Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.044466

ARTICLE

Improved Speech Emotion Recognition Focusing on High-Level Data Representations and Swift Feature Extraction Calculation

Akmalbek Abdusalomov[1], Alpamis Kutlimuratov[2], Rashid Nasimov[3] and Taeg Keun Whangbo[1,*]

[1]Department of Computer Engineering, Gachon University, Sujeong-Gu, Seongnam-Si, Gyeonggi-Do, 13120, Korea

[2]Department of AI.Software, Gachon University, Seongnam-Si, 13120, Korea

[3]Department of Artificial Intelligence, Tashkent State University of Economics, Tashkent, 100066, Uzbekistan

*Corresponding Author: Taeg Keun Whangbo. Email: tkwhangbo@gachon.ac.kr

Received: 31 July 2023 Accepted: 18 October 2023 Published: 26 December 2023

ABSTRACT

The performance of a speech emotion recognition (SER) system is heavily influenced by the efficacy of its feature extraction techniques. The study was designed to advance the field of SER by optimizing feature extraction techniques, specifically through the incorporation of high-resolution Mel-spectrograms and the expedited calculation of Mel Frequency Cepstral Coefficients (MFCC). This initiative aimed to refine the system's accuracy by identifying and mitigating the shortcomings commonly found in current approaches. Ultimately, the primary objective was to elevate both the intricacy and effectiveness of our SER model, with a focus on augmenting its proficiency in the accurate identification of emotions in spoken language. The research employed a dual-strategy approach for feature extraction. Firstly, a rapid computation technique for MFCC was implemented and integrated with a Bi-LSTM layer to optimize the encoding of MFCC features. Secondly, a pretrained ResNet model was utilized in conjunction with feature Stats pooling and dense layers for the effective encoding of Mel-spectrogram attributes. These two sets of features underwent separate processing before being combined in a Convolutional Neural Network (CNN) outfitted with a dense layer, with the aim of enhancing their representational richness. The model was rigorously evaluated using two prominent databases: CMU-MOSEI and RAVDESS. Notable findings include an accuracy rate of 93.2% on the CMU-MOSEI database and 95.3% on the RAVDESS database. Such exceptional performance underscores the efficacy of this innovative approach, which not only meets but also exceeds the accuracy benchmarks established by traditional models in the field of speech emotion recognition.

KEYWORDS

Feature extraction; MFCC; ResNet; speech emotion recognition

1 Introduction

The affective disposition of human beings, in other words, their emotional state, serves as a significant determinant in how they interact with one another and with machines. This emotional underpinning is not just trivial but rather it deeply informs and molds a multitude of communication pathways. Such pathways extend from visual cues evident in facial expressions, auditory signals highlighted in vocal characteristics, to the semantic structures embedded within verbal exchanges.

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Notably, spoken language, an integral component of human communication, is not simply a means of exchanging information. It serves as a critical medium for the articulation and conveyance of a myriad of human emotions. This emotive dimension of speech is often implicitly encoded in tonality, pitch, pace, and volume, thereby making it a rich and multifaceted source of information [1]. Consequently, when we consider the interaction between humans and machines, especially in the context of developing interfaces that are natural and user-friendly, the ability to correctly understand, interpret, and respond to these emotional nuances becomes increasingly paramount. The pursuit of enabling machines with this proficiency akin to an emotionally intelligent human listener has the potential to profoundly enhance the level of empathy, engagement, and overall effectiveness of human-machine interactions. In this light, the affective computing field stands as a promising domain, committed to endowing machines with these emotionally cognizant capabilities.

The process of discerning emotional cues in verbal communication, also known as Speech Emotion Recognition (SER), bears immense relevance for gaining insights into human communicative behaviors. At its core, SER provides a mechanism to interpret a speaker's emotional condition utilizing the acoustic and prosodic attributes of their utterances. This creates an intriguing nexus of linguistics, psychology, and artificial intelligence [2–4]. The broad-ranging implications and applications of SER techniques have permeated various fields, such as trading [5], tele-consultation (health prediction) [6], education [7], each with its unique requirements and objectives.

Emotion recognition hinges on pinpointing features that capture emotional cues within datasets. Essentially, finding a precise feature set for this task is intricate due to emotions' multifaceted and personal nature. These 'emotion cues' in data symbolize attributes that resonate with specific emotions. Recognizing these cues, like pitch variations in speech, is vital for creating emotion-savvy models. While references [8–9] show no consensus on the best features, the quest is to identify features broad yet precise enough for different emotions, a continuing challenge in the field.

In recent years, a predominant portion of studies have leaned towards the use of deep learning models, trained specifically to distill relevant feature sets from the data corpus [10–12]. The precision of categorization in the realm of SER issues is predominantly influenced by the procurement and choice of effective features. For instance, the extraction of features from speech signals leverages techniques such as MFCC, mel-scale spectrogram, tonal power, and spectral flux. To enhance learning performance by reducing feature size, the Deer Hunting with Adaptive Search (DH-AS) algorithm is employed for optimal feature selection in the research [11]. These selected features are then subjected to emotion classification via the Hybrid Deep Learning (HDL) approach, which combines both Deep Neural Network (DNN) and Recurrent Neural Network (RNN).

Many trusted studies in SER have leaned on acoustic features like Pitch, Energy, MFCC, Discrete Fourier Transform (DFT), among others [13]. Digging deeper, the effectiveness of emotion classification in speech emotion recognition is primarily anchored in the capability to distill and pinpoint the most influential features within speech data. These distinguishing traits shed light on a speaker's emotional state. Therefore, the steps of feature extraction and selection are paramount, often shaping the classification algorithm's outcome. Notably, MFCCs furnish a detailed insight into speech signals compared to rudimentary acoustic attributes. At its core, MFCCs capture the power spread across frequencies in a speech signal, presenting a glimpse of the speaker's unique vocal tract dynamics.

Our proposed solution employs a bifurcated strategy to mitigate these significant issues. First, capitalizing on our pioneering rapid computation of MFCC, we introduce an expedited method for the extraction of MFCC from vocal signals, serving as our MFCC feature encoder. The primary objective of this strategy is to enhance the efficiency of the feature extraction process, thereby enabling accurate and swift decomposition of speech data. This novel computational tactic emphasizes improving the velocity and effectiveness of MFCC attribute extraction. We further incorporate a bidirectional Long Short-Term Memory (Bi-LSTM) layer, tasked with capturing and encoding the complex temporal dependencies within the MFCC sequence. Second, serving as our Mel-spectrogram feature encoder, we exploit a pre-trained Residual Neural Network (ResNet) along with feature Stats pooling and fully connected layers to extract high-resolution spectrograms from Mel-spectrogram features.

Subsequently, the outputs from both feature encoders are concatenated and introduced to a CNN, and subsequently to a Fully Connected Network (FCN). Ultimately, a softmax function is employed to facilitate emotion classification. This all-encompassing strategy is formulated to enhance the intricacy and efficiency of our speech emotion recognition paradigm.

The main contributions of the proposed system are as follows:

- A novel system for SER has been introduced, exhibiting remarkable accuracy when benchmarked against existing base models. This cutting-edge approach signifies a promising trajectory for subsequent research within the SER sphere.
- Pioneering techniques were employed to extract fast MFCC features and Mel-spectrogram features from audio signals.
- A novel method for the swift calculation of MFCC features was formulated. The accelerated computation of MFCC features greatly improves the efficiency of the feature extraction phase in the SER process, thereby reducing the overall processing time and increasing system responsiveness. The extracted MFCC features provide essential insights into the spectral properties of the speech signal, making them invaluable for emotion detection and recognition tasks.
- A parallel processing methodology was introduced for the implementation of the Hanning window and value reduction operations.
- Collectively, the outcomes of this research have significantly enriched our comprehension of SER, offering crucial insights into the development of more proficient speech recognition models. The ramifications of these results span various fields, including emotion identification in human-robot exchanges, progressions in speech therapy methodologies, and enhancements in psychiatric health diagnostics.

The structure of this manuscript is as follows: Section 2 provides a comprehensive review of current research on SER modeling employing deep learning methodologies. Sections 3 and 4 are dedicated to a comprehensive elucidation of the proposed SER model, supplemented by empirical substantiation of its effectiveness and comparative analyses against established benchmarks. The aim is to equip the reader with an in-depth comprehension of the model's structure and competencies, as well as positioning it within the broader context of the SER domain. In Section 5, a definitive summary is provided, along with discussions on potential avenues for future exploration. The document culminates with a reference list that includes a broad spectrum of recent academic publications related to SER.

2 Related Work

In this section, we offer a synopsis of the prevailing scholarly works pertinent to the subject of SER. By examining the current state of research in this area, we hope to provide valuable context and insights, and to highlight key areas where further investigation is needed. This overview should serve as a valuable foundation upon which to build future research endeavors, driving innovation and advancement in this exciting field of study.

Over the recent timespan, systems underpinned by deep neural network architectures [14–17] have demonstrated significant triumphs in discerning emotions from vocal signals. In particular, the combination of CNNs and LSTMs in end-to-end methods [18–19] provided a robust way to capture both spatial features (through CNN) and temporal dynamics (through LSTM) in speech data. This fusion allowed for the thorough analysis of the speech signal, leveraging the strength of both models to yield a more accurate and nuanced understanding of the emotional content. This innovative blend of CNN and LSTM architectures paved the way for more sophisticated and effective speech emotion recognition models, revolutionizing the field's landscape.

The research [20] detailed a SER-focused integrated deep neural network model. Developed using advanced multi-task learning, it sets new SER standards with remarkable results on the renowned IEMOCAP dataset. It efficiently utilizes pretrained wav2vec-2.0 for speech feature extraction, refined on SER data. This model serves a dual purpose: emotion identification and automatic speech recognition, also producing valuable speech transcriptions. The research [21] investigated the nuances of implementing dilation/stride in 2D dilated convolution. It presents a method for the efficient execution of the inference section, free from constraints on input size, filter size, dilation factor, or stride parameters. This approach is built on a versatile 2D convolution architecture and reimagines 2D-dilated convolution through strategic matrix manipulation. Notably, its computational complexity remains constant regardless of dilation factor changes. Additionally, the method seamlessly integrates stride, resulting in a framework proficient in handling both dilation and stride simultaneously. The scholarly investigation [22] orchestrated a fusion of MFCCs and time-domain characteristics, generating an innovative hybrid feature set aimed at amplifying the performance matrix of SER systems. The resultant hybrid features, coined MFCCT, are employed within the architecture of a CNN to create a sophisticated SER model. Notably, this synergistic amalgamation of MFCCT features with the CNN model markedly transcends the effectiveness of standalone MFCCs and time-domain elements across universally acknowledged datasets.

Furthermore, research [23] addressed the challenging task of effectively merging multimodal data due to their inherent differences. While past methods like feature-level and decision-level fusion often missed intricate modal interactions, a new technique named 'multimodal transformer augmented fusion' is introduced. This method combines feature-level with model-level fusion, ensuring a deep exchange of information between various modalities. Central to this model is the fusion module, housing three Cross-Transformer Encoders, which generate multi-modal emotional representations to enhance data integration. Notably, the hybrid approach uses multi-modal features from feature-level fusion and text data to better capture nuances in speech.

The operational efficiency of SER systems often encounters roadblocks owing to the intricate complexity inherent in these systems, the lack of distinctiveness in features, and the intrusion of noise. In an attempt to overcome these hurdles, the research [24] introduced an enhanced acoustic feature set, which is a composite of MFCC, Linear Prediction Cepstral Coefficients (LPCC), Wavelet Packet Transform (WPT), Zero Crossing Rate (ZCR), spectrum centroid, spectral roll-off, spectral kurtosis, Root Mean Square (RMS), pitch, jitter, and shimmer. These collectively serve to magnify the distinctive nature of the features. Further augmenting this proposition is the deployment of a streamlined one-dimensional deep convolutional neural network 1-D DCNN, designed to both reduce computational complexity and effectively encapsulate the long-term dependencies embedded within speech emotion signals. Acoustic parameters, typically embodied in the form of a feature vector, play a pivotal role in determining the salient characteristics of speech. The research [25] unfolded a pioneering SER model adept at simultaneously learning the Mel Spectrogram (MelSpec) and acoustic parameters, thereby harnessing their respective advantages while curbing their potential shortcomings. For the acoustic parameters, the model leverages the Geneva Minimalistic Acoustic Parameter Set (GeMAPS), a comprehensive compilation of 88 parameters acclaimed for their efficacy in SER. The model, as proposed, is a multi-input deep learning architecture comprising a trinity of networks, each catering to a specific function: one dedicated to the processing of MelSpec in image format, another engineered to handle GeMAPS in vector format, and the final one synergizing the outputs from the preceding two to forecast emotions.

In spite of the individualistic models posited by authors within the previously referenced literature for the SER task, the persistent presence of certain limitations and the obstacle of sub-optimal prediction accuracy continue to warrant further exploration and resolution. The ensuing segments of this document thoroughly illuminate the exhaustive process flow of the proposed system, buttressed by detailed empirical results that serve as corroborative evidence.

3 The Proposed SER System

The section explicates the complex nuances inherent within the proposed system, explicitly engineered to discern emotional indications within vocal articulations. The system consists of two primary constituents, each integral to generating a precise interpretation of the speaker's emotional predilection. A thorough operational sequence is illustrated in Fig. 1 portraying the ordered progression of stages involved in the system's deployment. The disparate components of the model synergistically operate to fulfill the goal of detecting emotional cues in speech. The architecture is constructed via the incorporation of MFCC and Mel-spectrogram characteristics, employing diverse deep learning techniques in accordance with their designated objectives. In sum, the proposed model embodies a holistic and robust strategy for auditory emotion identification, demonstrating versatility across a wide spectrum of pragmatic applications. This adaptability enhances the model's operational capacity and positions it as a potent tool within the ever-evolving field of auditory emotion recognition.

3.1 MFCC Feature Encoder

3.1.1 Accelerated MFCC

MFCCs have gained recognition as informative attributes for analyzing speech signals, finding widespread usage in the field of speech recognition. These characteristics are built upon two fundamental principles: cepstral analysis and the Mel scales. With their ability to capture crucial aspects of the speech signal, MFCC features have become a cornerstone of speech recognition systems. The process of extracting MFCC features involves separating them from the recorded speech signals. This separation allows for the isolation of specific acoustic properties that contribute to the discriminative power of the features. By focusing on these distinctive aspects, the MFCC algorithm enhances the accuracy and effectiveness of speech recognition. MFCCs offer a concise yet informative representation of the speech signal, facilitating robust speech recognition. These features serve as vital inputs to classification algorithms and models, enabling accurate identification and understanding of spoken words and phrases.

This study introduces an expedited approach for extracting the MFCC from speech signals. The primary objective is to streamline the process of feature extraction, enabling efficient and accurate analysis of speech data. The proposed approach focuses on optimizing the speed and effectiveness of MFCC feature extraction. By employing innovative techniques and algorithms, the study aims to reduce the computational complexity and time required for extracting MFCC features from speech signals.

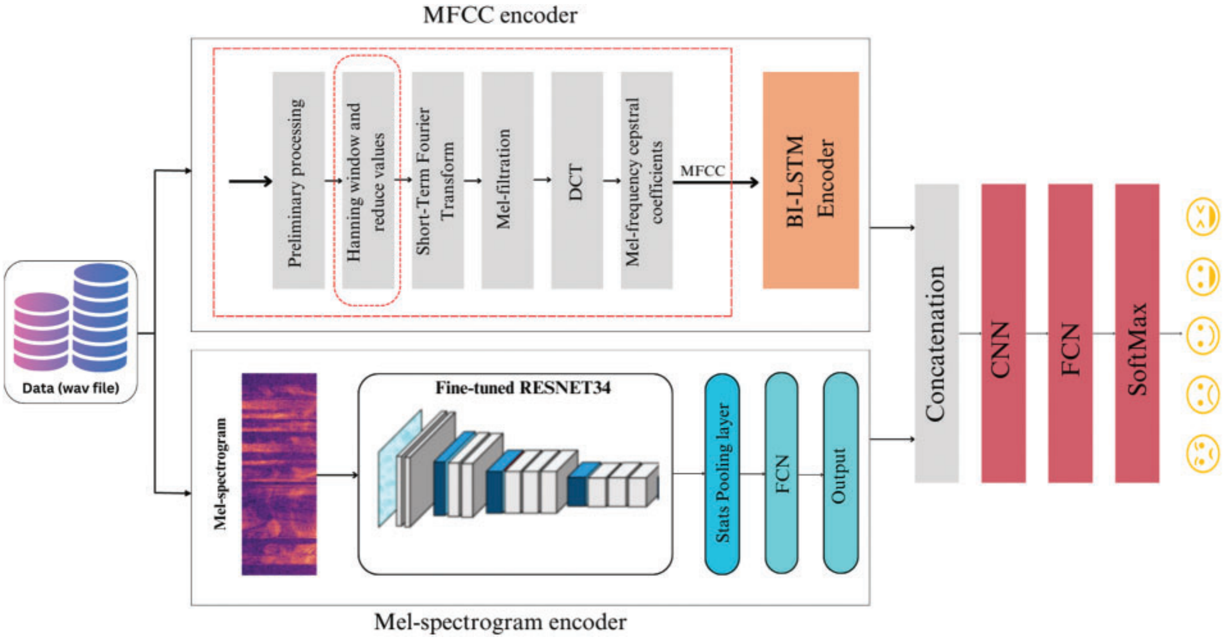

Figure 1: Operational sequence of the proposed system

The research delves into the development of efficient algorithm that leverage parallel processing and optimization strategies to expedite the extraction process. These advancements enable real-time or near real-time extraction of MFCC features, making it more practical for applications requiring swift processing, such as speech recognition, audio classification, and voice-based systems.

We contemplate the operational order of the suggested approach (Fig. 2) for swift derivation of the MFCC features from spoken discourse.

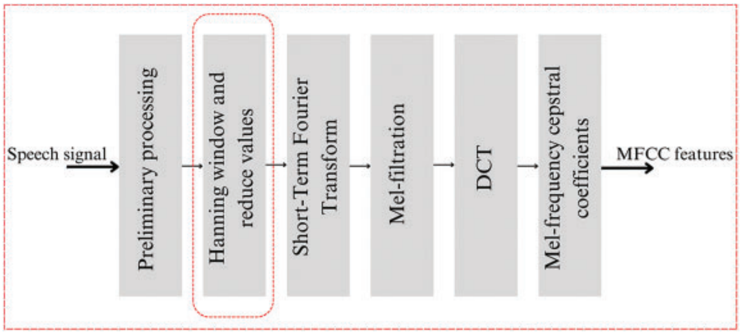
Speech signal
Preliminary processing
Hanning window and reduce values
Short-Term Fourier Transform
Mel-filtration
DCT
Mel-frequency cepstral coefficients
MFCC features

Figure 2: Suggested fast calculation approach of the MFCC

1) *Framing*–after initial filtering, the speech signal undergoes segmentation into frames of 16 milliseconds. Except for the initial frame, each subsequent frame includes the last 10 milliseconds of the frame before it, thereby generating a seamless and overlapping sequence of frames that covers the whole length of the signal. In this particular study, the frame length (N) is set to 256 samples, considering the speech signal's sampling rate of 16 kHz. The offset length (M) is defined as 160 samples, indicating the displacement between consecutive frames. As a result, there is a 62.5% overlap between adjacent frames, which means that each frame shares a significant portion of data with the preceding and subsequent frames. Opting for this degree of overlap guarantees that 50% to 75% of the frame length is covered, conforming to the advised span for the analysis of speech signals. By including this level of overlap, the study aims to capture sufficient temporal information and provide a more comprehensive representation of the speech signal within each frame.
2) *Parallel processing of Hanning window and Reducing values* (Fig. 3). A Hanning window with a size of 1D was employed in this study. The application of the Hanning window aims to curtail disruptions from high-frequency constituents and diminish energy seepage. The structure of the Hanning window permits its side lobes to counterbalance each other, consequently mitigating the effect of high-frequency disturbances on the intended signal. By doing so, it helps to achieve improved spectral resolution and minimize the phenomenon of energy leakage, where signal energy spills into adjacent frequency bins. To minimize distortion and ensure smoother transitions within individual frames, a weight box is employed in the context of this study. This weight box serves the purpose of reducing abrupt changes and promoting a more gradual variation in the signal. The signal under examination in this study is a floating signal that comprises a dense and continuous tone. The strength or amplitude of this unvarying tone is primarily influenced by the amplitude of a pure tone at a distinct frequency, represented as f. This pure tone is subjected to filtering through the Hanning window, which helps shape and modify its characteristics. Incorporating the effects of the window's frequency response, the magnitude of the flat tone is influenced when the Hanning window is applied to the pure tone. This process allows for the manipulation and adjustment of the signal's spectral properties, leading to a more controlled and refined representation of the floating signal. An important feature of this window is that it establishes zero boundaries for the frames. This facilitates the computation of short-term energies as the signal traverses through the window. Subsequently, these energies can be retrieved from the sequence to ascertain the lowest amplitude energies. The aim is to exclude low-energy signals from the total signal by evaluating the signal's energy and simultaneously refining it with this window. The subsequent formula (1) is essential for calculating the signal energy in this procedure:

$$E_n = \sum_{.}^{N} x_i^2 \tag{1}$$

In this equation, E_n represents the energy of the input signal fragment, while x_i denotes the signal value. In the subsequent step, the signal undergoes processing that effectively reduces the quantity of values that are passed into the processor. The window size, which is determined by both the number of samples and the duration, serves as a crucial parameter in the analysis. It is influenced by factors such as the fundamental frequency, intensity, and variations within the signal.

3) Short-Time Fourier Transform (STFT)–the concept of high or low height in relation to the STFT has an intuitive interpretation. The STFT is a transform technique closely associated with the Fourier transform. It is employed to analyze the frequency and phase characteristics

of localized portions of a signal as they vary over time. In practical terms, the computation of the STFT involves dividing a longer time signal into shorter segments of equal length. Each segment is then individually subjected to Fourier transformation, thereby revealing the Fourier spectrum of each segment. In practical applications, discrete-time signals are commonly used. The corresponding conversion from time domain to frequency domain is achieved through a discrete Fourier transformation, where the length of the signal X_n represents the complex value frequency domain of N coefficients. The STFT is a widely utilized tool in speech analysis and processing, as it captures the time-varying evolution of frequency components. One notable advantage of the STFT, similar to the spectrum itself, is that its parameters have meaningful and intuitive interpretations. To visualize the STFT, it is often represented using the logarithmic spectra $20 \log 10\,(\mathrm{X}\,(\mathrm{h}, \mathrm{j}))$. These 2D log spectra can then be displayed as a spectrogram using a color map known as a thermal map. During the third stage of the algorithm, the frames that have undergone the weight windowing process are subjected to the STFT spectral switching procedure. This involves opening the windows and calculating the Discrete Fourier Transform (DFT) of each window, resulting in the STFT of the signal. The transformation of the X_n and W_n windows of the input signal can be defined as follows:

$$STFT = \{x_n\}\,(h, k) = X\,(h, k) = \sum_{n=0}^{N-1} x_{n+h} w_n e^{-i2\pi \frac{kn}{N}} \tag{2}$$

where the k-index represents the frequency values, x_n is the signal window, w_n is the window function and N is the total number of samples in the window.

4) Mel-Filterbank. In the fourth phase, the signal, now transformed to the frequency spectrum, is divided into segments using triangular filters, the boundaries of which are determined by the Mel frequency scale. The transition to the Mel frequency scale is guided by the following formula:

$$M\,(f) = 1127 \times \ln\left(1 + \frac{f}{700}\right) \tag{3}$$

where f—represents the frequency band.

3.1.2 Bi-LSTM Encoder

The processing pipeline for the MFCC sequence is meticulously crafted to harness the power of recurrent neural networks. We employ a bidirectional LSTM layer, imbued with a dropout rate of 0.5, to capture and encode the intricate temporal dependencies within the MFCC sequence. This bidirectional nature enables the model to effectively leverage both past and future context, ensuring a holistic understanding of the speech data. Furthermore, to counteract the risk of overfitting and enhance model generalization, a dropout rate of 0.1 is introduced in a subsequent linear layer, leveraging the rectified linear unit (ReLU) activation function to facilitate non-linear transformations and foster expressive feature representations. This thoughtful design seamlessly integrates regularization techniques, underscoring our commitment to achieving robust and reliable SER performance. The output from the Bi-LSTM encoder is subsequently amalgamated with the resultant output from the Mel-spectrogram feature encoder. This fusion ensures a comprehensive representation of the data by combining temporal sequence learning from Bi-LSTM with the frequency-based understanding from the Mel-spectrogram feature encoder. It optimizes the system's learning capability by exploiting the complementary information inherent in these two distinct yet interrelated sources.

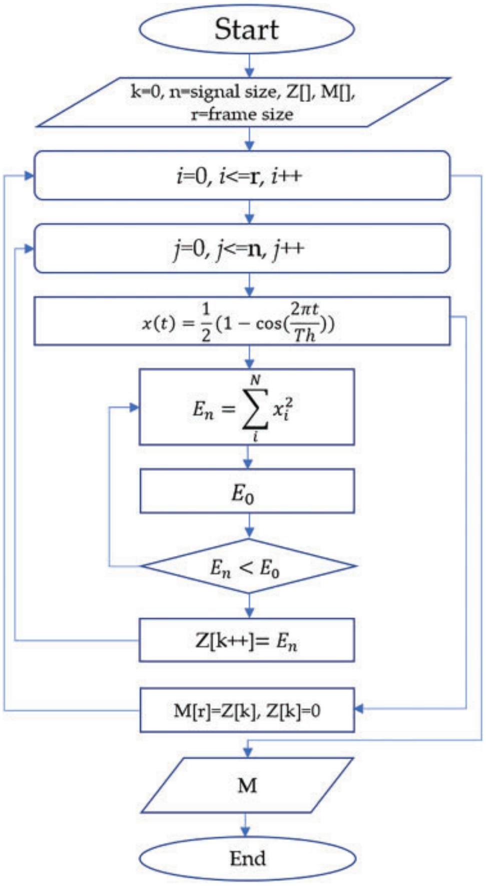

Figure 3: Parallel processing of Hanning window and Reducing values

3.2 Mel-Spectrogram Feature Encoder

Conventional SER methodologies have conventionally relied upon an extensive repertoire of low-level time- and frequency-domain features to effectively capture and represent the multifaceted tapestry of emotions conveyed through speech. However, recent advancements have witnessed a paradigm shift towards cutting-edge SER systems that harness the formidable prowess of complex neural network architectures, enabling direct learning from spectrograms or even raw waveforms. In the pursuit of furthering this research frontier, our study capitalizes on the ResNet [26] architecture, skillfully employing mel-spectrograms as input features. By extracting high-resolution spectrograms, our model adeptly encodes the subtle intricacies of the spectral envelope and the coarse harmonic structures, deftly unraveling the rich tapestry of emotions permeating speech signals. Through this sophisticated framework (Fig. 4), our model transcends the limitations of traditional feature-based approaches, culminating in an elevated degree of accuracy and efficacy in the discernment and recognition of emotions in speech data.

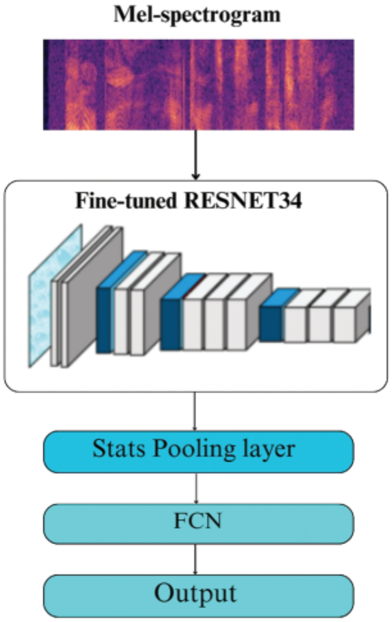

Figure 4: Mel-spectrogram feature encoder

Our Mel-spectrogram feature encoder commences with a meticulous two-step process. Firstly, we subject a ResNet model to pre-training on the expansive Librispeech dataset [27]. This initial phase endows the model with a comprehensive foundation of knowledge, enabling it to glean essential insights into the underlying speech representations. Subsequently, we strategically replace the fully connected (FC) layers of the pre-trained model with Stats Pooling and FC layers. This deliberate replacement serves to prime the model for the precise task at hand-SER-employing the CMU-MOSEI and RAVDESS datasets as our experimental bedrock. To holistically capture the intricate temporal dynamics and contextual information pervasive within the speech data, our proposed system leverages the indispensable statistics pooling layer [28]. Serving as a pivotal component within the architecture, this layer adeptly assimilates frame-level information over temporal sequences. Through the astute concatenation of the mean and standard deviation computed over the frames, it ingeniously distills the sequence of frames into a single, compact vector representation. This judiciously crafted vector encapsulates vital statistical information that encapsulates the nuanced emotional content ingrained within the speech signal. Notably, our system operates across distinct levels of granularity, intelligently harnessing the disparate capabilities of the ResNet model's components. The convolutional layers, meticulously designed to extract salient features, operate at the frame level, diligently capturing local patterns and structures that underpin the speech signal's intrinsic characteristics. In a complementary fashion, the FC layers assume the role of segment-level interpreters, harmoniously synthesizing the accumulated frame-level information within a given segment of speech. This segment-based perspective engenders a holistic grasp of the temporal dynamics while facilitating a comprehensive interpretation of higher-level emotional patterns. As a result, our SER system manifests heightened discrimination capabilities, adroitly striking a balance between fine-grained temporal dynamics and the discernment of overarching emotional patterns.

In order to capture a rich and detailed representation of the audio signals, we employ a meticulous procedure for extracting high-resolution log-mel spectrograms. These spectrograms are meticulously engineered to possess a dimensionality of 128, allowing for a comprehensive encoding of the acoustic features embedded within the speech signals. The extraction process entails a frame-based approach, where each frame spans a duration of 25 ms and is sampled at a rate of 100 Hz, effectively capturing temporal dynamics with fine-grained precision at intervals of 10 ms. To ensure appropriate feature normalization, we employ a segment-level mean and variance normalization technique. However, we acknowledge that this normalization approach falls short of the optimal scenario where normalization is conducted at the recording or conversation level. In light of this limitation, we recognize the value of a more holistic normalization strategy that takes into account the broader context of the recording or conversation. By considering statistics computed at the conversation level, such as mean and variance, we can effectively capture the inherent variations and nuances within the conversation. Significantly, our meticulous experimentation has unveiled a compelling finding: normalizing the segments using conversation-level statistics yields substantial enhancements in the performance of the SER system, particularly when applied to the CMU-MOSEI and RAVDESS datasets. This empirical observation underscores the criticality of incorporating context-aware normalization techniques in order to effectively capture the subtle emotional cues embedded within real-world conversational scenarios and elevate the overall accuracy of SER systems.

The ResNet model architecture employed in our study incorporates a first block comprising 32 channels, signifying the number of parallel convolutional filters utilized. To optimize the training process, we leverage the stochastic gradient descent (SGD) optimizer with a momentum value of 0.9, ensuring efficient convergence towards an optimal solution. Concurrently, a batch size of 32 is employed, allowing for parallel processing and expedited training. For the purpose of fine-tuning the convolutional layers, we adopt a learning rate of 0.001, enabling precise adjustments to the network parameters during this critical phase. Notably, the learning rate strategy is intelligently modulated, remaining constant for the initial 10 epochs to establish a stable training foundation. Subsequently, for each subsequent epoch, the learning rate is halved, facilitating finer parameter updates as the training progresses. To introduce the crucial element of non-linearity and enhance the model's expressive capabilities, ReLU activation functions are applied across all layers, excluding the output layer. This choice of activation function enables the ResNet model to effectively capture complex patterns and salient features, facilitating the extraction of meaningful representations. In order to expedite the training process and bolster the model's generalization properties, we integrate layer-wise batch normalization. This technique normalizes the inputs to each layer, ensuring consistent distribution and alleviating internal covariate shift, thereby accelerating model convergence and enhancing its ability to generalize to unseen data.

4 Experiments and Discussion

4.1 Datasets

4.1.1 The CMU-MOSEI

The CMU-MOSEI [29] dataset stands as an expansive multimodal corpus tailored for emotion analysis in the realm of conversational video data. Spanning an impressive collection of over 23,000 video clips sourced from 1,000 distinct sessions involving a diverse cohort of over 1,200 participants, this dataset offers an unparalleled resource for exploring emotion dynamics within the conversational context. Complementing the video data, it encompasses an array of auxiliary features including speech transcripts, audio features, visual features, and annotated labels that denote the valence and arousal levels associated with each video clip. The CMU-MOSEI dataset encompasses six distinct emotion classes (Table 1): anger, happiness, sadness, disgust, fear, and surprise, characterized by varying sample distributions: angry (4,600 samples), sadness (5,601 samples), disgust (3,755 samples), surprise (2,055 samples), happy (10,752 samples), and fear (1,803 samples).

Table 1: RAVDESS and the CMU-MOSEI dataset content

	RAVDESS		The CMU-MOSEI	
Emotion	Number of samples	Percentage (%)	Number of samples	Percentage (%)
Angry	192	13.3	4600	16
Sadness	192	13.3	5601	20
Disgust	192	13.3	3755	13
Fear	192	13.3	1803	6
Happy	192	13.3	10752	38
Surprise	192	13.3	2055	7
Calm	192	13.3	–	–
Neutral	96	6.9	–	–

4.1.2 RAVDESS

The RAVDESS [30] dataset stands as a notable collection designed specifically for cutting-edge research pursuits in emotion recognition. Distinguished by its accessibility and public availability, the RAVDESS dataset encompasses a rich reservoir of recordings that encompasses both emotional speech and song, rendering it an invaluable resource within the scientific community. A notable facet of the dataset lies in its diverse repertoire of emotional expressions, spanning a spectrum that encompasses neutral, calm, happy, sad, angry, fearful, surprised, and disgusted states, each meticulously articulated in both English and French languages. Actors were chosen to deliver monologues consisting of 13 sentences, both statements and questions, each conveying specific emotional tones.

The data collection process is marked by meticulousness, with a cohort of 24 highly skilled individuals, equally balanced between males and females, actively participating. This gender parity ensures an equitable representation, fostering a holistic understanding of emotional expressions across diverse demographics. The speech dataset has 1440 files, derived from 60 trials for each of the 24 actors. The dataset is in WAV format with a 16-bit bitrate and a 48 kHz sampling rate. In Table 1 of the RAVDESS dataset, while each emotion is represented by 192 samples, the "Neutral" emotion has only 96 samples.

4.2 Implementation Details

To fairly evaluate our model on the CMU-MOSEI and RAVDESS datasets, we strictly trained it using the method described in [31]. We split the data into 80% for training and 20% for testing. This allows the model to train on a large part of the data and still have a significant portion for validation, ensuring a complete evaluation of its performance. We obtained 1152 training samples and 288 testing samples from the RAVDESS dataset through this process. In a similar vein, for the CMU-MOSEI dataset, we assigned 22852 training instances and 5714 testing instances. Unlike the method in [32], we did not use 10-fold cross-validation. This was due to the challenges and resources needed to apply cross-validation to deep learning models.

In speech emotion recognition, precision, recall, and accuracy are common metrics to evaluate model performance. They are widely accepted in the SER field because they offer a well-rounded view of how a model performs, considering both its relevance and comprehensiveness.

Our system uses TensorFlow with Python, a popular open-source machine learning platform. We chose the Adam optimizer with a learning rate of 0.0001 and used 'categorical_crossentropy' for the loss function. We also applied L2 regularization for better efficiency and reliable convergence.

We trained our model for 200 epochs using batches of 32 on a system with an Nvidia GeForce GTX 1660 Ti 16 GB GPU and an Intel Core i7-1265UE 10-Core CPU. Running on Windows 11 with 64 GB RAM, this setup allowed us to efficiently carry out deep learning tasks during both training and testing.

4.3 Recognition Results

Tables 2 and 3 represent a detailed breakdown of the performance of the proposed emotion recognition system for each class of emotion. The classes include "Angry", "Sadness", "Disgust", "Fear", "Happy", "Surprise", "Calm", and "Neutral". The performance measures used are precision, recall, and F1 score.

Table 2: The system's recognition performance on different emotions of the RAVDESS dataset

Emotion	Angry	Sadness	Disgust	Fear	Happy	Surprise	Calm	Neutral
Precision	95.9	94.4	93.7	94.1	96.2	96.6	97.0	89.9
Recall	97.5	95.6	95.1	93.7	95.9	97.3	96.1	93.8
F1	96.7	95.0	94.4	93.9	96.0	96.9	96.5	91.8

Table 3: The system's recognition performance on different emotions of the CMU-MOSEI dataset

Emotion	Angry	Sadness	Disgust	Fear	Happy	Surprise
Precision	92.6	94.9	92.0	91.7	98.2	88.5
Recall	91.2	95.3	90.8	92.5	96.8	89.7
F1	91.8	95.0	91.3	92.0	97.7	89.0

From the Table 2, we can infer that the model has a relatively high precision, recall, and F1 score for all the emotion classes of the RAVDESS dataset, indicating that it is performing well in recognizing different emotions from the speech data. The lowest precision is for the "Neutral" class at 89.9%, but this is still quite high. The lowest recall is for the "Fear" class at 93.7%, but again, this is relatively high. The F1 scores also indicate that there's a good balance between precision and recall across all classes, with the lowest F1 score being 91.8% for the "Neutral" class. For example, in the "Surprise" class, the model correctly predicted anger with a precision of 96.6% of the times. It identified 97.3% of all actual instances of surprise (recall). The F1 score of 96.9% suggests a good balance between precision and recall.

Overall, these numbers suggest the model performs well across different emotion categories, successfully recognizing each type of emotion from the given speech data.

In the case of the CMU-MOSEI dataset, we evaluated the performance of our emotion recognition model across six different emotions: Angry, Sadness, Disgust, Fear, Happy, and Surprise. The results, as outlined in Table 3, show strong performance across all tested emotions. Our model achieved exceptional precision, particularly for 'Happy' emotion, which scored 98.2%. Similarly, recall was highest for 'Happy' at 96.8%, indicating the model's robustness in identifying instances of this emotion. The model demonstrated balanced performance in the 'Sadness' category, with both precision and recall scoring around 95.0%. Furthermore, the 'Happy' emotion resulted in the highest F1 score (97.7%), denoting an excellent harmony between precision and recall.

However, our analysis also highlighted some areas for potential improvement. Both 'Surprise' and 'Disgust' had relatively lower precision, recall, and F1 scores compared to other emotions, which suggests room for further optimization. This robust evaluation provides important insights into our model's performance, emphasizing its strengths and revealing potential areas for future enhancements. These results are an encouraging step forward for the development of more effective and accurate emotion recognition models.

The evaluation process was expanded by employing a confusion matrix, as shown in Tables 4 and 5. These tables supplied a visual interpretation and detailed explanation of how the model performed. It illustrated that the model surpassed a 92% accuracy rate on the RAVDESS dataset and 90% on the CMU-MOSEI dataset for each unique emotion class. These results suggest a high level of precision in the classification tasks, indicating the model's sturdy and trustworthy ability to categorize emotions.

Table 4: The confusion matrix on the RAVDESS dataset

	Angry	Sadness	Disgust	Fear	Happy	Surprise	Calm	Neutral
Angry	0.97	0.00	0.01	0.01	0.00	0.00	0.00	0.00
Sadness	0.00	0.95	0.00	0.03	0.00	0.00	0.01	0.00
Disgust	0.00	0.01	0.94	0.00	0.00	0.00	0.01	0.02
Fear	0.03	0.00	0.01	0.94	0.00	0.00	0.0	0.00
Happy	0.00	0.00	0.00	0.00	0.96	0.03	0.00	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.01
Calm	0.00	0.01	0.00	0.00	0.00	0.00	0.96	0.02
Neutral	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.92

Table 5: The confusion matrix on the CMU-MOSEI dataset

	Angry	Sadness	Disgust	Fear	Happy	Surprise
Angry	0.92	0.00	0.04	0.01	0.00	0.00
Sadness	0.02	0.95	0.00	0.01	0.00	0.00
Disgust	0.00	0.04	0.92	0.00	0.00	0.00
Fear	0.03	0.00	0.02	0.91	0.00	0.00
Happy	0.00	0.00	0.00	0.00	0.98	0.00
Surprise	0.00	0.00	0.00	0.00	0.05	0.90

The following Table 6 presents a comparative analysis of the proposed SER system against benchmark methods applied to two distinct datasets: RAVDESS and CMU-MOSEI.

Table 6: Comparison with the benchmark SER methods

Datasets	SER methods	Accuracy (%)
RAVDESS	Bhangale et al. [24]	94.2
	Ephrem et al. [33]	82.71
	Pulatov et al. [34]	94.8
	UA et al. [35]	89.0
	The proposed system	95.3
CMU-MOSEI	Mittal et al. [36]	89.0
	Xia et al. [37]	88.2
	Jing et al. [38]	87.5
	Lian et al. [39]	86.82
	Fang et al. [40]	85.40
	The proposed system	93.2

For the RAVDESS dataset, we compare four SER methods, namely Bhangale et al. [24], Ephrem et al. [33], Pulatov et al. [34], and UA et al. [35]. The proposed system is also included for reference. Notably, the proposed system exhibits the highest accuracy of 95.3%, outperforming the other methods.

On the CMU-MOSEI dataset, we assess the performance of three SER methods, specifically Mittal et al. [36], Xia et al. [37], and Jing et al. [38], along with the proposed system. Here, the proposed system achieves an accuracy of 93.2%, which surpasses the results obtained by the other methods.

The findings from this comparative analysis demonstrate the efficacy of the proposed system in both datasets, RAVDESS and CMU-MOSEI, showcasing its robustness in recognizing emotions from speech signals. It is important to consider the limitations and biases in each dataset, as they can impact the performance of SER methods. Careful evaluation and validation on diverse datasets are essential for developing robust emotion recognition models.

4.4 Discussion and Limitations

The results of our study demonstrate that the speech emotion recognition system we developed outperforms existing models in terms of detection accuracy. By leveraging high-resolution Mel-spectrograms for feature extraction and swiftly computing MFCCs, our system streamlines the emotion recognition process. This innovation effectively tackles the pressing challenge of feature extraction, a crucial component that significantly impacts the effectiveness of SER systems.

Upon evaluating our proposed model, we garnered several compelling insights. The system showcased an impressive level of accuracy, as evidenced by the data in the accompanying tables. Across multiple emotion categories, our model consistently achieved Precision, Recall, and F1 scores exceeding 90% for the majority of them. These metrics highlight the model's robustness and skill in emotion classification tasks.

The categories of ‘Happy’ emotions stood out, with exceptional Precision and Recall rates of 98.2% and 96.8%, respectively. This underlines the model’s proficiency in correctly identifying and categorizing instances of happiness. Conversely, the categories for ‘Surprise’ and ‘Disgust’ showed slightly weaker performance metrics, suggesting that the model may face challenges in accurately categorizing these specific emotions.

These variations could be attributed to the inherently complex and subjective nature of human emotions, which can differ significantly across individual experiences and settings. Nonetheless, the overall high performance metrics affirm the model’s potential and efficacy.

While our findings are encouraging, there are several limitations to consider. Our study primarily utilized the CMU-MOSEI and RAVDESS databases, which mainly consist of acted emotions that may not fully represent the nuances of spontaneous emotional expressions. For future research, extending to databases that capture more naturalistic emotional behavior would be beneficial. Additionally, our system is optimized for English and may display varying performance levels in different linguistic and cultural contexts. Future work should aim to improve the system’s linguistic and cultural adaptability. Moreover, our current model is audio-focused, overlooking the potential benefits of integrating visual or textual cues. Investigating multi-modal systems could offer a more holistic approach to emotion recognition in future studies.

Finally, it is worth noting that the performance of our system can be influenced by various factors like background noise, distance of the speaker from the microphone, and other environmental elements. To enhance robustness, future iterations of this model could incorporate noise reduction techniques and additional preprocessing measures to maintain high recognition accuracy under diverse recording conditions.

5 Conclusion

The primary focus of our study was to design a sophisticated SER system that makes the most of Mel-spectrograms for intricate spectrogram extraction, coupled with the swift computation capabilities of MFCC, making the feature extraction phase both efficient and effective. At the heart of our effort was a commitment to go beyond existing benchmarks. We sought to address and overcome the limitations of current techniques, driven by an unwavering commitment to heightened accuracy in emotion recognition. In order to validate our advancements, we subjected our proposed system to rigorous evaluations using two distinct databases: The CMU-MOSEI and RAVDESS. The ensuing results not only met our expectations but in many respects, exceeded them. The system showcased its mettle by recording an accuracy rate of 93.2% on the CMU-MOSEI dataset and an even more commendable 95.3% on the RAVDESS dataset.

Our findings in this research signify more than just technical advancements; they herald a new era in speech recognition systems. The insights we have garnered underscore several compelling avenues that warrant deeper investigation. To elaborate, we are currently delving into the fusion of our established model with the nuanced mechanisms of transformer architectures and self-attention. Additionally, there is a concerted effort underway to harness the power of pretrained audio models. Our overarching aim remains clear: to sift through speech and extract features that are not only abundant but also meaningful, thereby elevating the finesse of emotion detection. Recognizing the evolving landscape of spoken language and its myriad emotional undertones, we are also directing our energies towards assimilating a broader emotional speech database. Such a move is anticipated to fortify our model’s adaptability, ensuring it remains robust when faced with a spectrum of emotional expressions and intricate speech variations. By doing so, we aim to make our model not just technically adept but also practically invaluable in diverse real-world scenarios.

In closing, the advancements birthed from this research project hold profound potential. The ripple effects of our work are anticipated to be felt far and wide, from making machine-human interactions more intuitive and genuine to refining therapeutic speech interventions and offering sharper mental health evaluations. We stand at the cusp of a transformative era, and our work seeks to be a beacon, lighting the way for future explorations and innovations that have the power to enrich and reshape the tapestry of our daily interactions and experiences.

Acknowledgement: The authors A.A., A.K. and R.N. would like to express their sincere gratitude and appreciation to the supervisor, Taeg Keun Whangbo (Gachon University) for his support, comments, remarks, and engagement over the period in which this manuscript was written. Moreover, the authors would like to thank the editor and anonymous referees for the constructive comments in improving the contents and presentation of this paper.

Funding Statement: This work was supported by the GRRC program of Gyeonggi Province (GRRC-Gachon2023(B02), Development of AI-based medical service technology).

Author Contributions: A.A. developed the method; A.A., A.K. and R.N. performed the experiments and analysis; A.A. wrote the paper; T.K.W. supervised the study and contributed to the analysis and discussion of the algorithm and experimental results. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All datasets are available publicly at the website www.kaggle.com.

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

References

[1] S. Langari, H. Marvi and M. Zahedi, "Efficient speech emotion recognition using modified feature extraction," *Informatics in Medicine Unlocked*, vol. 20, pp. 100424, 2020. https://doi.org/10.1016/j.imu.2020.100424

[2] H. Zhao, L. Li, X. Zha, Y. Wang, Z. Xie *et al.,* "ACG-EmoCluster: A novel framework to capture spatial and temporal information from emotional speech enhanced by DeepCluster," *Sensors*, vol. 23, pp. 4777, 2023. https://doi.org/10.3390/s23104777

[3] L. Yuan, G. Huang, F. Li, X. Yuan, C. M. Pun *et al.,* "RBA-GCN: Relational bilevel aggregation graph convolutional network for emotion recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023. https://doi.org/10.1109/TASLP.2023.3284509

[4] C. Deepika and S. Kuchibhotla, "Feature extraction model for speech emotion detection with prodigious precedence assortment model using fuzzy-based convolution neural networks," *Soft Computing*, 2023. https://doi.org/10.1007/s00500-023-08458-5

[5] B. Tripathi and R. K. Sharma, "EEG-based emotion classification in financial trading using deep learning: Effects of risk control measures," *Sensors*, vol. 23, pp. 3474, 2023. https://doi.org/10.3390/s23073474

[6] N. Azam, T. Ahmad and N. Ul Haq, "Automatic emotion recognition in healthcare data using supervised machine learning," *PeerJ-Computer Science*, vol. 7, pp. e751, 2021.

[7] X. L. He and Z. H. Liu, "Design of blended teaching model based on emotion recognition and language learning," *Frontiers in Psychology.*, vol. 13, pp. 917517, 2022. https://doi.org/10.3389/fpsyg.2022.917517

[8] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 917517, pp. 221640–221653, 2020. https://doi.org/10.1109/ACCESS.2020.3043201

[9] M. Kotha and E. Logashanmugam, "Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm," *Knowledge-Based Systems*, vol. 246, pp. 108659, 2022. https://doi.org/10.1016/j.knosys.2022.108659

[10] F. Makhmudov, A. Kutlimuratov, F. Akhmedov, M. S. Abdallah and Y. I. Cho, "Modeling speech emotion recognition via attention-oriented parallel CNN encoders," *Electronics*, vol. 11, pp. 4047, 2022. https://doi.org/10.3390/electronics11234047

[11] M. Sharafi, M. Yazdchi, R. Rasti and F. Nasimi, "A novel spatio-temporal convolutional neural framework for multimodal emotion recognition," *Biomedical Signal Processing and Control*, vol. 78, pp. 103970, 2022. https://doi.org/10.1016/j.bspc.2022.103970

[12] J. Singh, L. B. Saheer and O. Faust, "Speech emotion recognition using attention model," *International Journal of Environmental Research and Public Health*, vol. 20, pp. 5140, 2023. https://doi.org/10.3390/ijerph20065140

[13] T. Liu and X. Yuan, "Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, pp. 23, 2023. https://doi.org/10.1186/s13636-023-00290-x

[14] M. T. Yan, F. Yuan and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125 868–125 881, 2019.

[15] A. Farkhod, A. B. Abdusalomov, M. Mukhiddinov and Y. I. Cho, "Development of real-time landmark-based emotion recognition CNN for masked faces," *Sensors*, vol. 22, pp. 8704, 2022. https://doi.org/10.3390/s22228704

[16] B. Mocanu, R. Tapu and T. Zaharia, "Utterance level feature aggregation with deep metric learning for speech emotion recognition," *Sensors*, vol. 21, pp. 4233, 2021. https://doi.org/10.3390/s21124233

[17] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov and T. K. Whangbo, "Multimodal emotion detection via attention-based fusion of extracted facial and speech features," *Sensors*, vol. 23, pp. 5475, 2023. https://doi.org/10.3390/s23125475

[18] J. Li, X. Zhang, L. Huang, F. Li, S. Duan *et al.,* "Speech emotion recognition using a dual-channel complementary spectrogram and the CNN-SSAE neutral network," *Applied Sciences*, vol. 12, pp. 9518, 2022. https://doi.org/10.3390/app12199518

[19] H. Zhang, H. Huang and H. Han, "A novel heterogeneous parallel convolution Bi-LSTM for speech emotion recognition," *Applied Sciences*, vol. 11, pp. 9897, 2021.

[20] X. Cai, J. Yuan, R. Zheng, L. Huang and K. Church, "Speech emotion recognition with multi-task learning," in *Proc. of the Interspeech*, Brno, Czechia, 2021.

[21] V. H. E. and S. Ghanekar, "An efficient method for generic DSP implementation of dilated convolution," in *ICASSP 2022–2022 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, pp. 51–55, 2022. https://doi.org/10.1109/ICASSP43922.2022.9747844

[22] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman and O. S. Neffati, "Speech emotion recognition through hybrid features and convolutional neural network," *Applied Sciences*, vol. 13, pp. 4750, 2023. https://doi.org/10.3390/app13084750

[23] Y. Wang, Y. Gu, Y. Yin, Y. Han, H. Zhang *et al.,* "Multimodal transformer augmented fusion for speech emotion recognition," *Frontiers in Neurorobotics*, vol. 17, pp. 1662–5218, 2023. https://doi.org/10.3389/fnbot.2023.1181598

[24] K. Bhangale and M. Kothandaraman, "Speech emotion recognition based on multiple acoustic features and deep convolutional neural network," *Electronics*, vol. 12, pp. 839, 2023. https://doi.org/10.3390/electronics12040839

[25] I. Toyoshima, Y. Okada, M. Ishimaru, R. Uchiyama and M. Tada, "Multi-input speech emotion recognition model using mel spectrogram and GeMAPS," *Sensors*, vol. 23, pp. 1743, 2023. https://doi.org/10.3390/s23031743

[26] He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 770–778, 2016.

[27] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, pp. 5206–5210, 2015. https://doi.org/10.1109/ICASSP.2015.7178964

[28] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in *Proc. of IEEE ICASSP*, Calgary, AB, Canada, pp. 5329–5333, 2018.

[29] A. Zadeh and P. Pu, “Multimodal language analysis in the wild: CMU-mosei dataset and interpretable dynamic fusion graph,” in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, VIC, pp. 2236–2246, 2018.

[30] S. R. Livingstone and 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*, vol. 13, no. 5, pp. e0196391, 2018.

[31] W. Dai, S. Cahyawijaya, Z. Liu and P. Fung, “Multimodal end-to-end sparse model for emotion recognition,” in *Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Association for Computational Linguistics, pp. 5305–5316, 2021.

[32] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh *et al.,* “Multimodal sentiment analysis: Addressing key issues and setting up the baselines,” *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.

[33] E. A. Retta, R. Sutcliffe, J. Mahmood, M. A. Berwo, E. Almekhlafi *et al.,* “Cross-corpus multilingual speech emotion recognition: Amharic vs. other languages,” 2023. https://doi.org/10.48550/arXiv.2307.10814

[34] I. Pulatov, R. Oteniyazov, F. Makhmudov and Y. I. Cho, “Enhancing speech emotion recognition using dual feature extraction encoders,” *Sensors*, vol. 23, pp. 6640, 2023. https://doi.org/10.3390/s23146640

[35] A. A. U. and K. K. V., “Speech emotion recognition-a deep learning approach,” in *Fifth Int. Conf. on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Palladam, India, pp. 867–871, 2021. https://doi.org/10.1109/I-SMAC52330.2021.9640995

[36] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera and D. Manocha, “M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues,” in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 34, no. 2, pp. 1359–1367, 2020. https://doi.org/10.1609/aaai.v34i02.5492

[37] X. Liu, Z. Xu and K. Huang, “Multimodal emotion recognition based on cascaded multichannel and hierarchical fusion,” *Computational Intelligence and Neuroscience*, vol. 2023, pp. 9645611, 2023. https://doi.org/10.1155/2023/9645611

[38] J. He, H. Yanga, C. Zhang, H. Chen and Y. Xua, “Dynamic invariant-specific representation fusion network for multimodal sentiment analysis,” *Computational Intelligence and Neuroscience*, vol. 2022, pp. 2105593, 2022. https://doi.org/10.1155/2022/2105593

[39] Z. Lian, B. Liu and J. Tao, “SMIN: Semi-supervised multi-modal interaction network for conversational emotion recognition,” *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2415–2429, 2023. https://doi.org/10.1109/TAFFC.2022.3141237

[40] L. Fang, G. Liu and R. Zhang, “Sense-aware BERT and multi-task fine-tuning for multimodal sentiment analysis,” in *2022 Int. Joint Conf. on Neural Networks (IJCNN)*, Padua, Italy, pp. 1–8, 2022. https://doi.org/10.1109/IJCNN55064.2022.9892116