IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 September 2023, accepted 28 September 2023, date of publication 2 October 2023, date of current version 9 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3321122

RESEARCH ARTICLE

Mel-MViTv2: Enhanced Speech Emotion Recognition With Mel Spectrogram and Improved Multiscale Vision Transformers

KAH LIANG ONG[1], CHIN POO LEE[1], (Senior Member, IEEE), HENG SIONG LIM[2], (Senior Member, IEEE), KIAN MING LIM[1], (Senior Member, IEEE), AND ALI ALQAHTANI[3,4]

[1]Faculty of Information Science and Technology, Multimedia University, Malacca 75450, Malaysia
[2]Faculty of Engineering and Technology, Multimedia University, Malacca 75450, Malaysia
[3]Department of Computer Science, King Khalid University, Abha 61421, Saudi Arabia
[4]Center for Artificial Intelligence (CAI), King Khalid University, Abha 61421, Saudi Arabia

Corresponding author: Chin Poo Lee (cplee@mmu.edu.my)

This work was supported in part by Telekom Malaysia Berhad's Research and Development under Grant RDTC/231075 and Grant RDTC/221064; and in part by the Deanship of Scientific Research, King Khalid University, Saudi Arabia, under Grant RGP2/332/44.

ABSTRACT Speech emotion recognition aims to automatically identify and classify emotions from speech signals. It plays a crucial role in various applications such as human-computer interaction, affective computing, and social robotics. Over the years, researchers have proposed different approaches for speech emotion recognition, leveraging various classifiers and features. However, despite the advancements, existing methods in speech emotion recognition still have certain limitations. Some approaches rely on handcrafted features that may not capture the full complexity of emotional information present in speech signals, while others may suffer from a lack of robustness and generalization when applied to different datasets. To address these challenges, this paper proposes a speech emotion recognition method that combines Mel spectrogram with Short-Term Fourier Transform (Mel-STFT) and the Improved Multiscale Vision Transformers (MViTv2). The Mel-STFT spectrograms capture both the frequency and temporal information of speech signals, providing a more comprehensive representation of the emotional content. The MViTv2 classifier introduces multi-scale visual modeling with different stages and pooling attention mechanisms. MViTv2 incorporates relative positional embeddings and a residual pooling connection to effectively model the interactions between tokens in the space-time structure, preserve essential information, and improve the efficiency of the model. Experimental results demonstrate that the proposed method generalizes well on different datasets, achieving an accuracy of 91.51% on the Emo-DB dataset, 81.75% on the RAVDESS dataset, and 64.03% on the IEMOCAP dataset.

INDEX TERMS Speech, speech emotion, speech emotion recognition, spectrogram, mel spectrogram, mel spectrogram with short-time Fourier transform, vision transformer, improved multiscale vision transformers, Emo-DB, RAVDESS, IEMOCAP.

I. INTRODUCTION

Speech emotion recognition is a significant task within the field of signal processing and machine learning, focused on detecting and analyzing emotional information conveyed through speech signals. Previous research has employed various techniques to extract emotional cues and classify them into discrete emotional states. However, most existing approaches rely on handcrafted time and frequency domain features, which possess potential limitations. These limitations include limited resolution in the time and frequency domains, impacting classification accuracy and discriminative power.

To overcome these limitations, some studies have explored the use of deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno M. Garcia.

(LSTM) networks for speech emotion recognition. CNNs excel at capturing local patterns and spectral information, but struggle with modeling long-term dependencies and sequential dynamics present in speech signals. In contrast, LSTM networks effectively model temporal dependencies but face challenges in capturing intricate spectral characteristics. These potential limitations emphasize the necessity for novel approaches that overcome the shortcomings of handcrafted features and explore sophisticated architectures capable of capturing both spectral and temporal dynamics in speech emotion recognition.

To address this, this paper proposes the ''Mel-MViTv2'' method, which combines the strengths of Mel spectrogram with short-time Fourier transform (Mel-STFT) and Improved Multiscale Vision Transformers (MViTv2). Mel-STFT merges the concepts of short-time Fourier transform (STFT) and Mel-frequency spectrogram to provide a descriptive representation of speech signals. STFT captures the sinusoidal frequency and phase content of local signal sections over time, while the Mel-frequency spectrogram applies a non-linear transform to the frequency axis using the Mel-scale, emphasizing perceptually important frequency ranges.

Additionally, the paper employs Improved Multiscale Vision Transformers (MViTv2) to learn and classify the Mel-STFT representation. MViTv2 incorporates multi-scale visual modeling by utilizing different stages instead of single-scale blocks in the Vision Transformer architecture. The network gradually expands the channel width while reducing the resolution from input to output stages. MViTv2 also incorporates relative positional embeddings to capture relative location distance between input tokens, enhancing shift-invariance properties. Furthermore, it employs residual pooling connections to enhance information flow and aid in training pooling attention blocks while maintaining low-complexity attention computation. This approach allows MViTv2 to extract features from multiple scales and resolutions, capturing fine details and larger context simultaneously, thereby enhancing the accuracy of speech emotion recognition.

The combination of Mel-STFT and MViTv2 demonstrates remarkable performance in speech emotion recognition. Mel spectrograms provide visual representations of speech signals, highlighting frequency content relevant for emotion analysis. MViTv2, with its multiscale feature hierarchies and transformer-based architecture, classifies and recognizes emotional patterns within the Mel spectrograms. The hierarchical nature of the MViTv2 model captures complex relationships between different scales of features, facilitating the identification and understanding of emotional cues in the visual representations of speech. Subsequently, the generalization capabilities of the Mel-MViTv2 method are evaluated on three speech emotion datasets that encompass diverse characteristics such as gender, language, and recording environments. This evaluation assesses how effectively the Mel-MViTv2 method can adapt to variations in these important factors, providing insights into its robustness and applicability across different contexts. The main contributions of this paper are:

- The utilization of Mel-STFT in speech representation provides a descriptive and informative representation that effectively captures the intricate temporal and spectral characteristics of speech signals, thereby enhancing the discriminative power and interpretability of speech analysis in the context of speech emotion recognition. Mel-STFT comprehensively captures both the nuanced temporal dynamics and the perceptually significant frequency components, facilitating precise characterization of emotional content within speech signals.
- The adoption of Improved Multiscale Vision Transformers (MViTv2) for representation learning and classification of Mel-STFT further enhances the capability of the model. MViTv2 excels in extracting hierarchical and multi-scale features, enabling the model to capture fine-grained variations and contextual information in the representation. This multi-scale feature extraction allows the model to capture both fine details and larger context simultaneously, resulting in higher accuracy and improved generalization ability in speech emotion classification.
- The performance evaluation on three diversified speech emotion datasets, namely the Berlin Database of Emotional Speech (Emo-DB), the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) to assess the generalization capabilities of the model.

II. RELATED WORKS

Recognizing emotions from speech signals is a complex task due to the inherent variability in speech patterns and the subjective nature of emotional expression. Researchers have made significant progress in this field by employing various approaches ranging from traditional machine learning techniques to deep learning models. However, there are still challenges that need to be addressed, such as the impact of cross-cultural differences in emotional expression and the need for larger annotated datasets. This section provides an overview of the existing research in speech emotion recognition.

Zeng et al. [1] introduced a deep neural network architecture called Gated Residual Neural Networks (GResNets) for recognizing emotions in speech. This architecture combines the power of Deep Residual Networks with a gate mechanism that helps minimize the gradient exploding problem and identify more informative features from the spectrogram. The output of this feature representation was evaluated on two tasks: emotion recognition from speech and speaker accent recognition. On the RAVDESS dataset, GResNets recorded an accuracy of 64.48% for multi-task recognition, demonstrating its potential in speech emotion recognition tasks.

A deep belief networks (DBN) for speech emotion recognition was proposed by Latif et al. [2]. The proposed method utilized a cross-language and cross-corpus transfer learning technique to improve the performance of speech emotion recognition. The speech signals were represented using eGeMAPS feature set, which includes 88 features such as frequency, energy, spectral, cepstral and dynamic information. The proposed DBN obtained a recognition accuracy of 54.77% on the IEMOCAP dataset and 72.38% on the Emo-DB dataset.

Singh et al. [3] performed speech emotion recognition using a kernel-based support vector machine (SVM) with radial basis function (RBF). The researchers utilized the scattering transform to extract both frequency domain and time domain feature representations from the audio signals. Three feature representations were extracted using the scattering transform, namely frequency scattering (F-ScatNet), time-domain scattering (ScatNet), and mel-frequency cepstral coefficients (MFCC). The performance of the proposed RBF with SVM model was evaluated using these features. The F-ScatNet feature representation achieved the highest accuracy among all three features, with 74.59% accuracy on the Emo-DB dataset, 51.81% accuracy on the RAVDESS dataset, and 61.55% accuracy on the IEMOCAP dataset.

Han et al. [4] suggested a parallel method for speech emotion recognition, called ResNet CNN-Transformer which combines the Residual Neural Network (ResNet) and Convolutional Neural Network (CNN). A transformer encoder was implemented to classify the frequency distribution of each emotion. The proposed method utilized mel spectrograms and MFCC feature representations for training. The performance of the proposed ResNet-CNN-Transformer was evaluated on the RAVDESS dataset and yielded an accuracy of 80.89%.

Similarly, Pandey et al. [5] proposed a combination speech emotion recognition model by combining the CNN and Bi-directional Long Short-Term Memory (Bi-LSTM) classifier. The proposed model was trained on three different feature representations, namely spectrogram, mel spectrogram, and MFCC. The results showed that the best performance was achieved with MFCC features, with an accuracy of 82.35% on the Emo-DB dataset.

Moreover, Swain et al. [6] devised a Concatenated Convolution Neural Network model that applied Gated Recurrent Unit (CGRU). The GRU was utilized to learn the small sequences of information from the prosodic and spectral features. The proposed method was evaluated on the RAVDESS dataset and recorded an accuracy of 73.26%.

Kerkeni et al. [7] utilized multiple classifiers and majority voting approach for speech emotion recognition. The approach used two classifiers, K-Nearest Neighbors (KNN) and SVM in combination with multiple feature extraction techniques. To select the most relevant features, an iterative neighborhood component analysis (INCA) technique was applied. The INCA worked by selecting a subset of features that capture the most relevant information in the input. Then, the highest correlation target input was passed into the two classifiers for voting. The proposed INCA with majority voting method achieved 80.76% accuracy on the RAVDESS dataset.

Likewise, Jha et al. [8] leveraged various machine learning models for speech emotion recognition, such as Gaussian Naive Bayes (GNB), Random Forest (RF), KNN, SVM, and Multilayer Perceptron (MLP). The proposed approach combined prosodic and spectral features, including MFCC, linear frequency cepstral coefficients (LFCC), spectral centroids, formants, pitch, and intensity. The RAVDESS dataset was used to evaluate the performance of the machine learning models. The best performance was achieved by MLP classifier using combined features, achieving an accuracy of 79.62%.

Zhang et al. [9] proposed a parallel model named Heterogeneous Parallel Convolution Bi-LSTM (HPCB) that combined CNN and Bi-LSTM classifiers for speech emotion recognition. The model used low-level descriptors (LLD) and high-level statistical functions (HSF), including Chroma, MFCC, mean of Chroma, and variance of MFCC. A total of 70 acoustic features were used to train the HPCB model. The HPCB model obtained an accuracy of 84.65% on the Emo-DB dataset.

Andayani et al. [10] combined the Long-Short Term Memory with Transformer (LSTM-Transformer) for speech emotion recognition. The audio signals were first processed using MFCC to extract the relevant features and then fed into the proposed LSTM-Transformer model for classification. The proposed model shown an accuracy of 75.33% on the RAVDESS dataset.

In a recent study, Ong et al. [11] proposed a speech emotion recognition method with Light Gradient Boosting Machine (LightGBM) referred to as the Emo-LGBM method. The data augmentation techniques, time stretching and pitch shifting, were applied to expand the dataset for model training. Following that, seven frequency domain and time domain features were extracted from the augmented audio samples. Subsequently, the extracted features were utilized as input for the LightGBM method to classify the emotional state conveyed in speech. The proposed method achieved 84.91% accuracy on the Emo-DB dataset, 67.72% on the RAVDESS dataset, and 62.94% on the IEMOCAP dataset.

III. SPEECH EMOTION RECOGNITION WITH MEL-STFT AND IMPROVED MULTISCALE VISION TRANSFORMERS

The research paper presents a methodology for speech emotion recognition that utilizes the Mel Spectrogram with Short-Time Fourier Transform (Mel-STFT) as a feature representation technique. The Mel-STFT captures relevant acoustic features by applying the STFT equation to short-time frames of the audio signal and converting the magnitude spectrum into the Mel scale using triangular filterbanks.

To improve classification performance, the methodology incorporates the Improved Multiscale Vision Transformers (MViTv2) as the classifier. MViTv2 introduces multiple stages for capturing information at varying granularity

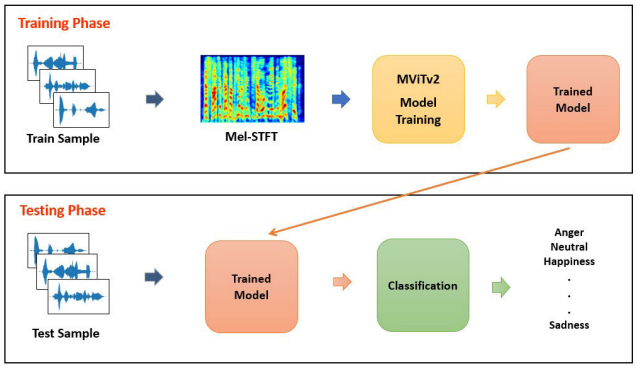


FIGURE 1. The system flow of Mel-MViTv2.

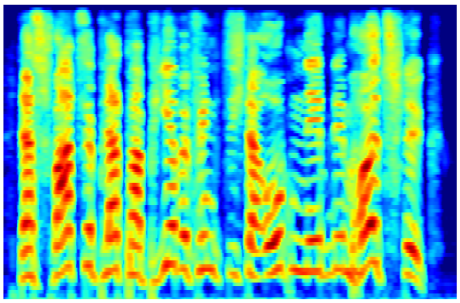


FIGURE 2. An example of a Mel-STFT showing the distribution of energy in different frequency bands over time.

levels, resulting in a comprehensive representation of the input data. It addresses limitations of its predecessor by incorporating relative positional embeddings to enhance space-time interaction modeling and utilizing a residual pooling connection technique to minimize information loss during pooling attention operations. Figure 1 shows the system flow of the proposed Mel-MViTv2.

A. MEL SPECTROGRAM WITH SHORT TIME FOURIER TRANSFORM

The Mel Spectrogram with Short-Time Fourier Transform (Mel-STFT) is a powerful audio visualization technique that converts audio signals into visual representations. The Mel-STFT combines the Short-Time Fourier Transform (STFT) with frequency-to-Mel scale conversion to create a more perceptually relevant representation of audio. To generate the Mel-STFT, the audio signal is first divided into short-time frames using the Hann window function, $w(n)$. This window function is applied to a small segment of the audio signal at a time and then shifted to cover the entire signal. The Hann window function helps in reducing the spectral leakage and improving frequency resolution. The STFT equation is used to calculate the frequency content of the audio signal at each point in time. The equation is given as follows:

$$X(m,k) = \sum_{n=0}^{N-1} x(n+mH) \times w(n) \times e^{-j2\pi n \frac{k}{N}} \quad (1)$$

where m is the frame index, k is the frequency bin index, and j is the imaginary unit. The N refers to the length of the window function used to segment the audio signal into shorter frames. The Hann window function $w(n)$ is applied to each frame to reduce spectral leakage and improve frequency resolution. The hop size H determines the amount of overlap between adjacent frames, affecting the time resolution of the STFT.

To convert the frequency bin index k into Mel-scale, the following equation is applied:

$$Mel(k) = 2595 \times \log_{10}(1 + f(k)/700) \quad (2)$$

where $f(k)$ is the frequency corresponding to the frequency bin in Hz. The resulting Mel-scale is used to apply a filterbank of triangular filters to the magnitude spectrum obtained from the STFT. The filterbank computes the energy in each filter, which represents the distribution of energy in different frequency bands over time. The resulting energy distribution is known as the Mel-spectrogram with Short-Time Fourier Transform.

The parameters of the Mel-STFT, such as the frame length and hop size, affect the time and frequency resolution of the spectrogram. In this experiment, a frame length of 4096 samples and a hop size of 256 samples are used. An example of a Mel-STFT is shown in Figure 2, illustrating the distribution of energy in different frequency bands over time.

B. IMPROVED MULTISCALE VISION TRANSFORMERS

The Improved Multiscale Vision Transformers (MViTv2) [12] are an upgrade to MViTv1 [13] that introduces different stages for multi-scale visual modeling, instead of single-scale blocks in Vision Transformer. The channel width is slowly expanded, while the resolution is reduced from input to output stages of the network. To perform downsampling within a transformer block, MViTv1 introduces Pooling Attention. This mechanism applies linear projections followed by pooling operators to query, key, and value tensors. Pooling attention enables resolution and computation reduction between different stages by pooling the query, key and value tensors. Given the input sequence $X \in \mathbb{R}^{L \times D}$ with sequence length L and channel width D, the pooling attention operations are defined as:

$$Q = \mathcal{P}_Q\left(XW_Q\right), K = \mathcal{P}_K\left(XW_K\right), V = \mathcal{P}_V\left(XW_V\right) \quad (3)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D \times D}$ denote the linear projections for query tensor Q, key tensor K and value tensor V, while $\mathcal{P}_Q, \mathcal{P}_K, \mathcal{P}_V$ denote the pooling operators for Q, K, and V.

However, MViTv1 has some potential improvements, which are addressed in the MViTv2. The first issue addressed is the modeling of the interactions between tokens in the space-time structure, which relies solely on the "absolute" positional embedding to offer location information. This ignores the fundamental principle of shift-invariance in vision, where the way MViT models the interaction between two patches will change depending on their absolute position

in images even if their relative positions stay unchanged. To address this issue, MViTv2 incorporates relative positional embeddings, which only depend on the relative location distance between input tokens into the pooled self-attention computation. The relative distance between two input tokens i and j with spatial position $p(i)$ and $p(j)$ is encoded into a positional embedding $R_{p(i),p(j)}$. More specifically, the computation of $R_{p(i),p(j)}$ is decomposed as:

$$R_{p(i),p(j)} = R^{\mathrm{h}}_{h(i),h(j)} + R^{\mathrm{w}}_{w(i),w(j)} \tag{4}$$

where R^{h} and R^{w} are the positional embeddings along the height and width axes. The symbols $h(i)$ and $h(j)$ represent the vertical position of the token i and j, while $w(i)$ and $w(j)$ denote the horizontal position of the token i and j. The positional embedding is then integrated into the self-attention module as:

$$\begin{aligned} &\mathrm{Attn}(Q, K, V) = \mathrm{Softmax}\left(\left(QK^{\top} + E^{(\mathrm{rel})}\right)/\sqrt{d}\right) V \\ &\text{where} \quad E^{(\mathrm{rel})}_{ij} = Q_i \cdot R_{p(i),p(j)} \end{aligned} \tag{5}$$

The second issue addressed is the possible information loss in the pooling attention while reducing computation complexity and memory requirements in attention blocks. To this end, the MViTv2 model employs a pooling technique known as residual pooling connection with pooled Q tensor which significantly enhances the information flow and assists in training pooling attention blocks. This technique helps to maintain low-complexity attention computation with large strides in the key, K and value V pooling, thereby improving the overall efficiency of the model. By applying the pooling connection in the query Q tensor, the input sequence size remains unchanged and there is no additional learning computation cost. The equation for the residual pooling connection is defined as:

$$Z := Attn(Q, K, V) + Q \tag{6}$$

The MViTv2 model integrates features from multiple scales, which enables the model to capture information at different granularity levels, leading to improved accuracy and performance. Additionally, MViTv2 utilizes the decomposed relative position embedding and residual pooling connection to preserve essential information at a lower computation cost. Figure 3 illustrates the architecture of MViTv2.

C. DATASETS

In order to ensure an objective evaluation of the proposed Mel-MViTv2 method, it has been tested on three speech emotion datasets: the Berlin Database of Emotional Speech (Emo-DB), the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and the Interactive Emotional Dyadic Motion Capture (IEMOCAP).

The Emo-DB [14] is a well-established speech emotion recognition dataset consisting of 535 audio samples from ten professional German speakers, including an equal representation of 5 male and 5 female actors. This dataset covers 7 distinct emotions, namely anger, boredom, neutral, happiness, anxiety, sadness, and disgust.

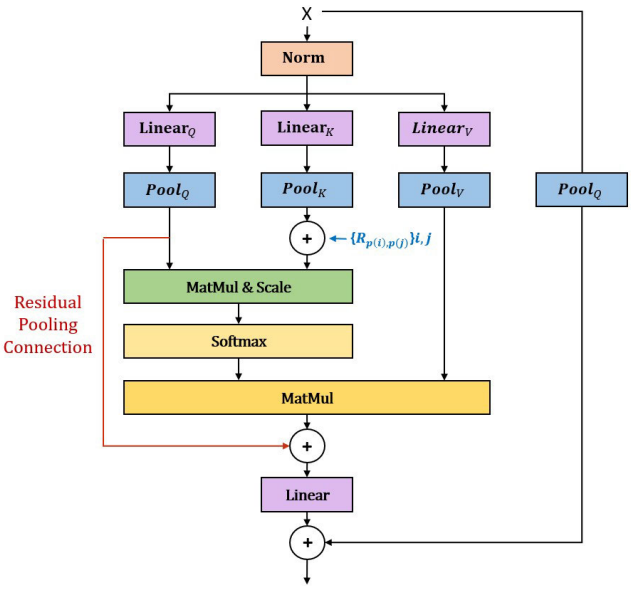


FIGURE 3. The architecture of improved multiscale vision transformers.

TABLE 1. Speech emotion recognition datasets.

Dataset	Speakers	Classes	Samples
Emo-DB [14]	10 (5M, 5F)	7	535
RAVDESS [15]	24 (12M, 12F)	8	1440
IEMOCAP [16]	10 (5M, 5F)	4	5507

Another widely recognized dataset employed in this evaluation is the RAVDESS [15]. It comprises 1440 audio samples in the English language, recorded by 24 professional actors, with an equal distribution of 12 male and 12 female speakers. RAVDESS covers a wide spectrum of emotions, encompassing neutral, calm, happy, sad, angry, fearful, disgust, and surprised.

The IEMOCAP [16] dataset, chosen for its extensive content and prior utilization in related research, consists of 5507 English-language audio samples recorded by five male and five female actors. In line with previous studies, our research focuses on analyzing four specific emotions within IEMOCAP: neutral, happiness, anger, and sadness.

IV. EXPERIMENTS AND ANALYSIS

To ensure consistency and compatibility, the data samples were resampled at a frequency of 44.1kHz. Subsequently, these samples were transformed into Mel-STFT spectrograms and resized to a resolution of 244 × 244, meeting the input size requirements of the MViTv2 classifier. The obtained spectrograms were then utilized as input for the MViTv2 model to perform speech emotion recognition. To ensure a fair comparison with existing works, the three datasets were divided into training and testing sets using an 80:20 ratio.

TABLE 2. Experimental results of different spectrograms with MViTv2 as the classifier. [Optimizer = Adam, Learning rate = 0.02].

Methods	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
Linear-STFT with MViTv2	79.25	63.16	57.95
MFCC with MViTv2	76.42	76.14	61.67
CQT with MViTv2	86.79	73.33	61.76
Mel-STFT with MViTv2	**90.57**	**81.75**	**63.49**

A. EXPERIMENTAL RESULTS WITH DIFFERENT SPECTROGRAMS

The effectiveness of speech emotion recognition using the MViTv2 model was evaluated by comparing four different spectrograms on three datasets: Emo-DB, RAVDESS, and IEMOCAP. Table 2 presents the results for the spectrograms, which were Linear-STFT, MFCC, CQT, and Mel-STFT with same optimizer and learning rate. The results indicate that among the four different spectrograms, the Linear-STFT spectrogram exhibited the lowest accuracy. In contrast, the Mel-STFT spectrogram displayed exceptional accuracy and efficiency across all three datasets. While the MFCC spectrogram provides reasonable accuracy, it exhibited comparatively lower performance than the other methods. Although the CQT spectrogram yielded satisfactory results, its overall performance across the three datasets fell short of the Mel-STFT spectrogram.

B. HYPERPARAMETER TUNING

Hyperparameter tuning is a critical process aimed at determining the optimal hyperparameter settings for the Mel-MViTv2 method. By utilizing grid search, the hyperparameter tuning phase focuses on two key hyperparameters: the optimizer and the learning rate. The optimizer plays a central role in the model training procedure, as it aims to minimize the loss function and guide the model towards attaining an optimal configuration of parameters that yield superior performance. Simultaneously, the learning rate, a scalar hyperparameter, governs the step size by which the optimizer adjusts the model's parameters during each training iteration. The optimizer and learning rate collectively determine the magnitude of parameter updates and exert influence over the speed and quality of the model's convergence. In the hyperparameter tuning phase, several popular optimizers were evaluated, including Adaptive Moment Estimation, Rectified Adam, and Quasi-Hyperbolic Adam. Additionally, three distinct learning rates, specifically 0.01, 0.02, and 0.03, were tested to explore their impact on model performance.

Table 3 shows the hyperparameter tuning results of the Emo-DB dataset, the highest accuracy of 91.51% was achieved using the Quasi-Hyperbolic Adam (QHAdam) [17] optimizer with a learning rate of 0.03. QHAdam is an optimizer that combines the advantages of the quasi-hyperbolic momentum (QHM) algorithm with Adam. While Adam is widely recognized for its effectiveness in large-scale training, QHAdam proves to be particularly valuable when working with small datasets. QHAdam introduces a new weight update rule by controlling the influence of the current and past gradient. By appropriately adjusting the unmodified and previous gradients, QHAdam strikes a balance between momentum and adaptive gradient scaling, leading to improved convergence behavior. Hence, the QHAdam optimizer and learning rate of 0.03 are chosen as the optimal optimizer for the proposed Mel-MViTv2 method on the Emo-DB dataset.

TABLE 3. Experimental results in accuracy (%) on Emo-DB dataset with different optimizers and learning rates.

Optimizer	Learning Rate		
	0.01	0.02	0.03
Adam	83.02	90.57	85.85
RAdam	80.19	88.68	84.91
QHAdam	87.74	84.91	**91.51**

The hyperparameter tuning results of the RAVDESS dataset are presented in Table 4, the highest accuracy of 81.75% was achieved using the Adaptive Moment Estimation (Adam) [18] optimizer with a learning rate of 0.02. Adam is a first-order gradient-based optimization algorithm that utilizes adaptive estimates of lower-order moments. While Adam performs well in various scenarios, it offers particular advantages for medium-sized datasets. One of its key features is adaptive learning rate adjustment, where the learning rate for each parameter is adapted based on the magnitude of the gradients and historical gradient information. This adaptivity enables Adam to automatically adjust the learning rate during training, making it well-suited for handling medium-sized datasets effectively. Therefore, the Adam optimizer and learning rate of 0.02 are the optimal selection for the proposed Mel-MViTv2 method when applied to the RAVDESS dataset.

Table 5 outlines the hyperparameter tuning results of the IEMOCAP dataset, the highest accuracy of 64.03% was achieved using the Rectified Adam (RAdam) [19] optimizer with a learning rate of 0.02. RAdam is a variant of the Adam optimizer that aims to address the limitations of the original Adam optimizer, specifically the issue of unstable adaptive learning rates during the early stages of training when dealing with large datasets. RAdam incorporates a rectified behavior beyond a specific threshold, which enhances the stability of the learning rate and improves the overall training process. By adjusting its behavior, RAdam provides improved performance during the early training iterations, resulting in enhanced optimization outcomes. The RAdam optimizer and learning rate of 0.02 are chosen for the proposed Mel-MViTv2 method when working with the IEMOCAP dataset.

C. COMPARATIVE RESULTS WITH EXISTING WORKS

Based on the results presented in Table 6, the proposed Mel-MViTv2 method demonstrated superior performance

TABLE 4. Experimental results in accuracy (%) on RAVDESS dataset with different optimizers and learning rates.

Optimizer	Learning Rate		
	0.01	0.02	0.03
Adam	76.49	**81.75**	80.70
RAdam	80.35	77.89	81.40
QHAdam	79.30	79.30	80.70

TABLE 5. Experimental results in accuracy (%) on IEMOCAP dataset with different optimizers and learning rates.

Optimizer	Learning Rate		
	0.01	0.02	0.03
Adam	59.58	63.49	63.40
RAdam	60.13	**64.03**	62.31
QHAdam	62.31	62.31	61.94

compared to existing methods in the field of emotion recognition in speech. On the Emo-DB dataset, the proposed method achieved an impressive accuracy of 91.51%. This accuracy outperformed all existing methods, whose accuracies ranged from 58.39% to 84.91%. The diverse nature of the emotional data in Emo-DB requires robust and adaptive methods for accurate recognition. The high accuracy achieved by the proposed Mel-STFT with MViTv2 approach, suggests that the method can effectively capture and analyze the discriminative emotional features present in the database.

For the RAVDESS dataset, the proposed Mel-MViTv2 method yielded an accuracy of 81.75%. This method demonstrated an improvement of 0.86% compared to the best-performing method, ResNet-CNN-Transformer [4]. The RAVDESS dataset is known for its relatively challenging nature, primarily due to the larger number of speakers it encompasses. This larger speaker count leads to inherently higher inter-subject variations, making emotion recognition more complex and demanding.

All existing methods tend to exhibit relatively lower performance on the IEMOCAP dataset, primarily due to its unique characteristics. The dataset captures emotional expressions in dyadic sessions, where interactions between actors occur, potentially resulting in mixtures of emotions within the samples. However, despite this challenge, the proposed Mel-STFT with MViTv2 method managed to record an accuracy of 64.03%. This accuracy surpassed the range of 55.54% to 62.94% achieved by the methods in comparison.

The results indicate that the proposed Mel-MViTv2 method outperforms existing methods on all three datasets, namely Emo-DB, RAVDESS, and IEMOCAP. The mel-STFT features are effective in capturing relevant acoustic information related to speech, such as spectral characteristics and energy distribution. The mel-frequency scale represents the perception of pitch and frequency in a manner more aligned with human hearing, allowing the model to focus

TABLE 6. Comparative results on Emo-DB, RAVDESS, IEMOCAP dataset.

Methods	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
GResNets [1]	-	64.48	-
DBN [2]	72.38	54.77	-
F-ScatNet features with SVM [3]	74.59	51.81	61.55
ScatNet features with SVM [3]	74.40	50.00	60.41
MFCC features with SVM [3]	58.39	36.74	55.54
ResNet-CNN-Transformer [4]	-	80.89	-
CNN-BLSTM [5]	82.35	-	-
CGRU [6]	-	73.26	-
INCA with Majority Voting [7]	-	80.76	-
MLP [8]	-	79.62	-
SVM [8]	-	77.86	-
HPCB [9]	84.65	-	-
LSTM-Transformer [10]	-	75.33	-
Emo-LGBM [11]	84.91	67.72	62.94
Mel-MViTv2 (Proposed)	**91.51**	**81.75**	**64.03**

on important aspects of the audio signal. The MViTv2 architecture leverages the power of the transformer model to capture complex patterns and long-range dependencies in sequential data, making them well-suited for analyzing mel-STFT spectrograms. Not only that, multiscale feature hierarchies, a key aspect of the MViTv2 architecture, facilitate the modeling of mel-STFT at multiple levels of abstraction. Mel-STFT spectrograms contain both local and global acoustic information, and the hierarchical nature of the MViTv2 architecture enables the extraction of meaningful features across different scales. The stages in the MViTv2 architecture hierarchically expand the channel capacity while reducing the spatial resolution. This allows the model to capture fine-grained details as well as higher-level semantic information from the spectrograms, enhancing the discriminative power of the features.

Figure 4 and Figure 5 present the confusion matrices obtained from evaluating the proposed Mel-MViTv2 method on the Emo-DB and RAVDESS datasets. These matrices offer a comprehensive view of the classification performance by comparing the predicted classes against the ground truth labels. It is noteworthy that the misclassification rate tends to be higher for classes with limited sample sizes, primarily due to the scarcity of training data available for these classes. This phenomenon can be attributed to the similarities in acoustic features, such as pitch variations, intensity fluctuations, and speech rate alterations, which pose challenges in distinguishing between certain emotions. Moreover, accurately classifying speech emotion is further complicated by the inter-individual variability in the expression of emotions.

In the case of the IEMOCAP dataset, the confusion matrix in Figure 6 provides insights into the performance of the proposed Mel-MViTv2 method. This dataset poses additional

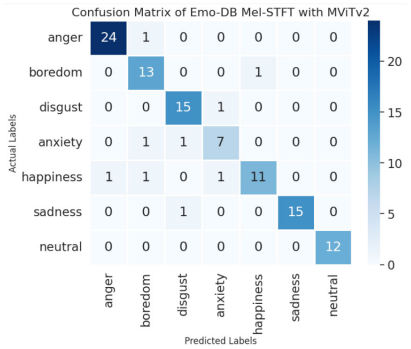


FIGURE 4. Confusion matrix of the Emo-DB.

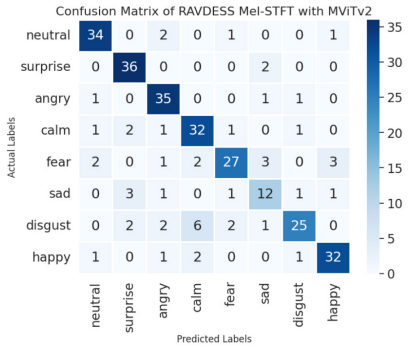


FIGURE 5. Confusion matrix of the RAVDESS dataset.

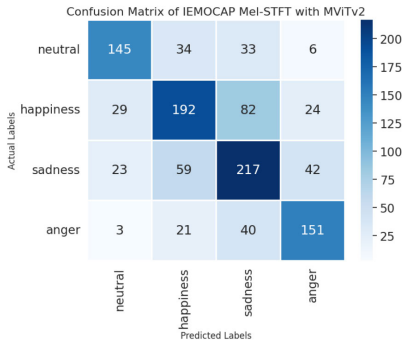


FIGURE 6. Confusion matrix of the IEMOCAP dataset.

challenges as it comprises dyadic speech with multiple sources simultaneously. Analyzing the confusion matrix reveals that happiness and sadness are particularly prone to misclassification. This can be attributed to the intricate nature of emotional expression, where inter-individual differences, the presence of mixed emotions within samples, and the subjectivity involved in emotion labeling contribute to the classification difficulties encountered.

V. CONCLUSION

This paper presents a speech emotion recognition method, known as "Mel-MViTv2" using Mel-STFT spectrograms and MViTv2 as a classifier. The Mel-STFT leverages the Short-Time Fourier Transform technique along with frequency-to-Mel scale conversion to generate perceptually relevant visual representations of audio signals. By dividing the audio signal into short-time frames using a window function, such as the Hann window, and applying the STFT equation, the frequency content of the audio signal at each time point is obtained. The resulting magnitude spectrum is then transformed into the Mel scale through triangular filterbanks, yielding the Mel-STFT representation. This technique effectively captures acoustic features such as pitch changes, intensity variations, and speech rate, which are relevant for speech emotion classification.

The MViTv2 introduces different stages for multi-scale visual modeling, improving the capture of information at different granularity levels. MViTv2 incorporates relative positional embeddings that consider the relative distance between input tokens, enhancing the modeling of token interactions. Furthermore, MViTv2 employs a pooling technique called residual pooling connection to mitigate potential information loss during pooling attention while maintaining computational efficiency. This technique enhances information flow and facilitates training of pooling attention blocks. By integrating features from multiple scales and incorporating relative positional embeddings and residual pooling connections, MViTv2 improves the accuracy and efficiency of speech emotion recognition models. The proposed Mel-STFT with MViTv2 model achieved promising results, recording the highest accuracy of 91.51%, 81.75%, and 64.03% on the Emo-DB, RAVDESS, and IEMOCAP datasets, respectively.

REFERENCES

[1] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019.

[2] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," 2018, *arXiv:1801.06353*.

[3] P. Singh, G. Saha, and M. Sahidullah, "Deep scattering network for speech emotion recognition," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 131–135.

[4] S. Han, F. Leng, and Z. Jin, "Speech emotion recognition with a ResNet-CNN-Transformer parallel neural network," in *Proc. Int. Conf. Commun., Inf. Syst. Comput. Eng. (CISCE)*, May 2021, pp. 803–807.

[5] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, "Deep learning techniques for speech emotion recognition: A review," in *Proc. 29th Int. Conf. Radioelektronika (RADIOELEKTRONIKA)*, Apr. 2019, pp. 1–6.

[6] M. Swain, B. Maji, and U. Das, "Convolutional gated recurrent units (CGRU) for emotion recognition in odia language," in *Proc. IEEE 19th Int. Conf. Smart Technol. (EUROCON)*, Jul. 2021, pp. 269–273.

[7] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using machine learning," in *Social Media and Machine Learning*. London, U.K.: IntechOpen, 2019.

[8] T. Jha, R. Kavya, J. Christopher, and V. Arunachalam, "Machine learning techniques for speech emotion recognition using paralinguistic acoustic features," *Int. J. Speech Technol.*, vol. 25, no. 3, pp. 707–725, Sep. 2022.

[9] H. Zhang, H. Huang, and H. Han, "A novel heterogeneous parallel convolution bi-LSTM for speech emotion recognition," *Appl. Sci.*, vol. 11, no. 21, p. 9897, Oct. 2021.

[10] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Recognition of emotion in speech-related audio files with LSTM-transformer," in *Proc. 5th Int. Conf. Comput. Informat. (ICCI)*, Mar. 2022, pp. 87–91.

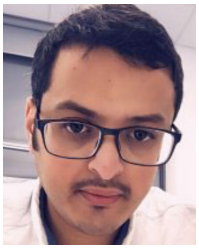
[11] K. L. Ong, C. P. Lee, H. S. Lim, and K. M. Lim, "Speech emotion recognition with light gradient boosting decision trees machine," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 13, no. 4, p. 4020, Aug. 2023.
[12] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "MViTv2: Improved multiscale vision transformers for classification and detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4794–4804.
[13] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6804–6815.
[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, vol. 5, 2005, pp. 1517–1520, doi: 10.21437/interspeech.2005-446.
[15] 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, May 2018, Art. no. e0196391, doi: 10.1371/journal.pone.0196391.
[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008, doi: 10.1007/s10579-008-9076-6.
[17] J. Ma and D. Yarats, "Quasi-hyperbolic momentum and Adam for deep learning," 2018, *arXiv:1810.06801*.
[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[19] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," 2019, *arXiv:1908.03265*.

KAH LIANG ONG received the bachelor's degree (Hons.) in information technology artificial intelligence from Multimedia University, Malaysia, in 2021. He is currently pursuing the full-time master's degree. His current research interests include speech emotion recognition which mainly involves audio pre-processing, feature extraction, and emotion classification.

CHIN POO LEE (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in information technology in the area of abnormal behavior detection and gait recognition. She has been a certified Professional Technologist, since 2018, has been a member of the International Association of Engineers, since 2020, and a Outcome-Based Education Consultant and a Trainer. She is currently a Senior Lecturer with the Faculty of Information Science and Technology, Multimedia University, Malaysia. Her research interests include action recognition, computer vision, gait recognition, natural language processing, and deep learning.

HENG SIONG LIM (Senior Member, IEEE) received the B.Eng. degree (Hons.) in electrical engineering from Universiti Teknologi Malaysia, in 1999, and the M.Eng.Sc. and Ph.D. degrees in engineering focusing on signal processing for wireless communications from Multimedia University, in 2002 and 2008, respectively. He is currently a Professor with the Faculty of Engineering and Technology, Multimedia University. His current research interests include signal processing for advanced communication systems, with emphasis on detection and estimation theory and their applications.

KIAN MING LIM (Senior Member, IEEE) received the B.I.T. degree (Hons.) in information systems engineering and the M.Eng.Sc. and Ph.D. (I.T.) degrees from Multimedia University. He is currently a Lecturer with the Faculty of Information Science and Technology, Multimedia University. His research and teaching interests include machine learning, deep learning, computer vision, and pattern recognition.



ALI ALQAHTANI received the Ph.D. degree in computer science from Swansea University, Swansea, U.K., in 2021. He is currently an Assistant Professor with the Department of Computer Science, King Khalid University, Abha, Saudi Arabia. He has published several refereed conference and journal publications. His research interests include various aspects of pattern recognition, deep learning, and machine intelligence and their applications to real-world problems.