.

Type	Features	Reported application
Prosody based features	HSFs of Raw Signal[a]	–
	HSFs of Pitch[a]	Busso et al. [6]
	Silence Length[a]	–
	Root Mean Square Energy (HSFs)*	Lee et al. [16]
	Zero-Crossing Rate (HSFs)[a]	Lee et al. [16]
	Auto-Correlation (HSFs)[a]	–
	Fundamental Frequency	Busso et al. [6]
Spectral features	Spectral Centroid (HSFs)[a]	–
	MFCCs (First 13)[a]	Ghazale and Hansen [24]
	GFCCs	Liu [25]
	HFCCs	Nagarajan et al. [26]
Voice quality features	Harmonics/HNR[a]	Cowie et al. [30]
	Jitter	Jacob [33]
	Shimmer	Jacob [33]

[a]Represents the used features in the proposed model.

2.2. Audio-only speech emotion recognition systems

Since the field of emotion recognition has garnered the interest of researchers, several audio-based frameworks have been proposed based on both conventional classifiers and neural networks. In 2000, Nicholson et al. [35] suggested a speaker-independent and context-independent one-class-in-one neural network [36] and achieved a 50% recognition rate on a Japanese database. This simple model of eight sub neural networks demonstrated that SER is feasible and neural networks are well suited for this task. With recent advancements in SER, the performance of neural network-based algorithms has shown promising results. Fayek et al. [37] proposed a straightforward deep neural network (DNN) model to recognize emotions from one-second frames of raw speech spectrogram and reported accuracies of 60.53% and 59.7% on the eNTERFACE'05 and SAVEE database, respectively. Following the successful application of DNNs, CNNs were also investigated, which offered excellent performances on several benchmark datasets leading to the development of several CNN-based approaches [38]. A 3D attention-based convolutional recurrent neural network (ACRNN), claiming respective recognition accuracies of 64.7% and 82.8% on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Emo-DB, as reported by Chen et al. [39]. Further, in [40], Hajarolasvadi and Demiral investigated the use of K-means clustering in conjugation with 3D CNN models and reported average recall values of 72.0% to 81.0% on different datasets. Other promising studies include 1D and 2D CNN networks combined with LSTM cells [10,11].

In addition to DNN frameworks, traditional classifiers such as the HMM [41], Gaussian mixture model (GMM) [42], SVM [43] have also been extensively used. Further, several new algorithms and pre-trained models have been researched. In 2011, Lee et al. put forward a structure to map an input speech utterance into one of the multiple emotion classes of the IEMOCAP database through subsequent layers of binary classifications, achieving an average unweighted recall (AUR) of 58.4% [16]. Similar to this approach, Liu et al. [44] introduced an SER system based on an extreme learning machine decision tree, separating one emotion at each level. They utilized correlation and Fisher Analysis as a feature selection method which reduced redundant features. This method yielded an accuracy of 89.6% on the Chinese Emotional Speech Corpus database. Recently, Liu et al. [45] proposed a brain emotion learning (BEL) model which offered an accuracy of 90.28% on the same dataset. In addition to the aforementioned, studies exploring bagged SVMs [46] and bag-of-visual words on audio spectrograms have also claimed accuracies up to 75% on various datasets [47].

2.3. Multimodal speech emotion recognition systems

As mentioned above, emotion is portrayed in humans via various modalities, including facial expressions, posture, words, and speech. Systems utilizing secondary features such as facial expressions and electroencephalogram (EEG) have been claimed to offer a higher recognition rate when used as a supporting modality to sound [7]. The uncertainty about the channel a user chooses to express their feelings reduces the risk of trusting a weak modality, making a multimodal approach more accurate than relying on a single modality [48].

In 2004, a novel SER system was proposed by Schuller et al. [49], which utilized speech and linguistic features as input modalities. In this study, an 8% reduction in the recognition error rate was reported compared to the unimodal baseline system. Further, when used as supporting modalities, facial expressions, body movement, and gestures resulted in a 10% improvement over unimodal systems [50]. Another important finding of this work was the superiority of feature-level fusion over a decision-level fusion of modalities. Developing on this approach, Soleymani et al. [51] used EEG, pupillary response, and gaze distance as inputs to a user-independent emotion recognition model. They reported recognition accuracies of 68.5% and 76.4% for three labels of valence and arousal, respectively. More recently, in 2020, Cui et al. proposed an end-to-end regional-asymmetric CNN model for emotion recognition from EEG and reported an accuracy of 95% on two public datasets [52]. The use of multiple advanced deep learning networks has also been linked to improved multimodal system efficiency. For instance, Tzirakis et al. [12] employed CNNs for emotion recognition from speech along with a deep residual network for visual features. They further integrated the network with LSTM cells to account for outliers and presented promising results. Similarly, Avots et al. [53] investigated the pre-trained AlexNet [54] model for extracting facial indicators from videos while incorporating the SVM framework for audio features. Unlike most databases that are recorded in simulated environments, the model was tested on the acted facial expressions in the wild's database, which contains audio and video recordings in environments close to the real world. In addition to visual features, video transcriptions have also been explored to automatically classify video segments and gain insights into emotional and contextual information [55].

Approaches based on acoustic and lexical features have also garnered interest in recent years. Yoon et al. investigated the use of audio-text modalities through dual recurrent neural networks, claiming accuracies of 68.8% to 71.8% on the IEMOCAP database [13]. By leveraging pre-trained word-embedding models such as the global vectors for word representation embedding (GloVe) and FastText for lexical features, Atmaja and Akagi [56] reported an accuracy of 49.6% on IEMOCAP. Gated recurrent units based RNNs have also shown state-of-the-art performance, achieving accuracies up to 76.9%, as reported by Ho et al. [57]. Similar to [56], they also utilized a pre-trained language model called bidirectional encoder representations (BERT) to compute the vector representation of textual features. When combined with MFCCs, recognition accuracy of 76.98% was claimed on an improvised IEMOCAP dataset. Lastly, the benefits of reinforcement learning and fuzzy commonsense have also been exploited in recent works. Li et al. [58] utilized a framework based on reinforcement learning for pre-selecting useful images for SER in FER and reported improvements over the state-of-the-art methods. Using a fuzzy logic classifier to predict the degree of a particular emotion, Chaturvedi et al. [59] successfully combined models of deep convolutional neural networks and fuzzy logic. Based on the above survey, the following three inferences are drawn.

(1) Suitability of DNNs for emotion recognition tasks is established.

(2) Performance improvement on the application of supporting modalities.

(3) Improvement in SER accuracy on an amalgamation of low-level descriptors and global level features.

In addition, the following research gaps are also observed. Prior SER research has been limited to employing multi-class classification, with very few articles on the hierarchical model approach. Further, most works employing multimodal systems have leveraged visual and biomedical signals as supporting modalities, while the use of textual transcriptions is relatively less researched. Furthermore, it appears that the potential of contextualized word embeddings has not been fully exploited.

Motivated by the above inferences, this work introduces a novel emotion classification method employing a hierarchical architecture of multiple DNN-based models. An amalgamation of audio-based features and word embeddings from ELMo v2 forms the input to the considered DNN networks. Further, the selected audio-based features are a combination of LLDs and HSFs of global level descriptors. The implemented hierarchy used in the model is based on the following factors: differences in acoustic features of emotions under consideration, the extent of ambiguity between the classes, and the number of data points available for each class. Finally, the suggested model utilizes feature-level concatenation of modalities that yield better performance than decision-level fusion.

3. Brief description of used datasets

This section presents a brief introduction of used datasets for the performance evaluation of the proposed model. To assess the model's adaptability, experiments were carried out on three emotional speech corpora of increasing complexity. For evaluation of the unimodal model, two audio-only datasets, namely, SAVEE and RAVDESS, were used. Similarly, for the multimodal system, the IEMOCAP dataset has been considered. Table 2 describes the original composition of each dataset along with baseline accuracy.

3.1. Surrey audio-visual expressed emotion (SAVEE) database

The SAVEE Database [61] is an emotional corpus recorded in the British-English language to develop automatic emotion recognition systems. It supports both audio and visual modalities, though, in this work, only the audio modality has been used. In total, it includes 480 utterances of 3 standard, 2 emotion-specific, and 10 generic sentences. It consists of 7 emotions — anger,

Table 2
Summary of used datasets.

Dataset	Number of utterances	Human recognition accuracy	Baseline accuracy
SAVEE	480 (7 emotions)	66 ± 2.5%	61%
RAVDESS	1440 (8 emotions)	62%	Not Available (N/A)
IEMOCAP	5331 (4 emotions)	72%	66.2% [60]

disgust, fear, happiness, surprise, neutral, and sadness, recorded from 4 native male speakers and evaluated by 10 independent subjects. The dataset has a reported human recognition accuracy of 66.5 ± 2.5% for audio modality.

3.2. Ryerson audio-visual database of emotional speech and song (RAVDESS)

RAVDESS [62] is an emotional speech and song dataset in English containing 7356 audio and video files. RAVDESS consists of two modalities, i.e., audio and audio–video, wherein, for this work, only the audio files have been used. Further, RAVDESS contains 3036 song files that have been ignored since the focus of this study is to detect emotion in speech, thereby reducing the files to 1440. The audio samples are categorized into 8 emotions: anger, disgust, fear, happiness, surprise, neutral, calm, and sadness, and recorded by 24 professional actors (12 male, 12 female). Additionally, each expression is recorded at two levels of emotional intensity. The evaluations were done by 247 individuals, with a further 72 people providing an evaluation for test-retest data.

3.3. Interactive emotional dyadic motion capture (IEMOCAP) database

IEMOCAP [63] is a multimodal, multispeaker English speech emotion recognition dataset comprising scripted and spontaneous utterances. In this dataset, the impromptu sessions consist of three 10-minute plays with explicit emotional content, which the actors are asked to memorize and rehearse. This comprises 55% of the corpus. For spontaneous sessions, speakers are given eight hypothetical scenarios and are asked to improvise emotions. These sessions constitute 45% of the corpus. In addition, IEMOCAP supports audio and visual modalities, including text transcriptions of all utterances. Out of these, only audio and text modalities have been used in this work. The dataset is recorded by 10 actors over 5 sessions, each 12 h long. Each session is recorded by one male and one female actor and evaluated by 3 separate annotators. While the dataset is categorized into 10 emotions: angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and other, the 4 most commonly used emotions (angry, happy, neutral, and sad) have been utilized, per similar works. Overall, this work utilizes audio and text transcription of 5331 utterances.

4. Proposed methodology

The proposed framework is described in this section. Used feature extraction methods for audio and textual features are introduced, followed by a description of the data preprocessing techniques. The detailed hierarchical tree structures and model architectures are also presented.

4.1. Feature extraction

The extraction techniques of audio and textual features are described in the following subsections. Fig. 2 depicts a flowchart outlining the proposed feature extraction methodology.

4.1.1. Audio based features

A total of 33 features: 16 prosody-based, 16 spectral-based, and one voice quality-based feature were extracted to encapsulate the distinct characteristics of each emotion in an audio sample. It should be noted that for all three datasets, the feature set was kept consistent. All audio files were sampled at 44.1 kHz. Further, framing and windowing techniques were applied to split the audio file into small intervals of time to extract local-level features. This was done using a fast Fourier transform (FFT) of window size 1024 and a hop size of 512 to facilitate overlap and reduce spectral leakage. These operations were implemented using Librosa [64]. Subsequently, feature extraction was carried out on each frame. The extracted values were also aggregated over the complete time interval using HSFs to include global level features for selected features. The mean, standard deviation, and the range of the raw audio signals were calculated and aggregated over the frames. These 3 selected HSFs were kept the same for all further aggregations.

A sample of the raw audio signal for 6 emotions from the SAVEE dataset is shown in Fig. 3. Post that, the root-mean-square energy (RMSE) of the signal was calculated for each frame using (1), wherein, $x(n)$ represents the signal, and N is the number of frames. The HSFs of RMSE values were considered as the features.

$$RMSE = \sqrt{(\frac{1}{N}\sum_{i=0}^{N-1}|x(n)^2|)} \tag{1}$$

Next, the duration of silence in the audio samples was calculated. For this work, the silence threshold was set at 0.4 times the RMSE value. Any value lower than the threshold was considered as silent, and the ratio of the total number of silent intervals to the number of intervals yielded the desired value for the silence of the sample. This was followed by the computation of the zero-crossing rate (ZCR) and its HSFs for each frame using (2).

$$ZCR_n = \frac{1}{2N}\sum_{m=N-n+1}^{N}|sgn[x(m)] - sgn[x(m+1)]| \tag{2}$$

where,

$$sgn[x(m)] = 1 \; for \; x(m) > 0$$
$$sgn[x(m)] = 0 \; for \; x(m) = 0$$
$$sgn[x(m)] = -1 \; for \; x(m) < 0$$

Further, the sample's pitch was computed using the Parselmouth [65] library and was aggregated for the feature vector. The final prosody-based feature, autocorrelation $R(t)$ of the signal, was computed using (3).

$$R(\tau) = \frac{1}{t_{max} - t_{min}}\int_{t_{min}}^{t_{max}} x(t) * x(t+\tau) * dt \tag{3}$$

where τ represents the introduced delay.

Furthermore, autocorrelation was also selected as it has been reported to play a pivotal part in human audio signal processing [66]. Additionally, to consider the voice quality features of the audio sample, the harmonics of the sample were extracted using Librosa. By default, Librosa decomposes the audio sample into both the harmonic and percussive components. For implementation, only the harmonics were considered. As reported by Rao and Koolagudi [29], the vocal tract characteristics in humans are best represented by spectral features. The proposed model also considers 16 spectral features (13 MFCCs and 3 HSFs of spectral centroid) in light of their work. The spectral centroids (SCs) were calculated using (4).

$$SC_t = \frac{\sum_{n=1}^{N} m_t(n) * n}{\sum_{n=1}^{N} m_t(n)} \tag{4}$$

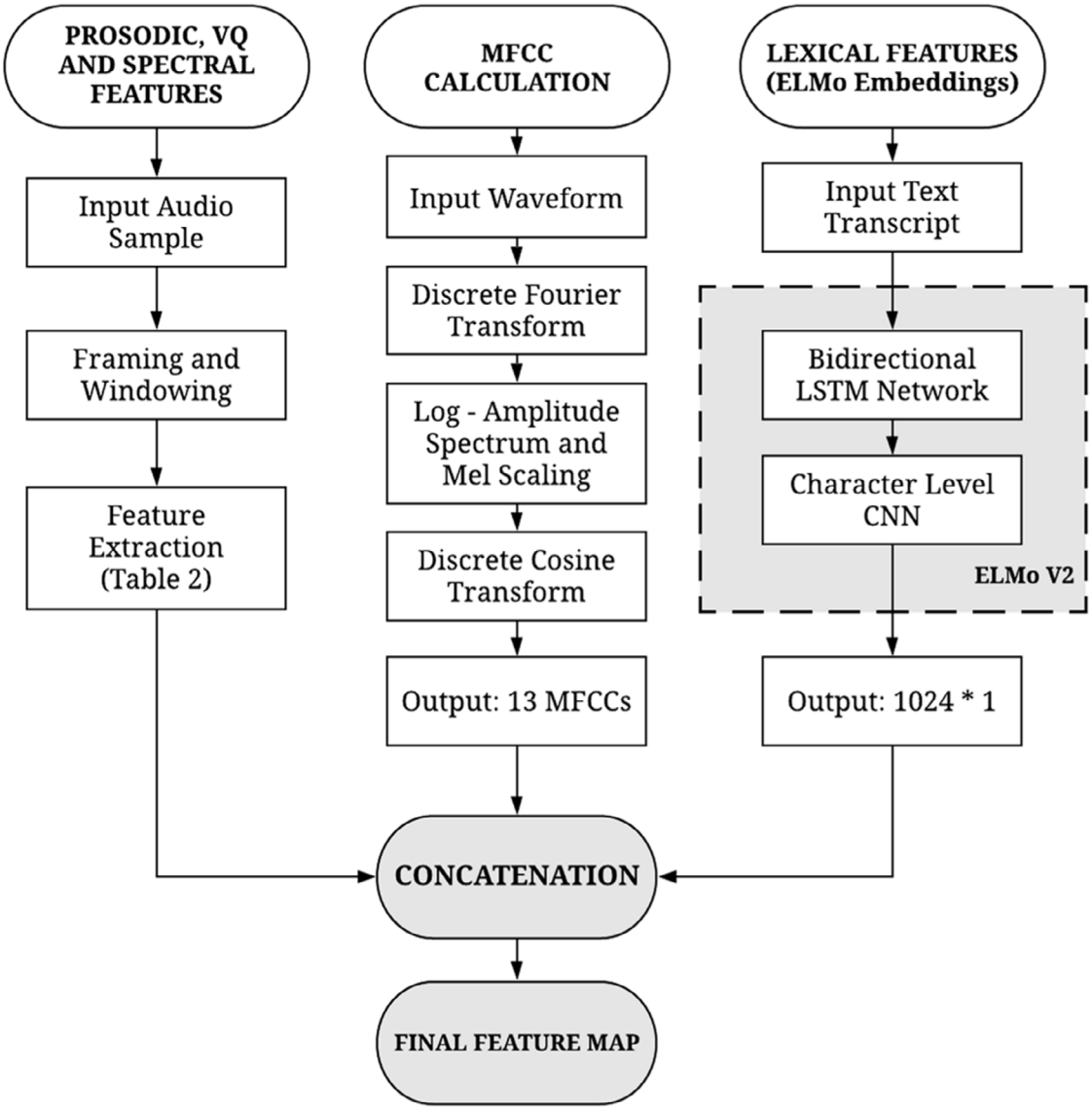

Fig. 2. Overview of the proposed feature extraction methodology.

where n represents the number of frequency bins and $m_t(n)$ represents the magnitude of that frequency bin.

After this, 13 Mel-scaled cepstral coefficients of the audio sample were calculated, as the human ear perceives frequency in the Mel scale. Framing and windowing were performed as mentioned before to compute the MFCCs for an audio sample. The time-domain samples were then transformed to the frequency domain using discrete Fourier transform (DFT). Next, the logarithm of the amplitude was computed to obtain the log-amplitude spectrum of the sample, which was further scaled to the Mel-frequency range (5).

$$M(f) = 1125 * \ln(1 + \frac{f}{700}) \tag{5}$$

Finally, the discrete cosine transform of the spectrum was computed to get the MFCCs of the audio samples. Only 13 MFFCs are considered for this implementation, in line with the present research [46]. All 33 features are concatenated to arrive at the audio modality feature vector. The MFCCs for the 6 emotions represented in SAVEE are shown in Fig. 4. It may be noted that the represented image contains 13 Mel coefficients displayed in dB (Decibels) per color bar.

4.1.2. Lexical features

In order to leverage information from textual data, vector representations of words were computed using ELMo. ELMo is a state-of-the-art NLP framework developed by AllenNLP and trained on the one Billion Word Benchmark. It computes word vectors on top of a two-layer bidirectional language model (biLM). The biLM model consists of two stacked layers of LSTM networks and uses character-level CNN to convert words of a given string into raw word vectors. As input to biLM is computed from characters rather than words, it can capture the inner structure of words. Unlike traditional word embeddings that produce the same vector for a word used in different contexts, ELMo employs the entire input sentence for computing word embeddings.

Moreover, the bidirectional approach considers the relation of a word with both the next and previous words. The higher-level LSTM states capture context-dependent aspects of word meaning. In contrast, the lower-level states model aspects of syntax [15], enabling a word to be represented by different word vectors under a different context. It also addresses distinguishing antonyms from synonyms, which is challenging for traditional language models such as GloVe and FastText due to the similar distributional information of these words [67]. The presented

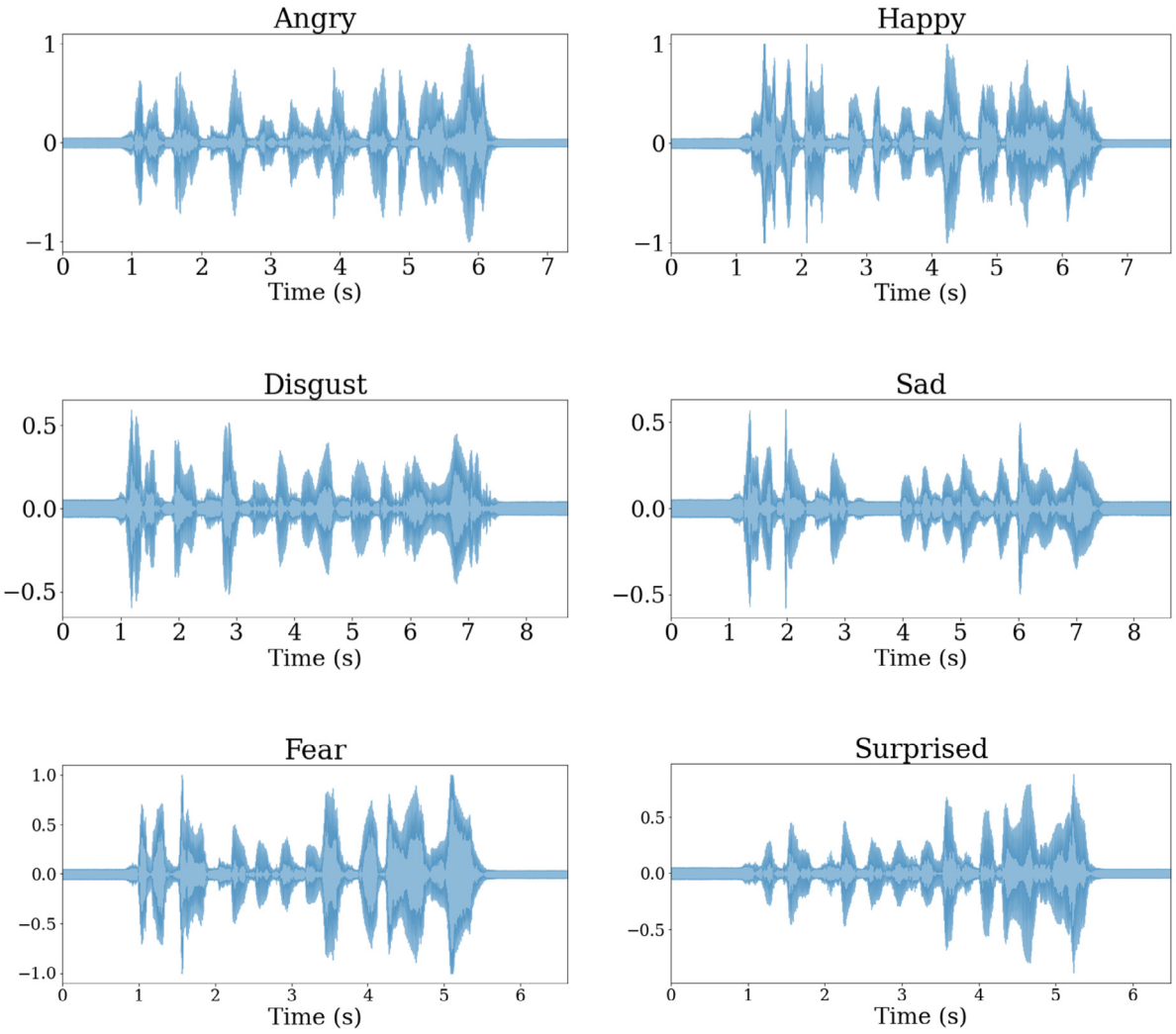

Fig. 3. Raw audio waveforms of 6 emotions: SAVEE dataset.

work uses the baseline ELMo v2 model for embedding textual information from IEMOCAP into a 1024-dimension vector, as shown in Fig. 2. For the combination of emotions selected in this work, 85.3% of all text transcriptions were unique, while the remaining 14.7% were repeated. When used independently, these repeated word embeddings may cause similar vectors to be produced for different emotions. These word embeddings are not used independently in this work but are instead concatenated with their corresponding audio features. As a result, the final input vector becomes unique even for repeated text transcriptions, enabling the model to better distinguish between their output classes.

4.2. Preprocessing

In order to ensure that data remained consistent throughout each experiment, various data preprocessing steps were applied before and after feature extraction. All audio samples were zero-padded to maintain the same length within an experiment. Further, framing and windowing of each audio sample were performed. A Hamming window of length M was used (6) to reduce spectral leakage during the computation of the FFT for each sample.

$$\omega(n) = 0.54 - (0.46 * \cos\frac{2\pi n}{M-1}) \quad (6)$$

During feature extraction, scaling was done to increase the effect of certain features. All extracted harmonics and computed autocorrelations were scaled by a factor of 1000. After feature extraction, the data was normalized to improve the feature set's generalization and reduce the effects of speaker and gender variation [7]. This normalization was explicitly required as datasets are recorded by multiple actors and speakers of varying gender and age. All features were scaled to the range (0,1) globally.

Post normalization, the feature set was split into training and testing datasets in an 80:20 ratio. Further, it was noted that apart from SAVEE, both IEMOCAP and RAVDESS displayed a significant imbalance of data points in several classes. Emotions like happy and angry contained multiple data points in contrast to less frequently used emotions like disgust. Data over-sampling was used to resolve this imbalance. It should be noted that this was performed only on training data. Further, for IEMOCAP, the emotion "Excitement" was merged with "Happy" while "Frustration" was merged with "Anger." The emotions "Disgust" and "Other" were not considered to establish a common comparing platform with other reported works that have utilized a similar combination. The composition of training and testing datasets are shown in Table 3 for all three datasets.

4.3. Description of proposed hierarchical model

This section describes the structure and hierarchy of the proposed model. The general hierarchical approach remains the same

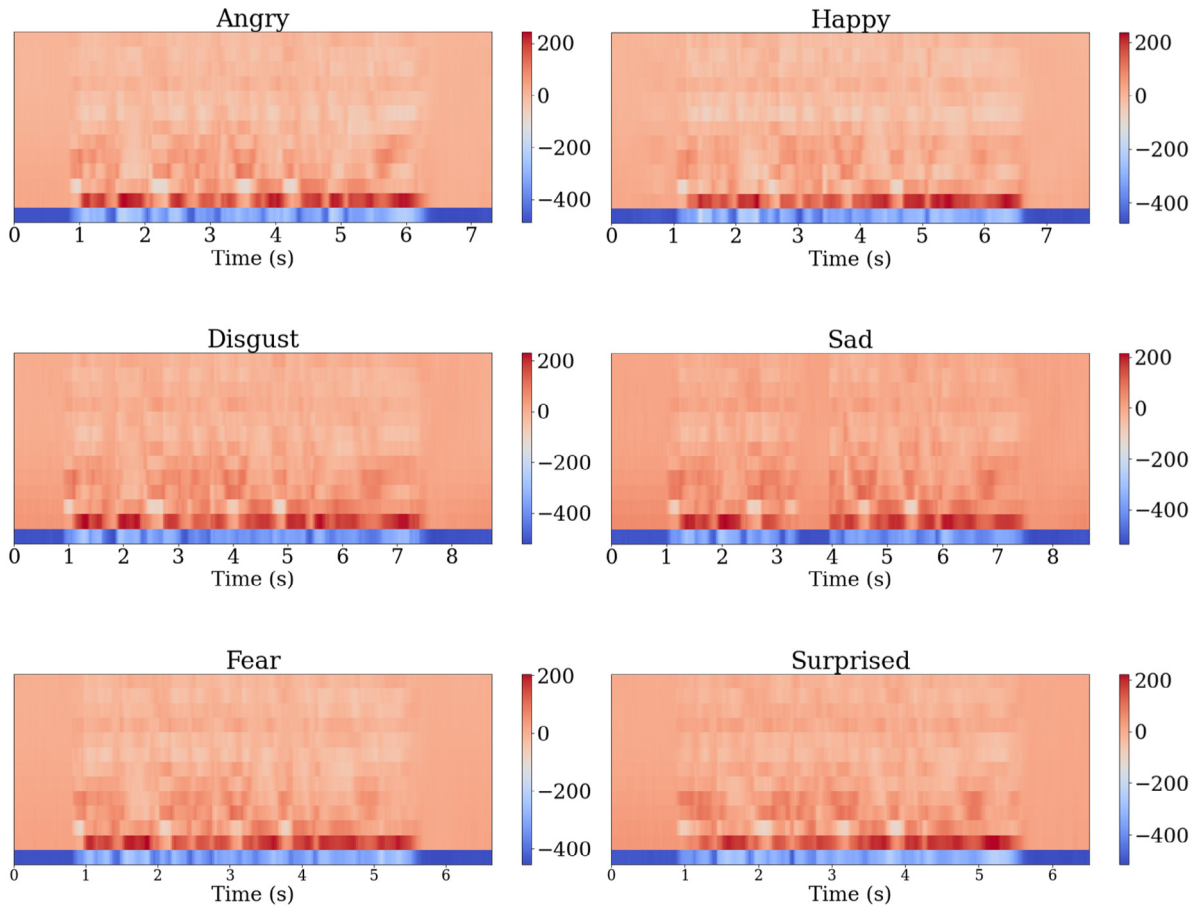

Fig. 4. Mel-scaled cepstral coefficients of 6 emotions: SAVEE dataset.

Table 3
Composition of each emotion class in training and testing datasets for SAVEE, RAVDESS, and IEMOCAP.

Dataset	Emotion	Training set	Testing set
IEMOCAP	Angry (Frustration Merged) (A)	2338	614
	Sad (SA)	837	211
	Happy (Excitement Merged) (H)	1304	332
	Neutral (N)	1389	319
SAVEE	Angry	48	12
	Sad	48	12
	Happy	48	12
	Surprised (SU)	48	12
	Disgust (D)	48	12
	Fear (F)	48	12
	Neutral	96	24
RAVDESS	Angry	154	38
	Sad	154	38
	Happy	154	38
	Surprised	153	39
	Disgust	154	38
	Fear	153	39
	Calm (C)	153	39
	Neutral	77	19

for each dataset, while the architecture of the specific model itself is presented in further respective subsections. Section 4.3.1 introduces the general steps that are adopted to build dataset-specific hierarchical models. Further, Sections 4.3.2 and 4.3.3 describe the specific models used on RAVDESS, SAVEE, and IEMOCAP datasets, respectively.

4.3.1. General hierarchy

While the proposed model is tested on three datasets, it can easily be adapted to other datasets. This section describes the method to select the hierarchy for a generalized dataset. For instance, for a dataset comprising five emotions – E1, E2, E3, E4, and E5, and two modalities – audio and text, the generalized hierarchy can be shown in Fig. 5. In the case of a unimodal dataset, only the audio features are used.

While the process of extracting features and embeddings remains the same, the hierarchical structure differs as it is developed based on a set of characteristics unique to each dataset. It should be noted that these are not exhaustive for any dataset. Following are the criteria for the development of a hierarchy:

(1) The aim is to increase the recognition rate of relatively easy tasks by classifying emotions that are easily differentiated early in the hierarchy. Emotions that are not easily differentiable are tackled lower in the tree to reduce error propagation [16]. For identification of these emotions, audio features described in Section 4.1.1 are used. An example of this is depicted in Fig. 6. As shown in Fig. 6(a), using RMSE plots, it can be seen that emotions "Happy" and "Sad" can easily be separated and are therefore given precedence. On the other hand, "Sad" and "Disgust" (Fig. 6(b)), are difficult to distinguish and are thus tackled lower in the tree.

(2) To improve adaptability and accuracy, multiple emotions can also be separated at the nodes. For example, in Fig. 5, model 1 separates two sets of emotions: (E1, E2, E3) v/s (E4, E5) based on their arousal. Highly empathic emotions like anger, happiness, and frustration can be pitted against a set of emotions with low arousal like sadness and fear. This ensures that the most

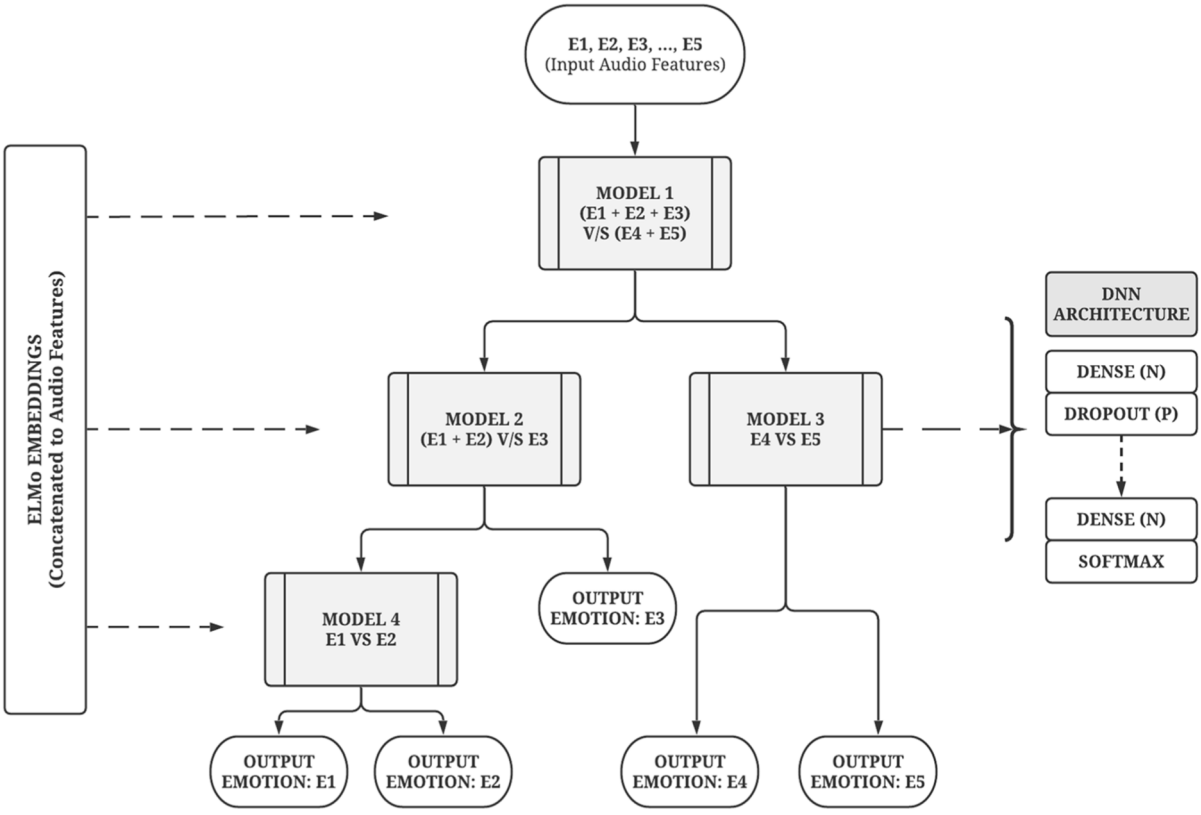

Fig. 5. Generic hierarchical structure for five emotions and multimodal inputs.

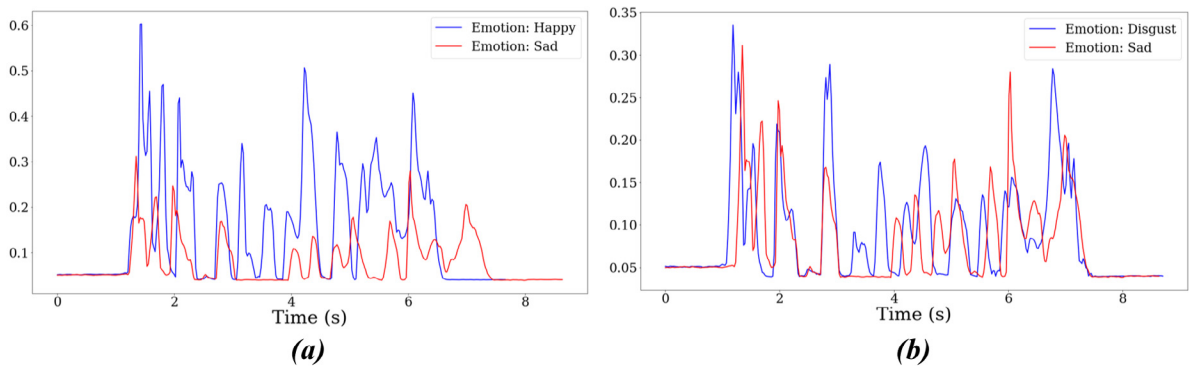

Fig. 6. RMSE plots of emotions: SAVEE Dataset. (a) Happy and Sad. (b) Disgust and Sad.

discriminate tasks are handled higher in the tree to prevent error propagation.

(3) Another important consideration is the sample size of each emotion. Emotions with a lower number of samples than others are placed lower in the tree as they are prone to a greater error rate in recognition.

Based on the above set of rules, a hierarchical model can be developed for any distribution of emotion.

4.3.2. Hierarchy for RAVDESS and SAVEE datasets

To evaluate audio-only performance, the proposed model is first tested on two datasets using only audio-based features. The original emotion set introduced by SAVEE (7 emotions) and RAVDESS (8 emotions) were used without any modifications for classification. The hierarchical structures for both SAVEE and RAVDESS are shown in Fig. 7(a) and Fig. 8, respectively. Since both datasets deal only with audio-based emotions, the input layer consists of a 33 1-dimensional vector consisting of features described in Section 4.1.1. Each dense layer comprises a fully connected hidden layer using the ReLU activation function. The complete architecture is shown in Fig. 7(b), and it remains the same for both datasets. Dropout layers are utilized to reduce overfitting, and both systems are trained on the Adam optimizer with a categorical cross-entropy loss function.

4.3.3. Hierarchy for IEMOCAP dataset

As mentioned before, the 4 most common emotions, i.e., angry, happy, neutral, and sad, have been used for classification for ease of comparison with similar works. Emotion "Excitement" was merged with happiness, while "Frustration" with anger. The structure of the hierarchical model is shown in Fig. 9(a). For the dataset under consideration, each classifier node in the figure has a multimodal DNN architecture described as follows. After the preprocessing of textual data, corresponding word embeddings obtained from ELMo are concatenated with the preprocessed audio features into a single vector. This vector is then fed into

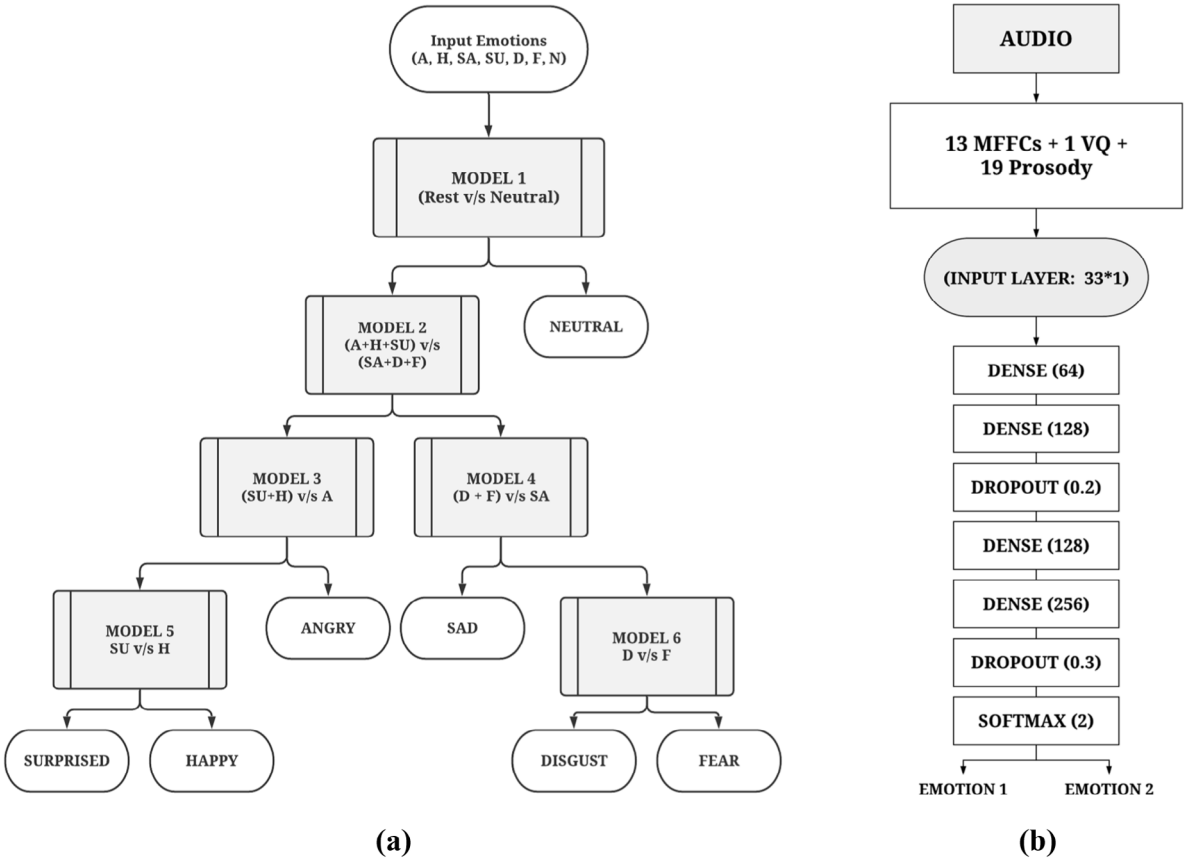

Fig. 7. (a) Hierarchy for SAVEE dataset. (b) Architecture of DNN nodes.

Table 4
Performance comparison: RAVDESS dataset.

S. no.	Reported application	Method	Train test split ratio	Validation criteria	Accuracy (%)	Average UR (%)
1.	Bhavan et al. [46]	Bagged ensemble of SVM on MFCCs, and spectral centroid.	90 - 10	10 Fold CV	75.7	N/A
2.	Zeng et al. [68]	DNN on audio spectrogram	N/A	5 Fold CV	64.3	64.5
3.	Shegokar and Sircar [69]	SVM on continuous wavelet transform	N/A	5 Fold CV	60.1	N/A
4.	**Proposed model**	**Hierarchical DNN based classifier on audio features**	**80 - 20**	**10 Fold CV**	**81.2**	**79.7**

the network, as shown in Fig. 9(b). Each dense layer shown in the figure utilizes a ReLU activation function. The dense layers are followed by dropout layers with a drop probability of 0.2 and 0.3, respectively, to avoid overfitting. The outputs of dense and dropout layers are finally used to predict the probabilities of the emotion classes using the softmax function. The proposed model is trained on the Adam optimizer and relies on categorical cross-entropy for the loss metric. Independent tests were conducted on both unimodal and multimodal inputs to benchmark their performances.

5. Results and discussion

The effectiveness of the proposed hierarchical classification system was evaluated on the datasets introduced in Section 3. For all training and testing purposes, the implementation of networks from the Keras [70] library for Python (TensorFlow Backend version 2.3.1) was used. Additionally, the pre-trained ELMo v2 model was obtained from the TensorFlow Hub [71] for lexical features. All three models were trained on the Adam optimizer with a categorical cross-entropy loss function. To prevent overfitting, early stopping was employed wherein training was halted when the difference between two successive validation accuracies was below a threshold. The receiving operator characteristic (ROC) curves were plotted for each task using Matplotlib [72]. As shown in [16], AUR is a more suitable metric for assessing performance in non-uniform distributions, and hence, both average accuracies and average unweighted recalls are reported.

Tables 4–6 present the performances of the developed model on three datasets and compare the same with recent potential works. Independent tests were performed on each dataset using 10-fold cross-validation (CV), and their average was recorded. It should be noted that all works considered for comparison have used different train, test, and validation splits, and this information has been indicated in the respective tables. Fig. 10 presents the individual class recognition accuracies for each emotion. For

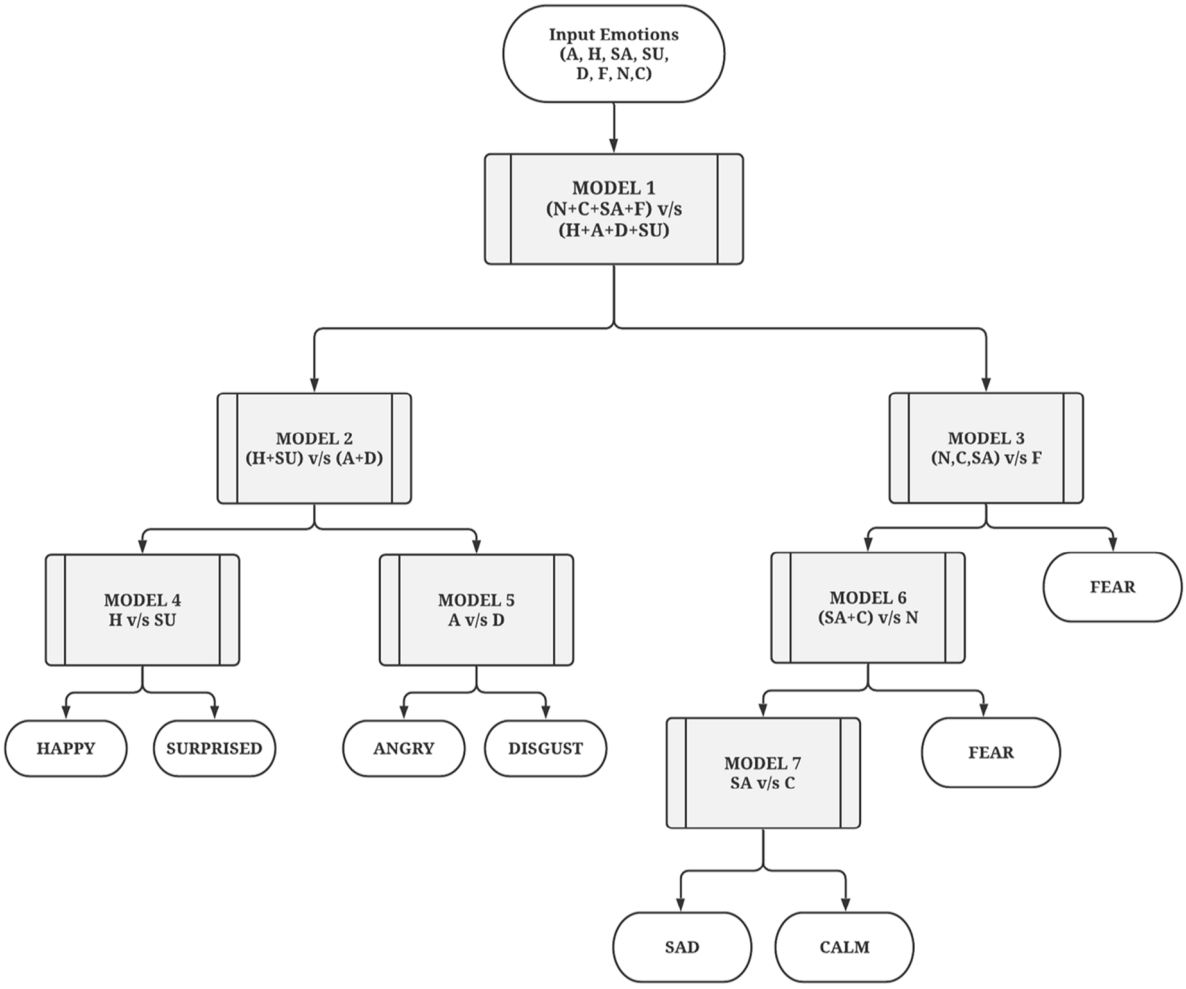

Fig. 8. Hierarchy for RAVDESS dataset.

Table 5
Performance comparison: SAVEE dataset.

S. No.	Reported application	Method	Train test split ratio	Validation criteria	Accuracy (%)	Average UR (%)
1.	Fayek et al. [37]	Real-time DNN on raw spectrogram	90 - 10	10 Fold CV	59.7	N/A
2.	Liu et al. [45]	Brain learning model using genetic algorithm	85 - 15	N/A	76.4	N/A
3.	Avots et al. [53]	SVM for audio features along with AlexNet for FER	90 - 10	10 Fold CV	79.4	N/A
4.	**Proposed model**	**Hierarchical DNN based classifier on audio features**	**80 - 20**	**10 Fold CV**	**81.7**	**80.5**

the IEMOCAP dataset (Table 6), the reported score of the proposed model corresponds to the scores obtained by the multimodal system, as shown in Fig. 9. Three independent tests were run on the IEMOCAP dataset to compare the unimodal and multimodal performance of the model. Results obtained for these experiments are shown in Table 7 and are further elaborated. Fig. 11 depicts the ROC plots of each model in the hierarchy of the RAVDESS dataset. The model nomenclature is kept similar to the one shown in Fig. 8. These results are further elaborated and analyzed to draw valuable conclusions.

For performance evaluation and relative assessment on RAVDESS, three notable works, Bhavan et al. [46], Zeng et al. [68], and Shegokar and Sircar [69], have been considered. A brief of these works, along with the comparative performance results, are summarized as follows. Bhavan et al. used a bagged ensemble of SVMs with a Gaussian kernel to recognize emotions from speech [46]. They utilized a feature set composed of MFCCs, spectral centroids, and high-level deltas and concluded that the use of MFCCs alone leads to poor recognition accuracy. The proposed model achieves an increase of 5.5% over [46] due to HSFs and LLDs of prosody, spectral, and VQ-based features being used. In contrast to [46], Zeng et al. used a combination of CNN and GResNets in a deep learning fashion on audio spectrograms to automatically extract features and recognize emotion [68]. The

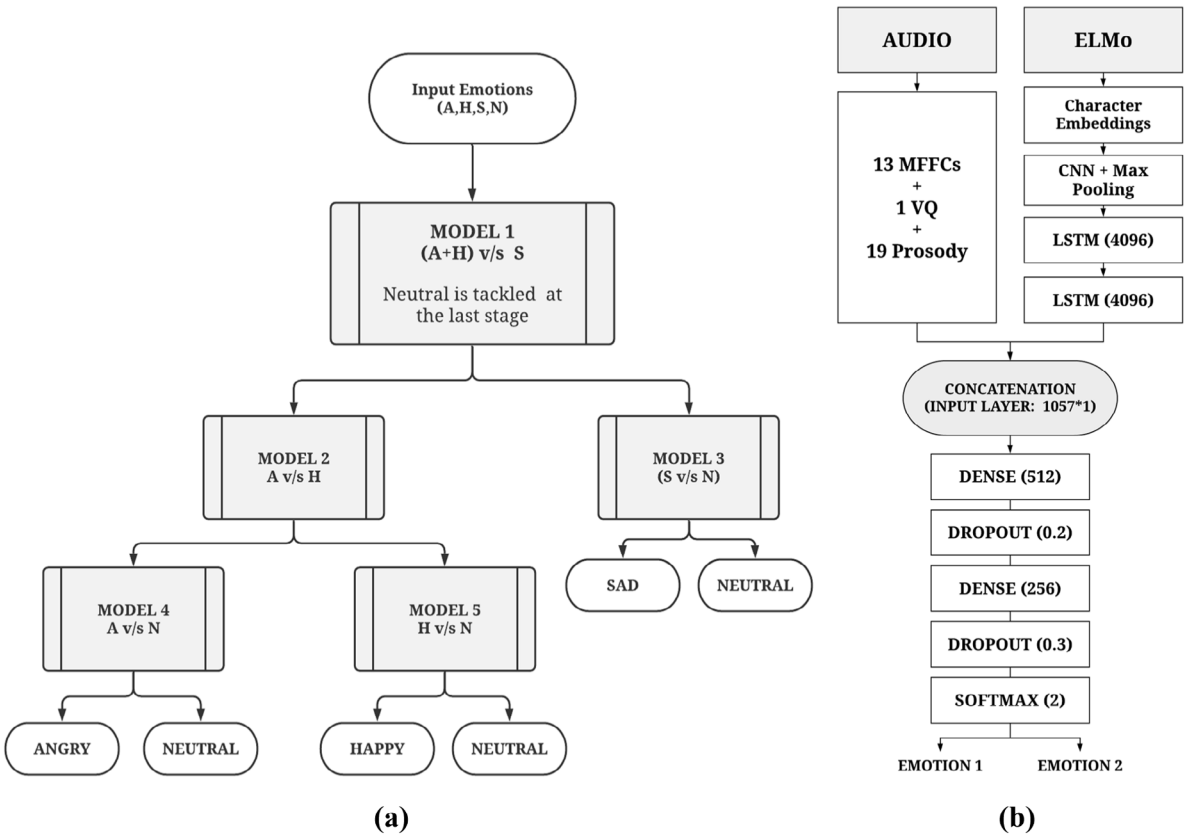

Fig. 9. (a) Hierarchy for IEMOCAP dataset. (b) Architecture of DNN nodes.

Table 6
Performance comparison: IEMOCAP dataset.

S. no.	Reported application	Method	Emotions selected	Train test split ratio	Validation criteria	Accuracy (%)	Average UR (%)
1.	Yao et al. [8]	Multi-task learning — CNN, DNN, and RNNs	A, H, S, N	N/A	Leave One Out CV	57.1	58.3
2.	Yoon et al. [13]	Dual RNN with acoustic features and text	A, H, S, N (Excitement merged with Happy)	80 - 15 - 5	5 Fold CV	71.8	N/A
3.	Lee et al. [16]	Binary hierarchical decision tree approach	A, H, S, N	90 - 10	Leave One Out CV	58.4	N/A
4.	Chen et al. [39]	3-D ACRNN using spectrogram	A, H, S, N (Improvised only)	80 - 10 - 10	10 Fold CV	64.7	N/A
5.	Ho et al. [57]	Multi-Level Multi-Head Attention-Based RNN using audio and text	A, H, S, N	N/A	Leave One Out 10 Fold CV	73.2	N/A
6.	**Proposed Multimodal model**	**Multimodal Hierarchical DNN based classifier on audio and lexical features**	A, H, S, N (Excitement Merged with Happy and Angry with Frustration)	**80 - 20**	**10 Fold CV**	**74.5**	**73.2**

model put forward in this work offers a 16.9% enhancement over their claimed accuracy of 64.3%. This clearly demonstrates that the use of a binary hierarchical approach works better than a multi-class approach. Similar to [46], Shegokar and Sircar [69] also used a support vector machine on a set of prosodic and continuous wavelet transform coefficients. They reported an accuracy of 60.1% compared to the 81.2% achieved by the proposed model leading to an improvement of 21.1%. Based on this comparison, it is inferred that DNN based networks are more suitable for SER tasks than traditional classifiers.

For SAVEE, similar to RAVDESS, three recent works, Fayek et al. [37], Liu et al. [45], and Avots et al. [53], have been considered for relative performance assessment. Similar to the architecture proposed in this work, Fayek et al. used a DNN on the

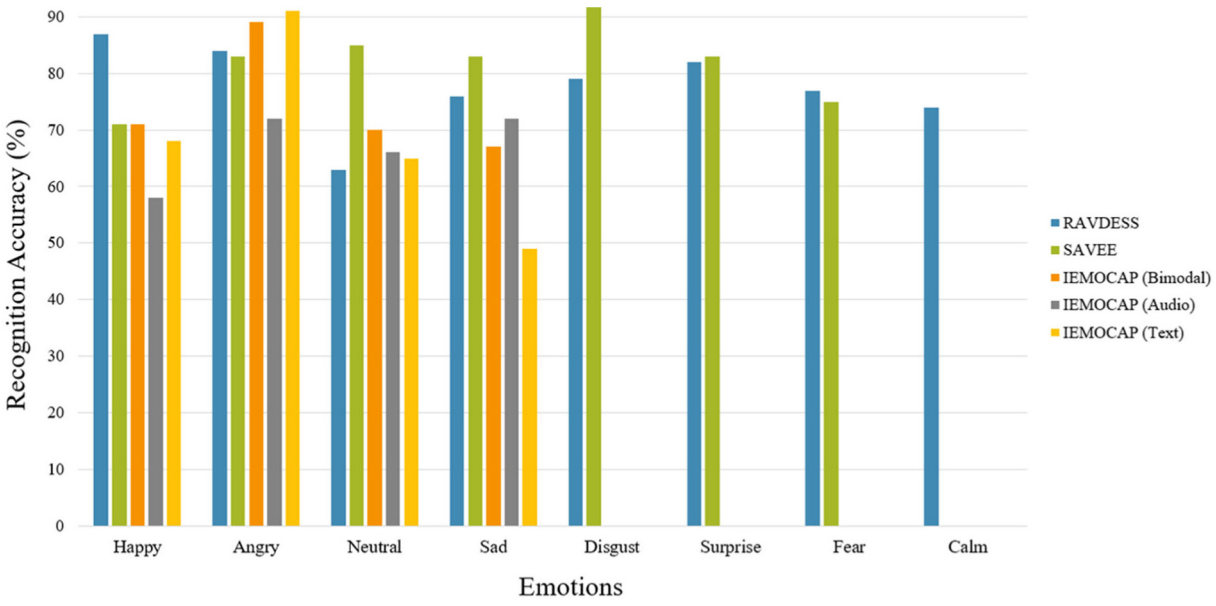

Fig. 10. Individual class accuracies for each emotion.

raw spectrogram of audio samples [37]. They obtained an average accuracy of 59.7% compared to the 81.74% of the proposed hierarchical model. This suggests that the application of spectrogram alone is not sufficient to encapsulate the emotion expressed in an audio sample. Deviating from commonly used neural networks, Liu et al. [45] used an improved BEL along with the genetic algorithm and achieved an accuracy of 76.4%. This promising result reaffirms the hypothesis that taking into account the way humans perceive emotions can yield better results than generic neural networks in SER. Utilizing both audio and facial features, Avots et al. employed a multimodal interface and achieved an accuracy of 79.4% [53]. This approach demonstrated that the addition of a supporting modality increases the recognition rate of models. However, the proposed work demonstrated that the consideration of textual features as an added modality could improve performance by 2.3%. The same is reflected in the forthcoming comparison of the unimodal and multimodal performance of the IEMOCAP dataset (Table 7).

The multimodal system, evaluated on IEMOCAP, has been compared with the following five recent works: Yao et al. [8], Yoon et al. [13], Lee et al. [16], Chen et al. [39], and Ho et al. [57]. Yao et al. [8] put forward a framework that integrated three distinct classifiers — DNN, CNN, and RNN. The extracted frame-level LLDs, segment-level Mel-spectrograms, and utterance-level outputs of HSFs were respectively passed through RNN, CNN, and DNN layers. Their fusion network achieved a weighted accuracy of 57.1% and unweighted accuracy of 58.3%, which was significantly higher than each network. Utilizing a similar audio feature set but an additional textual modality, the proposed model, achieved an improvement of 17.4%, proving the advantage of multiple modalities. As discussed before, Yoon et al. [13] employed a combination of audio and lexical features on the IEMOCAP dataset via a dual encoder RNN network and achieved an accuracy of 71.8%. The ELMo v2 model for lexical features used in this work increases the accuracy by 2.7%, proving its superiority. Similar to the proposed technique, Lee et al. first introduced a binary hierarchy-based classification approach and employed Bayesian logistic regression [16]. Their work tested the suggested framework on only acoustic features of IEMOCAP and reported an unweighted recall accuracy of 58.46%. As shown in Table 6, the proposed approach achieves a significant improvement of 16.1% over this baseline paper due to dual modalities.

Table 7
Comparison of proposed unimodal and multimodal systems: IEMOCAP dataset.

Modality	Accuracy (%)
Audio	68.3
Text	68.0
Audio + Text	**74.5**

In another experiment, Chen et al. [39] used delta and delta–delta spectrogram as features to reduce the influence of irrelevant emotional factors, leading to lesser instances of misclassification. Besides, they suggested using an attention layer to generate utterance-level features by discovering emotionally relevant parts of the CRNN features. They reported accuracy of 64.74% against the proposed 74.5%. Ho et al. [57] employed a multimodal method wherein they used the BERT model to extract higher dimensional word embeddings from the textual transcriptions. This was combined with audio features and fed to an RNN network for prediction. They achieved a comparable accuracy of 73.2%. The proposed model outperforms [57] by 1.2%, thereby demonstrating the superiority of the ELMO v2 model over BERT in this task.

Further, from the results presented in Table 7, it can be inferred that using a supporting modality improves recognition accuracy by 5.75% over the audio-only system and by 6.5% on the text-only system. The average class accuracies also improve on using a multimodal system, as shown in Fig. 10. It can also be observed that the use of textual features helps improve the recognition of emotions that can be identified from the general usage and context of words (Happy and Angry). The use of only audio features fails to completely distinguish between these emotions due to similar arousal and energy distribution. This causes the statistical features to average around a similar point, creating confusion for the audio-only classifier. In contrast, audio features are essential to separate emotions based on basic features like arousal (Sad). Because emotions, like sadness, differ vastly in energy distribution from happy and angry, it is easily identified by the audio-only classifier. It is also inferred that the performance of all three systems remains the same for an ambiguous class like “neutral”. This is one area where improvement can be achieved by incorporating more supporting modalities like facial expressions and body movements, which efficiently recognize apathy.

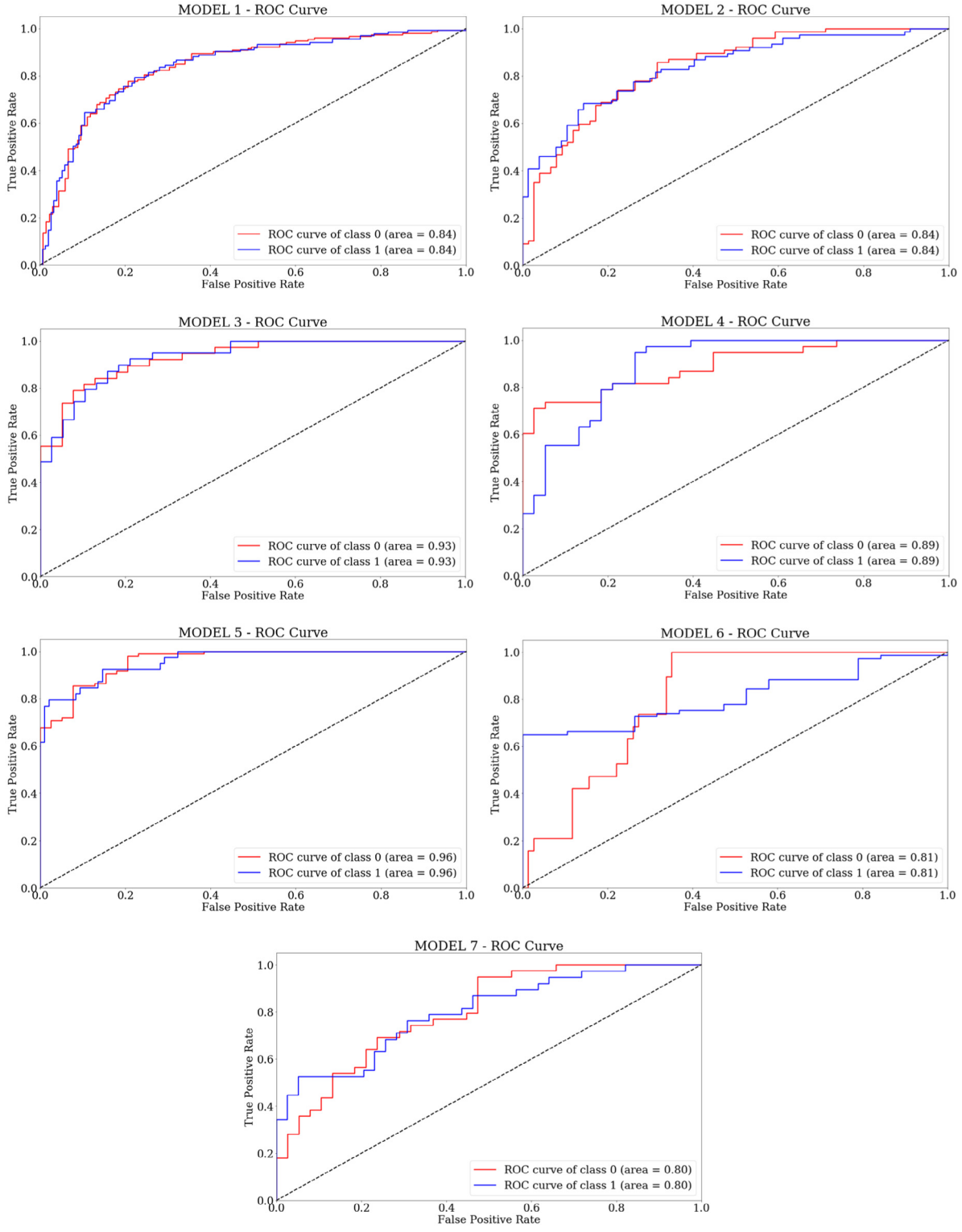

Fig. 11. ROC curves for the RAVDESS hierarchical structure.

Fig. 11 presents the ROCs of the RAVDESS model depicting the performance of each node in the model. As hypothesized before, the model manages to maintain a good accuracy for all nodes in the structures. While the primary nodes have significantly high accuracy, even the lower nodes achieve a respectable performance as the structure is built to ensure the least error propagation in the tree.

To further evaluate the developed models based on the training time, 10 independent trials were conducted, and their average times were recorded, as shown in Table 8. Comparison of the timing aspect was not possible as only a few papers have reported their training time.

From the above-presented results and discussions, the merits of the proposed method can be summarized as follows:

Table 8
Average training time for ten trials.

Dataset	Modality	Average training time (s)
RAVDESS	Audio only	79.0
SAVEE	Audio only	134.2
IEMOCAP	Audio only	176.1
IEMOCAP	Text only	44.7
IEMOCAP	Bimodal (Audio + Text)	1600.0

(1) As is evident from Tables 4–6, the hierarchical model presents superior accuracies and average unweighted recall values compared to notable recent works. The evaluation of the proposed approach on three different datasets further reinforces its credibility and adaptability.

(2) The developed procedure described in Section 4.3.1 is adaptable and can easily be followed to implement a general hierarchy for a given dataset.

(3) Since each node in the suggested model is characterized by a relatively simple neural network based on dense hidden layers, the presented framework is computationally less expensive than most recent works utilizing networks such as CNN, RNN, and ACRNN.

(4) The total training and testing time is also reasonably shorter as the datasets used in each step are smaller than the previous step. This offers a significant advantage for real-time applications and can contribute to an improved user experience.

(5) The hierarchical approach gives a clear indication of which emotions are being confused the most and to what degree, enabling faster and efficient optimization.

6. Conclusion and future work

This work successfully presented a hierarchical DNN based approach to emotion recognition from speech and text in both unimodal and multimodal systems. A combination of 33 acoustic features and their higher statistical functions were employed to encapsulate information from both local and global level segments. For the lexical features, the use of ELMo v2 word embeddings was investigated to extract contextual and character-based features from text transcriptions. The performances of both systems were investigated using the average unweighted recall, average recognition accuracies, and receiver operating characteristic curves on three popular datasets, namely, RAVDESS, SAVEE, and IEMOCAP. The proposed hierarchical unimodal model offered average unweighted recall scores of 81.2% and 81.7% on the RAVDESS and SAVEE datasets, respectively, while the proposed multimodal system offered an AUR of 74.5% on IEMOCAP. These results, being superior to their counterparts, clearly suggest that a hierarchical structure is a potential technique for efficiently recognizing emotion from speech with higher accuracy than traditional multi-class classification techniques. In addition, this work also demonstrated the feasibility of leveraging contextualized word embeddings to improve the performance of emotion recognition systems.

Though the proposed framework achieves superior performance over its potential counterparts, a natural progression of this work would be to develop a generalized architecture for the hierarchical model that may be fine-tuned across all datasets. Further, the application of other modalities such as facial features and body movements would be worth exploring in conjugation with the proposed model. Future studies could also assess the interdependence of polarity detection and emotion recognition and utilize the same to improve the model's accuracy. Furthermore, since this paper does not utilize automatic feature extraction using CNNs and RNNs, future works could usefully explore multi-task learning to combine automatic and manual feature extraction methods. Lastly, the investigation of various methods, such as feature reduction techniques, to reduce the training time of such hierarchical models would also be a fruitful area of research.

CRediT authorship contribution statement

Prabhav Singh: Conceptualization, Methodology, Data curation, Coding, Writing - draft preparation. **Ridam Srivastava:** Conceptualization, Methodology, Data curation, Coding, Writing - draft preparation. **K.P.S. Rana:** Conceptualization, Reviewing and editing, Supervision. **Vineet Kumar:** Editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Q. Luo, H. Tan, Facial and speech recognition emotion in distance education system, in: Proc. 2007 Int. Conf. Intell. Pervasive Comput. IPC 2007, 2007, pp. 483–486, http://dx.doi.org/10.1109/IPC.2007.55.
[2] S.N. Zisad, M.S. Hossain, K. Andersson, Speech Emotion Recognition in Neurological Disorders using Convolutional Neural Network, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Springer Science and Business Media Deutschland GmbH, 2020, pp. 287–296, http://dx.doi.org/10.1007/978-3-030-59277-6_26.
[3] S. Latif, J. Qadir, A. Qayyum, M. Usama, S. Younis, Speech technology for healthcare: Opportunities, challenges, and state of the art, IEEE Rev. Biomed. Eng. 14 (2021) 342–356, http://dx.doi.org/10.1109/RBME.2020.3006860.
[4] A. Ashok, J. John, Facial Expression Recognition System for Visually Impaired, in: Lect. Notes Data Eng. Commun. Technol., Springer, 2019, pp. 244–250, http://dx.doi.org/10.1007/978-3-030-03146-6_26.
[5] F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, N. Nguyen-Thien, Emotion on the road-necessity, acceptance, and feasibility of affective computing in the car, Adv. Human-Computer Interact. (2010) http://dx.doi.org/10.1155/2010/263593.
[6] C. Busso, S. Lee, S. Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection, IEEE Trans. Audio, Speech Lang. Process. 17 (2009) 582–596, http://dx.doi.org/10.1109/TASL.2008.2009578.
[7] M.B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Commun. 116 (2020) 56–76, http://dx.doi.org/10.1016/j.specom.2019.12.001.
[8] Z. Yao, Z. Wang, W. Liu, Y. Liu, J. Pan, Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN, Speech Commun. 120 (2020) 11–19, http://dx.doi.org/10.1016/j.specom.2020.03.005.
[9] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth, Acoustic emotion recognition: A benchmark comparison of performances, in: Proc. 2009 IEEE Work. Autom. Speech Recognit. Understanding, ASRU, 2009, pp. 552–557, http://dx.doi.org/10.1109/ASRU.2009.5372886.
[10] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomed. Signal Process. Control. 47 (2019) 312–323, http://dx.doi.org/10.1016/j.bspc.2018.08.035.
[11] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, B. Schuller, Speech emotion classification using attention-based LSTM, IEEE/ACM Trans. Audio Speech Lang. Process. 27 (2019) 1675–1685, http://dx.doi.org/10.1109/TASLP.2019.2925934.
[12] P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B.W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE J. Sel. Top. Signal Process. 11 (2017) 1301–1309, http://dx.doi.org/10.1109/JSTSP.2017.2764438.
[13] S. Yoon, S. Byun, K. Jung, Multimodal speech emotion recognition using audio and text, in: 2018 IEEE Spok. Lang. Technol. Work. SLT 2018 - Proc, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 112–118, http://dx.doi.org/10.1109/SLT.2018.8639583.
[14] D. Panda, D. Das Chakladar, T. Dasgupta, Multimodal system for emotion recognition using eeg and customer review, in: Adv. Intell. Syst. Comput, Springer, 2020, pp. 399–410, http://dx.doi.org/10.1007/978-981-15-2188-1_32.

[15] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, Deep contextualized word representations, 2018, https://arxiv.org/abs/1802.05365.

[16] C.C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan, Emotion recognition using a hierarchical binary decision tree approach, Speech Commun. 53 (2011) 1162–1171, http://dx.doi.org/10.1016/j.specom.2011.06.004.

[17] Z. Xiao, E. Dellandrea, W. Dou, L. Chen, Automatic hierarchical classification of emotional speech, in: Ninth IEEE Int. Symp. Multimed. Work. (ISMW 2007), Institute of Electrical and Electronics Engineers (IEEE), 2008, pp. 291–296, http://dx.doi.org/10.1109/ism.workshops.2007.56.

[18] Q.R. Mao, Y.Z. Zhan, A novel hierarchical speech emotion recognition method based on improved DDAGSVM, Comput. Sci. Inf. Syst. 7 (2010) 211–222, http://dx.doi.org/10.2298/CSIS1001211Q.

[19] A. Hassan, R.I. Damper, Multi-class and hierarchical SVMs for emotion recognition, in: Proc. 11th Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2010, 2010, pp. 2354–2357.

[20] R.S. Lazarus, Relational meaning and discrete emotions, in: K.R. Scherer, A. Schorr, T. Johnstone (Eds.), Appraisal Processes in Emotion: Theory, Methods, Research, Oxford University Press, 2001, pp. 37–67.

[21] S. Kuchibhotla, H.D. Vankayalapati, R.S. Vaddi, K.R. Anne, A comparative analysis of classifiers in emotion recognition through acoustic features, Int. J. Speech Technol. 17 (2014) 401–408, http://dx.doi.org/10.1007/s10772-014-9239-3.

[22] J. Martinez, H. Perez, E. Escamilla, M.M. Suzuki, Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) techniques, in: CONIELECOMP 2012-22nd Int. Conf. Electron. Commun. Comput., 2012, pp. 248–251, http://dx.doi.org/10.1109/CONIELECOMP.2012.6189918.

[23] N. Dave, Feature extraction methods LPC, PLP and MFCC in speech recognition, Int. J. Adv. Res. Eng. Technol. 1 (2013) 1–5.

[24] S.E. Bou-Ghazale, J.H.L. Hansen, A comparative study of traditional and newly proposed features for recognition of speech under stress, IEEE Trans. Speech Audio Process. 8 (2000) 429–442, http://dx.doi.org/10.1109/89.848224.

[25] G.K. Liu, Evaluating gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech, 2018, ArXiv. http://arxiv.org/abs/1806.09010.

[26] N. Sugan, N.S.S. Srinivas, N. Kar, L.S. Kumar, M.K. Nath, A. Kanhe, Performance comparison of different cepstral features for speech emotion recognition, in: 2018 Int. CET Conf. Control. Commun. Comput. IC4 2018, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 266–271, http://dx.doi.org/10.1109/CETIC4.2018.8531065.

[27] J.C. Lin, C.H. Wu, W.L. Wei, Error weighted semi-coupled hidden markov model for audio-visual emotion recognition, IEEE Trans. Multimed. 14 (2012) 142–156, http://dx.doi.org/10.1109/TMM.2011.2171334.

[28] B. Schuller, G. Rigoll, M. Lang, Hidden Markov model-based speech emotion recognition, in: Proc. - IEEE Int. Conf. Multimed. Expo, IEEE Computer Society, 2003, pp. I401–I404, http://dx.doi.org/10.1109/ICME.2003.1220939.

[29] K.S. Rao, S.G. Koolagudi, R.R. Vempada, Emotion recognition from speech using global and local prosodic features, Int. J. Speech Technol. 16 (2013) 143–160, http://dx.doi.org/10.1007/s10772-012-9172-2.

[30] R. Corive, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J.G. Taylor, Emotion recognition in human–computer interaction, IEEE Signal Process. Mag. 18 (2001) 32–80, http://dx.doi.org/10.1109/79.911197.

[31] M. Lugger, B. Yang, The relevance of voice quality features in speaker independent emotion recognition, in: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., 2007, http://dx.doi.org/10.1109/ICASSP.2007.367152.

[32] S. Zhang, Emotion recognition in chinese natural speech by combining prosody and voice quality features, in: Adv. Neural Networks - ISNN 2008, Springer Berlin Heidelberg, 2008, pp. 457–464, http://dx.doi.org/10.1007/978-3-540-87734-9_52.

[33] A. Jacob, Speech emotion recognition based on minimal voice quality features, in: Int. Conf. Commun. Signal Process. ICCSP 2016, Institute of Electrical and Electronics Engineers Inc., 2016, pp. 886–890, http://dx.doi.org/10.1109/ICCSP.2016.7754275.

[34] S. Latif, H. Cuayáhuitl, F. Pervez, F. Shamshad, H.S. Ali, E. Cambria, A survey on deep reinforcement learning for audio-based applications, 2021, http://arxiv.org/abs/2101.00240.

[35] J. Nicholson, K. Takahashi, R. Nakatsu, Emotion recognition in speech using neural networks, Neural Comput. Appl. 9 (2000) 290–296, http://dx.doi.org/10.1007/s005210070006.

[36] J.D. Markel, A.H. Gray, Linear Prediction of Speech, Springer Berlin Heidelberg, Berlin, Heidelberg, 1976, http://dx.doi.org/10.1007/978-3-642-66286-7.

[37] H.M. Fayek, M. Lech, L. Cavedon, Towards real-time speech emotion recognition using deep neural networks, in: 2015, 9th Int. Conf. Signal Process. Commun. Syst. ICSPCS 2015 - Proc, Institute of Electrical and Electronics Engineers Inc., 2015, http://dx.doi.org/10.1109/ICSPCS.2015.7391796.

[38] Q. Mao, M. Dong, Z. Huang, Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks, IEEE Trans. Multimed. 16 (2014) 2203–2213, http://dx.doi.org/10.1109/TMM.2014.2360798.

[39] M. Chen, X. He, J. Yang, H. Zhang, 3-D convolutional recurrent neural networks with attention model for speech emotion recognition, IEEE Signal Process. Lett. 25 (2018) 1440–1444, http://dx.doi.org/10.1109/LSP.2018.2860246.

[40] N. Hajarolasvadi, H. Demirel, 3D CNN-based speech emotion recognition using K-means clustering and spectrograms, Entropy 21 (2019) 479, http://dx.doi.org/10.3390/e21050479.

[41] A.D. Dileep, C. Chandra Sekhar, Hmm based intermediate matching kernel for classification of sequential patterns of speech using support vector machines, IEEE Trans. Audio, Speech Lang. Process. 21 (2013) 2570–2582, http://dx.doi.org/10.1109/TASL.2013.2279338.

[42] D. Neiberg, K. Elenius, K. Laskowski, Emotion recognition in spontaneous speech using GMMs, in: Proceedings ICSLP-2006, Pittsburgh, 2006, pp. 809–812.

[43] Y. Pan, P. Shen, L. Shen, Speech emotion recognition using support vector machine, Int. J. Smart Home. 6 (2012) 101–108, http://dx.doi.org/10.1109/kst.2013.6512793.

[44] Z.T. Liu, M. Wu, W.H. Cao, J.W. Mao, J.P. Xu, G.Z. Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, Neurocomputing 273 (2018) 271–280, http://dx.doi.org/10.1016/j.neucom.2017.07.050.

[45] Z.T. Liu, Q. Xie, M. Wu, W.H. Cao, Y. Mei, J.W. Mao, Speech emotion recognition based on an improved brain emotion learning model, Neurocomputing 309 (2018) 145–156, http://dx.doi.org/10.1016/j.neucom.2018.05.005.

[46] A. Bhavan, P. Chauhan, Hitkul, R.R. Shah, Bagged support vector machines for emotion recognition from speech, Knowledge-Based Syst. 184 (2019) 104886, http://dx.doi.org/10.1016/j.knosys.2019.104886.

[47] E. Spyrou, R. Nikopoulou, I. Vernikos, P. Mylonas, Emotion recognition from speech using the bag-of-visual words on audio segment spectrograms, Technologies 7 (2019) 20, http://dx.doi.org/10.3390/technologies7010020.

[48] J. Wagner, F. Lingenfelser, E. André, J. Kim, Exploring fusion methods for multimodal emotion recognition with missing data, IEEE Trans. Affect. Comput. 2 (2011) 206–218, http://dx.doi.org/10.1109/T-AFFC.2011.12.

[49] B. Schuller, G. Rigol, M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture, in: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc, 2004, http://dx.doi.org/10.1109/icassp.2004.1326051.

[50] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, K. Karpouzis, Multimodal emotion recognition from expressive faces, body gestures and speech, in: IFIP Int. Fed. Inf. Process., Springer, Boston, MA, 2007, pp. 375–388, http://dx.doi.org/10.1007/978-0-387-74161-1_41.

[51] M. Soleymani, M. Pantic, T. Pun, Multimodal emotion recognition in response to videos, IEEE Trans. Affect. Comput. 3 (2012) 211–223, http://dx.doi.org/10.1109/T-AFFC.2011.37.

[52] H. Cui, A. Liu, X. Zhang, X. Chen, K. Wang, X. Chen, EEG-Based emotion recognition using an end-to-end regional-asymmetric convolutional neural network, Knowledge-Based Syst. 205 (2020) 106243, http://dx.doi.org/10.1016/j.knosys.2020.106243.

[53] E. Avots, T. Sapiński, M. Bachmann, D. Kamińska, Audiovisual emotion recognition in wild, Mach. Vis. Appl. (2019) 975–985, http://dx.doi.org/10.1007/s00138-018-0960-9.

[54] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM. 60 (2017) 84–90, http://dx.doi.org/10.1145/3065386.

[55] L. Stappen, A. Baird, E. Cambria, B.W. Schuller, Sentiment analysis and topic recognition in video transcriptions, IEEE Intell. Syst. 36 (2) (2021) 88–95, http://dx.doi.org/10.1109/MIS.2021.3062200.

[56] B.T. Atmaja, M. Akagi, Dimensional speech emotion recognition from speech features and word embeddings by using multi-task learning, APSIPA Trans. Signal Inf. Process. 9 (2020) http://dx.doi.org/10.1017/ATSIP.2020.14.

[57] N.H. Ho, H.J. Yang, S.H. Kim, G. Lee, Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network, IEEE Access. 8 (2020) 61672–61686, http://dx.doi.org/10.1109/ACCESS.2020.2984368.

[58] H. Li, H. Xu, Deep reinforcement learning for robust emotional classification in facial expression recognition, Knowledge-Based Syst. 204 (2020) 106172, http://dx.doi.org/10.1016/j.knosys.2020.106172.

[59] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270, http://dx.doi.org/10.1016/j.patrec.2019.04.024.

[60] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, A. Hussain, Multimodal sentiment analysis: Addressing key issues and setting up the baselines, IEEE Intell. Syst. 33 (2018) 17–25, http://dx.doi.org/10.1109/MIS.2018.2882362.

[61] Jackson, Philip, ul haq, Sana, Surrey Audio-Visual Expressed Emotion (SAVEE) database, 2011, http://kahlan.eps.surrey.ac.uk/savee/.

[62] S.R. Livingstone, F.A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American English, PLoS One 13 (2018) e0196391, http://dx.doi.org/10.1371/journal.pone.0196391.
[63] 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, 2007.
[64] B. McFee, V. Lostanlen, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, J. Mason, D. Ellis, E. Battenberg, S. Seyfarth, R. Yamamoto, K. Choi, viktorandreevichmorozov, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Hereñú, F.-R. Stöter, P. Friesch, A. Weiss, M. Vollrath, T. Kim, Librosa/librosa: 0.8.0, 2020, http://dx.doi.org/10.5281/ZENODO.3955228.
[65] Y. Jadoul, B. Thompson, B. de Boer, Introducing parselmouth: A python interface to praat, J. Phon. 71 (2018) 1–15, http://dx.doi.org/10.1016/j.wocn.2018.07.001.
[66] Y. Soeta, Psychophysiological evidence of an autocorrelation mechanism in the human auditory system, in: Adv. Clin. Audiol. InTech, 2017, http://dx.doi.org/10.5772/66198.
[67] M. Etcheverry, D. Wonsever, Unraveling antonym's word vectors through a siamese-like network, in: Proc. 57th Annu. Meet. Assoc. Comput. Linguist, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 3297–3307, http://dx.doi.org/10.18653/v1/P19-1319.
[68] Y. Zeng, H. Mao, D. Peng, Z. Yi, Spectrogram based multi-task audio classification, Multimedia Tools Appl. 78 (2019) 3705–3722, http://dx.doi.org/10.1007/s11042-017-5539-3.
[69] P. Shegokar, P. Sircar, Continuous wavelet transform based speech emotion recognition, in: 2016, 10th Int. Conf. Signal Process. Commun. Syst. ICSPCS 2016 - Proc, Institute of Electrical and Electronics Engineers Inc., 2016, http://dx.doi.org/10.1109/ICSPCS.2016.7843306.
[70] F. Chollet, et al., Keras. GitHub, 2015, https://github.com/fchollet/keras.
[71] Elmo | TensorFlow Hub, (n.d.).https://tfhubdev/google/elmo/3.
[72] J. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. 9 (3) (2007) 90–95, http://dx.doi.org/10.5281/zenodo.592536.