IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 January 2024, accepted 27 January 2024, date of publication 31 January 2024, date of current version 7 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3360483

RESEARCH ARTICLE

MaxMViT-MLP: Multiaxis and Multiscale Vision Transformers Fusion Network for Speech Emotion Recognition

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, Melaka 75450, Malaysia
[2]Faculty of Engineering and Technology, Multimedia University, Melaka 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 the Telekom Malaysia Research and Development under Grant RDTC/221064 and Grant RDTC/231075; and in part by the Deanship of Scientific Research at King Khalid University, Saudi Arabia, under Grant RGP2/332/44.

ABSTRACT Vision Transformers, known for their innovative architectural design and modeling capabilities, have gained significant attention in computer vision. This paper presents a dual-path approach that leverages the strengths of the Multi-Axis Vision Transformer (MaxViT) and the Improved Multiscale Vision Transformer (MViTv2). It starts by encoding speech signals into Constant-Q Transform (CQT) spectrograms and Mel Spectrograms with Short-Time Fourier Transform (Mel-STFT). The CQT spectrogram is then fed into the MaxViT model, while the Mel-STFT is input to the MViTv2 model to extract informative features from the spectrograms. These features are integrated and passed into a Multilayer Perceptron (MLP) model for final classification. This hybrid model is named the "MaxViT and MViTv2 Fusion Network with Multilayer Perceptron (MaxMViT-MLP)." The MaxMViT-MLP model achieves remarkable results with an accuracy of 95.28% on the Emo-DB, 89.12% on the RAVDESS dataset, and 68.39% on the IEMOCAP dataset, substantiating the advantages of integrating multiple audio feature representations and Vision Transformers in speech emotion recognition.

INDEX TERMS Speech emotion recognition, ensemble learning, spectrogram, vision transformer, Emo-DB, RAVDESS, IEMOCAP.

I. INTRODUCTION

Speech emotion recognition stands at the confluence of signal processing and machine learning, addressing the automatic identification and classification of emotional expressions within spoken language. Signal processing forms the bedrock of speech emotion recognition, involving the extraction of significant information from speech signals, including spectral features, prosodic cues, and acoustic characteristics. Machine learning techniques play a pivotal role, enabling the development of models that recognize emotions by learning patterns and features associated with different emotional states. With applications spanning human-computer interaction, sentiment analysis, mental health assessment, and customer service enhancement, speech emotion recognition has captured the attention of researchers and practitioners alike.

In light of this, our paper introduces a dual-path approach referred to as the "MaxViT and MViTv2 Fusion Network with Multilayer Perceptron (MaxMViT-MLP)". The method encodes the speech signals into two representations: the Constant-Q Transform (CQT) spectrogram and the Mel Spectrogram via Short-Time Fourier Transform (Mel-STFT). This dual spectrogram strategy leverages the complementary attributes of CQT and Mel-STFT, providing a holistic and informative portrayal of the input data. The CQT spectrogram

The associate editor coordinating the review of this manuscript and approving it for publication was Sotirios Goudos.

is routed to the MaxViT model, while the Mel-STFT is channeled to the MViTv2 model. These Vision Transformers excel in extracting meaningful features from their respective spectrogram inputs. The resulting features are integrated, culminating in a comprehensive representation of the input data, which is then directed into a Multilayer Perceptron (MLP) for the final classification. The main contributions of this work can be summarized as follows:

- Introduces a dual-path architecture for speech emotion recognition, denoted as the "MaxViT and MViTv2 Fusion Network with Multilayer Perceptron (MaxMViT-MLP)". This approach involves encoding speech signals into two distinct representations. Subsequently, each representation is channeled into its respective Vision Transformer model. The features from the Vision Transformer models are amalgamated and subjected to a MLP for further representation learning and classification.
- Represents speech signals in dual spectrograms: CQT and Mel-STFT. CQT uses logarithmic frequency binning, and one of its key strengths is its constant-Q resolution, meaning it offers a higher resolution at lower frequencies and a coarser resolution at higher frequencies. This makes it particularly effective at capturing fine details in the low-frequency range, which is well-suited for harmonic analysis. Mel-STFT uses a non-linear frequency scale to capture broader spectral characteristics and is more intuitive for human interpretation of sound. Mel-STFT provides a detailed time-frequency representation of the audio signal, making it suitable for tasks that involve transient events. This dual spectrogram strategy capitalizes on the complementary characteristics of CQT and Mel-STFT, resulting in a comprehensive representation of the speech signals.
- Incorporates MaxViT and MViTv2 for representation learning on the CQT and Mel-STFT spectrograms, respectively. The hierarchical architecture of MaxViT allows it to capture multiscale information effectively. The block attention module partitions the input feature map into distinct windows and applies self-attention mechanisms to foster contextual understanding, allowing MaxViT to recognize intricate patterns and relationships within audio data. The grid attention module attends globally to pixels through a sparse grid, which is beneficial for capturing global contextual information. On the other hand, MViTv2 facilitates the capture of multiscale features with different channel-resolution scales. The utilization of relative positional embeddings in MViTv2 injects shift-invariance properties, enhancing its ability to understand spatial relationships within the spectrograms.

II. RELATED WORKS

Researchers have ventured into a multitude of learning approaches for speech emotion recognition. Nevertheless, the complexity of speech emotion recognition endures, primarily stemming from the myriad variations in speaking styles, gender, cultural influences, emotional expression, and other factors. This section presents an overview of the existing research in the field of speech emotion recognition.

Wen et al. [1] presented a fusion model that combined the strengths of a capsule network and a Convolutional Neural Network, denoted as CapCNN. The method involved a two-step pre-processing procedure comprising voice activity detection and a windowed framework. These steps speech emotion recognitionved to locate the speech segments and enhance the overall quality of the audio signals. The CapCNN architecture was trained using a diverse set of input features, encompassing MFCC, spectrogram, and spectral features. The proposed CapCNN model, when applied to spectral features, achieved remarkable performance on the Emo-DB with an accuracy of 82.90%.

A two-dimensional Convolutional Neural Network (2D-CNN) was proposed by Mujaddidurrahman et al. [2] for speech emotion recognition. The method involved data augmentation with noise injection and variations in loudness to enhance the diversity of the training data. Following augmentation, the speech signals were transformed into log-mel spectrogram features, which were used as input for the 2D-CNN model. The proposed model achieved an accuracy of 88% on the Emo-DB.

He and Ren [3] utilized various techniques for speech emotion recognition, including XGBoost, Convolutional Neural Network (CNN), and Bi-directional Long Short-term Memory (BiLSTM) with an attention model. The speech signals first went through the framing with a 25ms frame length. Subsequently, 34 low-level descriptors were extracted from each frame. These descriptors were then passed into the XGBoost classifier for feature selection. Subsequently, the chosen features were fed into a hybrid architecture combining both CNN and BiLSTM with attention model. The proposed CNN and BiLSTM with attention model achieved 86.87% accuracy on the Emo-DB.

In this study, Ancilin and Milton [4] employed mel-frequency magnitude coefficient (MFMC) features for speech emotion recognition. The MFMC features encapsulated a logarithmic representation of the magnitude spectrum aligned with the non-linear mel frequency scale. The MFMC features were then classified using a Support Vector Machine (SVM). The experimental results showed that the MFMC-SVM method yielded an accuracy of 81.50% on the Emo-DB and 64.31% on the RAVDESS dataset.

Pham et al. [5] explored various spectral features, namely mel frequency cepstral coefficients (MFCCs), mel scaled spectrogram, chromagram, spectral contrast feature, and tonnetz representation for speech emotion recognition. The method calculated the mean values of the features and stacked them together to form spectral mean vector features. Thereafter, CNN was engaged to learn and classify the spectral mean vector features. The proposed method yielded 76.40% accuracy on the Emo-DB and 70.80% accuracy on the RAVDESS dataset.

Singh et al. [6] presented an approach for speech emotion recognition using scattering transforms applied to speech signals. The method incorporated diverse features, including frequency-domain scattering representation (F-SCATNET), time-domain scattering representation (SCATNET), and MFCCs. The classifier utilized was a radial basis function kernel-based SVM classifier. The F-SCATNET with SVM achieved a recognition rate of 74.59% on the Emo-DB, 51.81% on the RAVDESS dataset, and 61.55% on the IEMOCAP dataset.

Tuncer et al. [7] presented a novel approach for speech emotion recognition. The method started with the transformation of speech signals through a Tunable Q wavelet transform (TQWT) together with a twine shuffle pattern feature generator (twine-shuf-pat) to extract pertinent features. Subsequently, the features underwent a feature selection process via iterative neighborhood component analysis (INCA), concluding in a refined set of features. These final features were then passed to an SVM classifier for model training. The proposed model attained an impressive accuracy of 90.09% on the Emo-DB and 87.42% on the RAVDESS dataset.

Thirumuru et al. [8] introduced a novel representation, known as the single frequency filtered-nonlinear energy cepstral coefficients (SFF-NEC) for speech emotion recognition. This representation is constructed by employing the nonlinear energy operator in conjunction with single frequency filtering on distinct frequency sub-bands. The SFF-NEC was then transformed into an identity vector (i-vector), making it a compact and low-dimensional representation. As for the classification, three models were evaluated: the Gaussian probabilistic linear discriminant analysis (G-PLDA), SVM, and Random Forest. The i-vector integrated with the SVM classifier achieved accuracies of 85.75% and 65.78% on the Emo-DB and IEMOCAP dataset, respectively.

A hybrid Long Short-Term Memory (LSTM) combined with a Transformer Encoder for speech emotion recognition was proposed by Andayani et al. [9]. The proposed model used the MFCCs extracted from speech signals as the features. As for the enhanced LSTM, the researchers replaced the single attention layer within the LSTM architecture with the multi-head attention mechanism inherent to the Transformer encoder. This adaptation improved the capability of the model to learn intricate patterns and features from the input data. The method achieved an accuracy of 85.55% on the Emo-DB and 75.62% on the RAVDESS dataset.

In another work, Hason Rudd et al. [10] first converted the speech signals into the mel spectrogram representation. Then, they applied a VGG16 model to extract feature maps with various dimensions and signal sampling ratios. The feature maps were subsequently fed into a multi-layer perceptron (MLP) architecture for classification. The proposed method achieved an accuracy of 92.79% with a bandwidth of 128 Hz, a frame rate of 128 fps, and a sampling rate of 88200 KHz on the Emo-DB.

Kakuba and Han [11] presented a multi-head attention machine with residual Bi-LSTM called (ResBLSTMA). The work leveraged spectral and voice representations. Spectral representations encompassed crucial features like MFCCs and chromagrams, while voice representations involved mel spectrograms. These features were then passed as feature vectors into the ResBLSTMA model, which effectively leveraged its multi-head attention mechanism and residual bidirectional long short-term memory for classification tasks. The proposed ResBLSTMA method achieved an accuracy of 90.57% on the Emo-DB and 84.50% on the RAVDESS dataset.

In the study by Singh et al. [12], the researchers investigated time-frequency-based features for speech emotion recognition. Three specific features were explored: mel-frequency spectral coefficients (MFSC), constant-Q transform (CQT), and continuous wavelet transform (CWT). Classification was performed by a 2D-CNN with LSTM architecture, referred to as Conv2D-LSTM. The proposed Conv2D-LSTM with CQT features demonstrated accuracies of 65.69%, 46.49%, and 54.83% on the Emo-DB, RAVDESS, and IEMOCAP datasets, respectively.

Vu et al. [13] applied principal component analysis (PCA) to pitch-related features for speech emotion recognition. PCA was used to reduce the dimensionality of feature vectors, mitigating the learning complexity of the model. The CQT spectrogram served as the input in this work. For classification, an MLP with two dense neural network layers activated by the ReLU function, followed by softmax layers, was employed. The proposed MLP with CQT, excluding PCA, achieved the highest accuracy of 66.67% on the RAVDESS dataset and 57.09% on IEMOCAP.

Sekkate et al. [14] introduced a statistical-based technique for speech emotion recognition. The researchers first converted the speech signals into MFCC where the mean values of MFCC were calculated as features. Subsequently, the MFCC features were subjected to a statistical distribution process to select pertinent features. The selected features were then employed as input for a combination of three CNNs for model training. Each CNN contains three 1D-CNN layers, followed by the max pooling and batch normalization layers. The decision scores from each classifier will be merged to determine the final class label. The method was evaluated on the RAVDESS dataset and achieved an accuracy of 87.08%.

Mishra et al. [15] performed a comparative analysis of different features and classifiers in speech emotion recognition. Their study extracted two features: MFCC and MFMC. These features were then employed for training and testing of deep neural networks (DNN) and CNN classifiers. The experimental results demonstrated that the MFMC features with DNN obtained the highest accuracy of 84.72% on the Emo-DB and 76.72% on the RAVDESS dataset.

Rehman et al. [16] employed the syllable-level feature extraction for speech emotion recognition. The speech signals were first transformed into mel-spectrograms, followed by segmentation into distinct syllable-level components. The syllable-level components were then encoded as the statistical features. A SVM was utilized in classification. The proposed method yielded promising performance achieving an accuracy of 62.90% when evaluated on the IEMOCAP dataset.

Tu et al. [17] performed speech emotion recognition using feature fusion and data augmentation techniques. Initially, the speech signals were augmented by randomly combining segments from various speech signals. The resulting augmented speech signals were then transformed into log-mel spectrograms and high-level statistical features. Subsequently, a LightGBM classifier was employed for global feature selection of the statistical features. A deep learning model incorporating the multi-head attention mechanism into the CNN and LSTM, named MHA-CRNN, was proposed. This MHA-CRNN was used to perform feature fusion and classification of the log-mel spectrogram features and the selected statistical features. The proposed method achieved an accuracy of 66.44% on the IEMOCAP dataset.

Feature fusion was also employed in Liu et al. [18], integrating acoustic and pre-trained features. The speech signals were segmented and encoded as acoustis features using the OpenSmile [19] library. The study also incorporated various pre-trained features, such as self-supervised learning of Transformer encoder representation (Tera) [20], A Lite BERT for self-supervised learning of audio representation (Audio ALBERT) [21], non-autoregressive predictive coding for learning speech representations (NPC) [22], unsupervised pre-training representations (Wav2Vec) [23], and self-supervised learning of discrete speech representations (Vq-wav2vec) [24]. The authors proposed a Transformer-inspired model with attention mechanism, along with two 1D convolutional layers, two Transformer modules, and two BiLSTM modules, forming the innovative Dual-TBNet architecture. The Dual-TBNet achieved accuracies of 84.10% on the Emo-DB and 64.80% on the IEMOCAP dataset.

Ong et al. [25] engaged a lightweight gradient boosting machine (LightGBM) approach for speech emotion recognition, called Emo-LGBM. The researchers first applied pitch shifting and time stretching data augmentation techniques to the input speech signals. Subsequently, both time and frequency domain features were extracted from the augmented signals as the input to the LightGBM. The proposed Emo-LGBM approach achieved an accuracy of 84.91% on the Emo-DB, 67.72% on the RAVDESS dataset, and 62.94% on the IEMOCAP dataset.

Singh et al. [26] introduced constant-Q based modulation spectral features (CQT-MSF) by combining CQT with modulation spectral features (MSF). The author utilized CNN to extract feature embeddings from CQT-MSF and employed SVM to classify the embeddings into emotions. The results highlight that the proposed DNN-SVM with CQT-MSF outperforms a single mel-scale-based spectrogram, achieving 79.86% accuracy on the Emo-DB and 52.24% on the RAVDESS dataset. Table 1 presents the summary of related works.

III. SPEECH EMOTION RECOGNITION WITH MULTIAXIS AND MULTISCALE VISION TRANSFORMERS FUSION NETWORK

This paper presents a novel dual-path architecture for speech emotion recognition. In the first path, speech signals are transformed into CQT spectrograms, which are subsequently processed by the MaxViT model for feature extraction and representation learning. Simultaneously, the second path encodes speech signals into Mel-STFT spectrograms, which are then subjected to feature extraction using the MViTv2 model. The feature maps generated by MaxViT and MViTv2 are concatenated to create a comprehensive representation of the input data. This combined feature representation is then subjected to classification using the MLP algorithm. The dual-path architecture, which incorporates both CQT and Mel-STFT spectrograms, leverages the strengths of Multiaxis and Multiscale Vision Transformers, resulting in improved performance for speech emotion recognition.Figure 1 illustrates the framework of the proposed MaxMViT-MLP.

A. CONSTANT-Q TRANSFORM SPECTROGRAM (CQT)

The CQT spectrogram is a time-frequency representation used in audio signal processing. It is designed to closely mimic the frequency resolution of the human auditory system. This makes CQT particularly well-suited for speech emotion recognition, as it aligns with how the ears perceive different frequencies. The process of creating a CQT spectrogram involves convolving the audio signal with a set of complex exponential functions. These functions are evenly distributed in logarithmic frequency steps while maintaining a consistent Q-factor (center frequency to bandwidth ratio) across all bins. This results in a time-frequency representation where each frequency bin corresponds to a specific Q value and center frequency. The CQT spectrogram captures the variations in the frequency content of the audio signal over time, providing valuable insights into its tonal and harmonic components.

Mathematically, the CQT at a specific frequency, f and time, t can be expressed as:

$$CQT(f, t) = \sum_{n=0}^{N-1} x(n) \cdot w^{*}(\frac{n-t}{s})e^{2\pi ifn} \tag{1}$$

where $x(n)$ is the input signal, w represents the wavelet, s represents the scale factor that controls the width of the wavelet, and i is the imaginary unit. By varying the frequency, f, and time, t, the CQT generates a time-frequency representation, where each entry corresponds to the energy of the signal at a particular frequency and time. This representation encapsulates both fine and broad frequency details, making it highly suitable for speech

TABLE 1. Summary of related works.

Reference	Feature Used	Classification Methods
Wen et al. (2020) [1]	MFCC, spectrogram, spectral features	CapCNN
Mujaddidurrahman et al. (2021) [2]	log-mel spectrogram	2D-CNN
He & Ren (2021) [3]	34 low-level descriptors	CNN with BiLSTM
Ancilin & Milton (2021) [4]	MFMC	SVM
Pham et al. (2021) [5]	MFCC, mel scaled spectrogram, chromagram, spectral contrast feature, tonnetz representation	CNN
Singh et al. (2021) [6]	F-SCATNET	SVM
Tuncer et al. (2021) [7]	TQWT	SVM
Thirumuru et al. (2022) [8]	SFF-NEC	SVM
Andayani et al. (2022) [9]	MFCC	LSTM with Transformer Encoder
Hason Rudd et al. (2022) [10]	mel spectrogram	MLP
Kakuba & Han (2022) [11]	MFCC, chromagrams	ResBLSTMA
Singh et al. (2022) [12]	CQT	Conv2D-LSTM
Vu et al. (2022) [13]	CQT	MLP
Sekkate et al. (2023) [14]	MFCC	CNN
Mishra et al. (2023) [15]	MFCC, MFMC	DNN
Rehman et al. (2023) [16]	mel-spectrograms	SVM
Tu et al. (2023) [17]	log-mel spectrograms, high-level statistical feature	MHA-CRNN
Liu et al. (2023) [18]	acoustic features	Dual-TBNet
Ong et al. (2023) [25]	time and frequency domain features	LightGBM
Singh et al. (2023) [26]	CQT-MSF	DNN-SVM

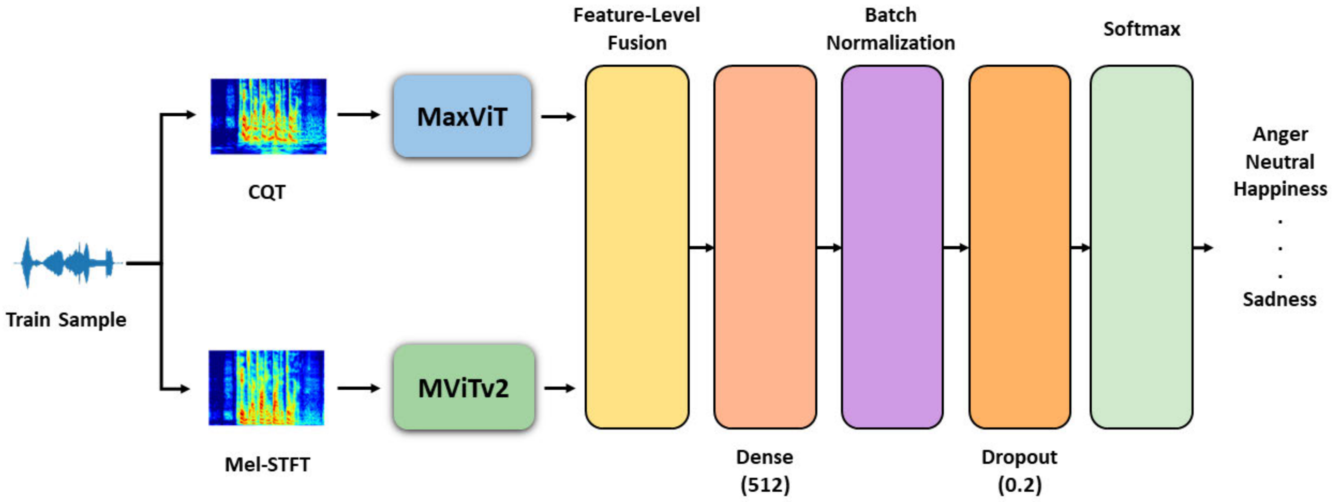


FIGURE 1. System flow of the proposed MaxMViT-MLP model.

emotion recognition. An example of the CQT spectrogram is depicted in Figure 2.

B. MULTI-AXIS VISION TRANSFORMER (MAXVIT)

The CQT spectrograms serve as the input data for the MaxViT model [27] for further representation learning. The MaxViT model is designed in a hierarchical architecture, as depicted in Figure 3. The CQT spectrograms are downsampled in the stem stage (S0), which is composed of two convolutional layers with a 3 × 3 kernel size (Conv3 × 3). The body of MaxViT comprises four stages (S1-S4), wherein each stage comprises a MaxViT block. Each MaxViT block consists of a Mobile Inverted Residual Bottleneck Convolution (MBConv) module, a block attention module, and a grid attention module.

The MBConv module begins with a 1 × 1 convolution. This initial layer is responsible for dimensionality expansion. It takes the input feature maps and projects them into a higher-dimensional space, typically with an expansion factor of 4. Following the expansion, the module applies Depthwise

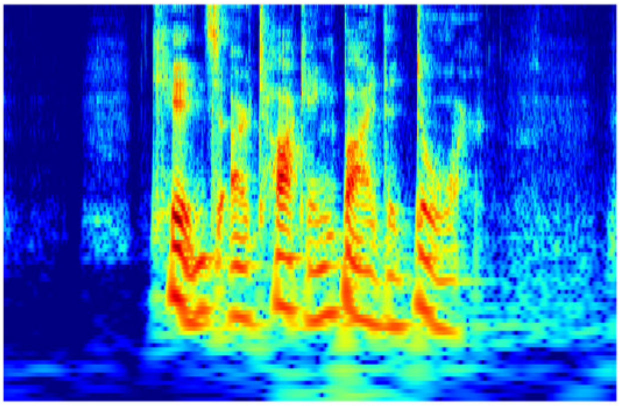


FIGURE 2. CQT spectrogram for the utterance 'Kids are talking by the door' from the RAVDESS dataset.

Convolution with a 3 × 3 kernel. Depthwise convolution operates on each input channel separately and captures spatial features across the input. Next, the module includes a Squeeze and Excitation (SE) mechanism. The SE mechanism is designed to adaptively recalibrate the importance of each channel in the feature maps. It consists of two steps: a global average pooling operation, which computes channel-wise statistics to capture the most informative features, and a set of fully connected layers. These fully connected layers learn channel-wise scaling factors, allowing the network to emphasize essential features while suppressing less relevant ones. Finally, the module concludes with another 1 × 1 convolution. This layer reduces the dimensionality back to the original channel dimension, achieving a shrink factor of 0.25. This reduction helps in managing computational complexity and memory usage while preserving the essential features learned throughout the module's operations.

The block attention module operates by first partitioning the input feature map into distinct windows. Specifically, given an input feature map X with dimensions $H \times W \times C$, where H represents height, W represents width, and C represents the number of channels, the block attention reshapes this input into a tensor with dimensions of $(\frac{H}{P} \times \frac{W}{P}, P \times P, C)$, where $P \times P$ denotes the dimension of each block. This transformation results in the creation of non-overlapping windows within the feature map, each of which is characterized by dimensions of $P \times P$. Within each of these windows, self-attention mechanisms are applied to capture interactions among the elements, fostering contextual understanding. Subsequently, a feedforward network (FFN) is employed to further process the information obtained from the block-based self-attention step. This FFN introduces non-linear transformations to the representations within each window, enabling the model to capture intricate patterns and relationships.

On the other hand, the grid attention module attends globally to pixels through a sparse, evenly distributed grid spanning the entire 2D space. The grid attention mechanism is applied by reshaping the tensor into dimensions of $(G \times G, \frac{H}{G} \times \frac{W}{G}, C)$, effectively gridding the feature maps into $G \times G$ partitions. In this work, fixed window and grid sizes ($P = G = 7$) are utilized. The output of the grid-based self-attention step is also passed into an FFN for further representation learning.

It is worth noting that the pre-normalized relative self-attention mechanism [28] is applied in the block attention and grid attention modules. The pre-normalized relative self-attention is a variant of self-attention that combines absolute positional encodings (standard positional encodings used in the Transformer) and learned relative positional biases before softmax normalization. The learned biases allow the model to attend differently to tokens that are at different relative distances from each other within the sequence. The pre-normalized relative self-attention is defined as:

$$y_i^{\text{pre}} = \sum_{j \in \mathcal{G}} \frac{\exp\left(x_i^\top x_j + w_{i-j}\right)}{\sum_{k \in \mathcal{G}} \exp\left(x_i^\top x_k + w_{i-k}\right)} x_j \tag{2}$$

where for each token at position i, it computes a weighted sum of the embeddings of all other tokens in the global spatial space $\mathcal{G}$ based on their pairwise relationships and learned positional biases w_{i-j}. This mechanism captures how different tokens interact and contribute to the representation of the token at position i in the context of the entire sequence.

Residual connections are also integrated into the MaxViT blocks. These residual connections allow the module to learn residual representations, which are essentially the differences between the original input and the output of the self-attention and feedforward operations. By adding these residuals back to the output, the model can effectively fine-tune the learned representations and mitigate the challenges associated with vanishing gradients during training.

C. MEL SPECTROGRAM USING SHORT TIME FOURIER TRANSFORM (MEL-STFT)

The Mel-STFT is another technique used in audio signal processing. It provides a visual representation of how the frequency content of a signal changes over time, enabling the analysis of the spectral characteristics of an audio signal. To generate a Mel-STFT, a sequence of steps is undertaken. Firstly, the audio signal is divided into overlapping frames of fixed duration. For each frame, the Fast Fourier Transform (FFT) is applied, which transforms the signal from the time domain to the frequency domain, providing a snapshot of its spectral components at that moment. The magnitude of the FFT output represents the energy present in different frequency bands.

To align the frequency representation, the linear frequency values obtained from the FFT are converted into the Mel scale using the formula:

$$M(f) = 2595 \times \log_{10}(1 + \frac{f}{700}) \tag{3}$$

where $M(f)$ is the frequency in Mel, and f is the linear frequency in Hertz. This transformation accounts for the nonlinear way in perceiving frequencies. Following the Mel

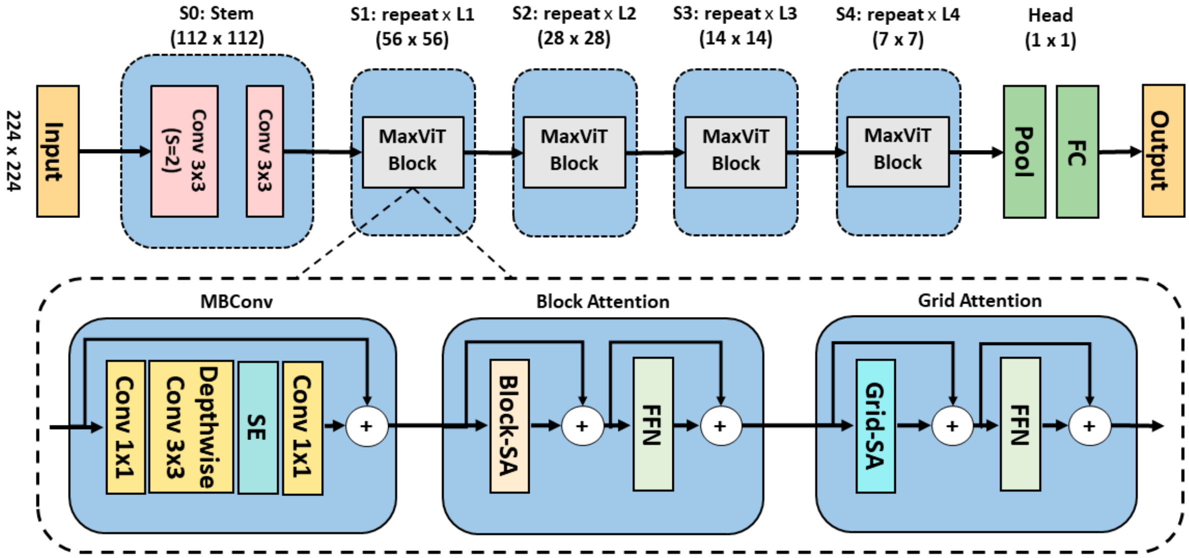


FIGURE 3. Model architecture of MaxViT.

scale conversion, a set of Mel filterbanks is applied to the transformed frequency values. Each filterbank is shaped as a triangle on the Mel scale and is used to capture the energy in a specific frequency range. The filterbank outputs are computed by multiplying the Mel spectrum with each triangular filterbank function. The logarithm of the energy values from the filterbanks is then calculated to approximate the logarithmic perception of loudness. This yields the Mel Spectrogram, a time-frequency representation where each pixel corresponds to the energy content of a specific frequency band within a given time frame. A frame length of 4096 samples and a hop size of 256 samples are applied in this work. An example of the Mel-STFT spectrogram is illustrated in Figure 4.

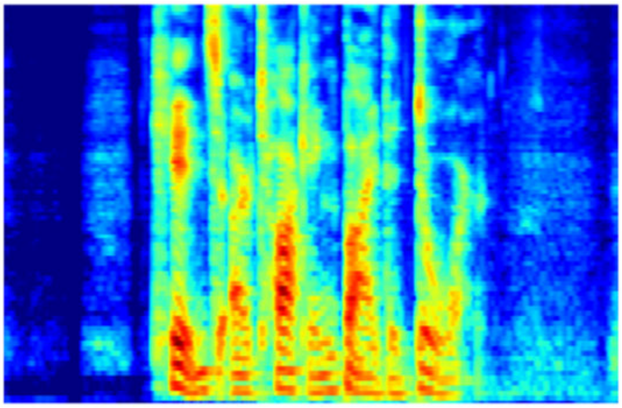


FIGURE 4. Mel-STFT spectrogram for the utterance 'Kids are talking by the door' from the RAVDESS dataset.

D. IMPROVED MULTISCALE VISION TRANSFORMERS (MVITV2)

The Mel-STFT spectrograms are channeled into the MViTv2 for representation learning. In contrast to Vision Transformers [29] characterized by a fixed channel capacity and spatial resolution across the network, Multiscale Vision Transformers (MViT) [30] introduce multiple stages with different channel-resolution scales. During the transition from input to output stages, MViT gradually expands the channel width while reducing resolution. Consequently, this construct forms a multiscale pyramid of feature maps within the transformer network. The initial layers can operate at high spatial resolution, modeling low-level visual information due to the lighter channel capacity. Conversely, deeper layers can focus on spatially coarse yet complex high-level features, effectively modeling visual semantics. Figure 5 depicts the architecture of MViTv2.

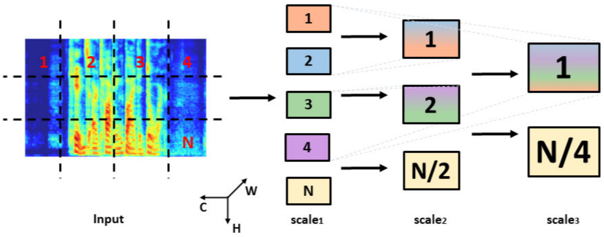


FIGURE 5. Model architecture of MViTv2.

To facilitate downsampling within a transformer block, MViT introduces Pooling Attention, a mechanism incorporating linear projections and pooling operators for query (Q), key (K), and value (V) tensors. The operations is defined as:

$$Q = \mathcal{P}_Q(XW_Q)$$
$$K = \mathcal{P}_K(XW_K)$$
$$V = \mathcal{P}_V(XW_V)$$

where $X \in \mathbb{R}^{L\times D}$ represents the input sequence with a sequence length of L and channel width of D, and W_Q, W_K, and W_V are the linear projections for Q, K, and V

tensors, respectively. Additionally, $\mathcal{P}_Q$, $\mathcal{P}_K$, and $\mathcal{P}_V$ denote the pooling operators for Q, K, and V tensors. The attention in MViT is computed by:

$$Z_1 := \mathrm{Attn}_1(Q, K, V) = \mathrm{Softmax}\left(QK^\top/\sqrt{D}\right)V \tag{4}$$

The Improved Multiscale Vision Transformers (MViTv2) [31] proposes some enhancements over MViT. Firstly, MViT relies solely on "absolute" positional embeddings to convey location information, which is less robust to shift-invariance. To rectify this, MViTv2 introduces relative positional embeddings that capture the relative location between input tokens. The computation of relative positional embeddings $R_{p(i),p(j)} \in \mathbb{R}^d$ of element i and j is decomposed as follows:

$$R_{p(i),p(j)} = R^{\mathrm{h}}_{h(i),h(j)} + R^{\mathrm{w}}_{w(i),w(j)} \tag{5}$$

where $p(i)$ and $p(j)$ are the spatial position of element i and j, R^{h} and R^{w} are the positional embeddings along the height and width axes, and $h(i)$, $h(j)$, $w(i)$, and $w(j)$ denote the vertical and horizontal positions of tokens i and j, respectively. These positional embeddings are then integrated into the self-attention module as follows:

$$\mathrm{Attn}_2(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top + E^{(\mathrm{rel})}}{\sqrt{d}}\right)V$$
$$\text{where } E^{(\mathrm{rel})}_{ij} = Q_i \cdot R_{p(i),p(j)}$$

Secondly, MViTv2 employs a residual pooling connection with the pooled Q tensor, enhancing information flow and reducing potential information loss during pooling attention. This technique maintains low-complexity attention computation with large strides in key (K) and value (V) pooling, thus improving overall model efficiency. The residual pooling connection can be formulated as:

$$Z_2 := \mathrm{Attn}_2(Q, K, V) + Q \tag{6}$$

The improved pooling attention is illustrated in Figure 6. The MViTv2 model integrates features from multiple scales, enabling the model to capture information at varying granularity levels. Additionally, MViTv2 utilizes the decomposed relative position embedding and residual pooling connection to preserve essential information at a lower computational cost.

In addition, the channel dimension expansion, previously located within the last MLP block of MViT's preceding stage, has been replaced with the attention computation embedded in the initial transformer block of each stage. This adjustment maintains a comparable level of accuracy while significantly reducing the model's parameter count and its floating point operations per second (FLOPs). Subsequently, the output tokens from the last transformer block are averaged, and the final classification head is employed, replacing the default class token in MViT. These modifications have led to a shortened training time and a reduction in computational resource demands.

1) MULTILAYER PERCEPTRON (MLP)

The feature maps of MaxViT and MViTv2 are concatenated and directed into the MLP for representation learning and classification. The MLP comprises a series of key components, including a dense layer, a batch normalization layer, a dropout layer, and a classification layer.

The dense layer plays a central role by applying a linear transformation to the concatenated feature maps, enabling the capture of intricate patterns within the data. Simultaneously, the batch normalization layer normalizes the activations within mini-batches, enhancing training stability, expediting convergence, and mitigating overfitting. To prevent overfitting and encourage robust feature learning, the dropout layer selectively deactivates a fraction of neurons during training.

In the final stage, the classification layer computes the probabilities associated with speech emotion classes through the softmax function. This transformation converts raw scores into a probability distribution, resulting in the ultimate prediction. The dense layer in this study comprises 512 hidden units, and a dropout rate of 0.2 is employed to achieve the desired balance between feature learning and regularization.

E. DATASETS

The proposed MaxMViT-MLP was evaluated on three publicly available 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 [27] comprises 535 audio samples collected from ten proficient German speakers, including five male and five female actors. These recordings encompass seven distinct emotions: anger, boredom, neutrality, happiness, anxiety, sadness, and disgust.

The RAVDESS [32] dataset consists of a diverse collection of 1440 audio samples recorded in English. The dataset showcases the performances of 24 professional actors, with 12 male and 12 female speakers. The samples in the RAVDESS dataset are categorized into eight classes: neutrality, calmness, happiness, sadness, anger, fear, disgust, and surprise.

The IEMOCAP [33] comprises 5507 English-language audio samples recorded by ten professional actors, with five male and five female speakers. There are four emotion classes commonly used by existing works, namely neutrality, happiness, anger, and sadness.

IV. EXPERIMENTS AND ANALYSIS

In the experiments, the datasets were divided into an 80% training set and a 20% testing set to facilitate a systematic comparison with existing works. To ensure uniformity, the data samples were resampled to a frequency of 44.1kHz. Subsequently, the samples underwent transformation into both CQT and Mel-STFT spectrograms, which were then

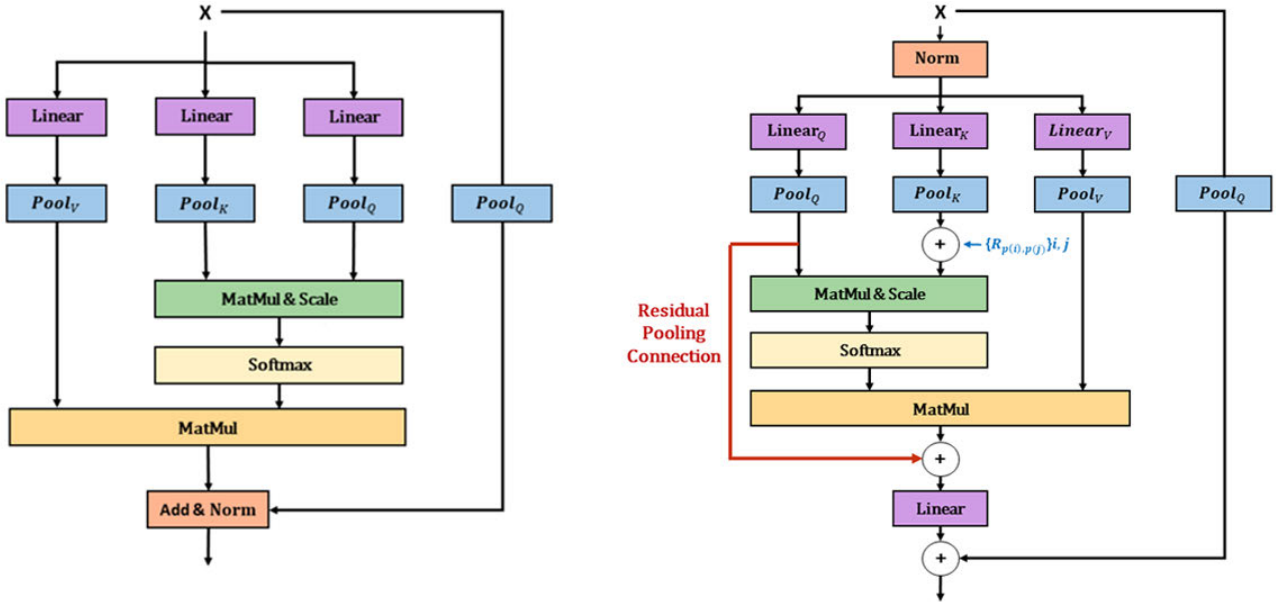


FIGURE 6. Pooling attention of MViT (left) and residual pooling attention of MViTv2 (right).

resized to 244 × 244 pixels in compliance with the input size requirements of MaxViT and MViTv2.

A. HYPERPARAMETER TUNING

Hyperparameter tuning was conducted to determine the optimal settings for the proposed MaxMViT-MLP. The process involved tuning five hyperparameters: the optimizer (O_1) and learning rate (R_1) for MaxViT, the optimizer (O_2) and learning rate (R_2) for MViTv2, and the number of hidden nodes in the MLP (N). The grid search mechanism was employed, where the experiments encompassed a range of values for each hyperparameter, allowing for a comprehensive examination of the settings. Table 2 provides a summary of the hyperparameter settings for MaxMViT-MLP.

TABLE 2. Summary of hyperparameter settings for MaxMViT-MLP.

Hyperparameter	Measured Values	Optimal Value
Optimizer (O_1)	Adam, RAdam, QHAdam	Adam
Learning Rate (R_1)	0.01, 0.02, 0.03	0.02
Optimizer (O_2)	Adam, RAdam, QHAdam	RAdam
Learning Rate (R_2)	0.01, 0.02, 0.03	0.02
Hidden Nodes (N)	128, 256, 512, 1024	512

The results in Table 3 showcase the performance of different optimizers on MaxViT, O_1. Three optimizers, namely Adam, Rectified Adam optimizer (RAdam), and Quasi-Hyperbolic Adam (QHAdam), were evaluated on the datasets. The Adam optimizer demonstrated superior performance with accuracy rates of 95.28%, 89.12%, and 68.39% for Emo-DB, RAVDESS, and IEMOCAP, respectively. Adam is known for its adaptive learning rate mechanism, combining the benefits of both momentum and root mean square propagation. This adaptability allows Adam to efficiently navigate complex optimization landscapes, contributing to its success in enhancing the MaxViT model's accuracy. RAdam incorporates a rectified term in its adaptive learning rate, enhancing robustness during training. Although slightly outperformed by Adam, RAdam's results underscore its effectiveness in optimizing MaxViT. QHAdam introduces quasi-hyperbolic terms to Adam's optimization strategy, striking a balance between stability and adaptability. Despite a lower accuracy than Adam, QHAdam's performance highlights its relevance as a competitive optimizer for the MaxViT model.

TABLE 3. Experimental results of different optimizers of MaxViT, O_1 [$R_1 = 0.02$, $O_2 =$ RAdam, $R_2 = 0.02$, $N = 512$].

Optimizer, O_1	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
Adam	**95.28**	**89.12**	**68.39**
RAdam	92.45	85.96	65.76
QHAdam	90.57	88.77	67.48

The results displayed in Table 4 provide insights into the impact of different learning rates on the performance of MaxViT. A learning rate of 0.02 emerges as the optimal choice, yielding the highest accuracy rates for the datasets. This finding suggests that the selected learning rate strikes a balance between the model's convergence speed and stability during training. A learning rate that is too low may hinder convergence, while one that is too high can lead to overshooting and instability. The superior performance at the 0.02 learning rate indicates its effectiveness in guiding the optimization process for MaxViT under the specified hyperparameter configuration. Comparatively, a learning rate of 0.01 produces slightly lower accuracy across the three datasets, emphasizing the sensitivity of the model to variations in the learning rate. Meanwhile, a learning rate of 0.03 exhibits a similar trend, suggesting that higher learning rates may lead to suboptimal convergence.

TABLE 4. Experimental results of different learning rates of MaxViT, R_1 [O_1 = Adam, O_2 = RAdam, R_2 = 0.02, N = 512].

Learning Rate, R_1	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
0.01	92.45	85.96	63.85
0.02	**95.28**	**89.12**	**68.39**
0.03	93.40	87.72	63.49

Table 5 presents experimental results showcasing the impact of different optimizers on the performance of MaxMViT-MLP. The incorporation of RAdam in MViTv2 achieves the highest accuracy rates on the datasets, demonstrating superior performance compared to both Adam and QHAdam. RAdam enhances the standard Adam optimizer by introducing a rectification term in its adaptive learning rate mechanism. This modification contributes to the optimizer's stability during training, preventing potential convergence issues and improving overall optimization efficiency.

TABLE 5. Experimental results of different optimizers of MViTv2, O_2 [O_1 = Adam, R_1 = 0.02, R_2 = 0.02, N = 512].

Optimizer, O_2	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
Adam	93.40	88.42	67.30
RAdam	**95.28**	**89.12**	**68.39**
QHAdam	94.34	86.32	60.67

Table 6 displays experimental results elucidating the influence of different learning rates on the performance of MViTv2. A learning rate of 0.02 emerges as the most effective choice, yielding the highest accuracy rates across the datasets. This finding suggests that the specified learning rate strikes a delicate balance, allowing for a harmonious convergence of the optimization process.

TABLE 6. Experimental results of different learning rates of MViTv2, R_2 [O_1 = Adam, R_1 = 0.02, O_2 = RAdam, N = 512].

Learning Rate, R_2	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
0.01	92.45	87.72	67.39
0.02	**95.28**	**89.12**	**68.39**
0.03	91.51	84.91	62.13

The experimental findings in Table 7 highlight the critical role of selecting an optimal number of hidden nodes of MLP. Among the tested configurations, the number of hidden nodes set at 512 emerges as the optimal choice. This finding underscores the impact of the hidden layer's capacity on the model's ability to capture complex patterns within the data. A hidden layer with 512 nodes strikes an effective balance, providing sufficient representational capacity without introducing excessive complexity that might lead to overfitting.

TABLE 7. Experimental results of different hidden nodes of MLP, N [O_1 = Adam, R_1 = 0.02, O_2 = RAdam, R_2 = 0.02].

Hidden Nodes, N	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
128	84.91	88.07	63.67
256	93.40	88.42	64.03
512	**95.28**	**89.12**	**68.39**
1024	94.34	87.37	64.21

B. ABLATION STUDY

The ablation study presented in Table 8 systematically explores the impact of different configurations on the performance of the proposed MaxMViT-MLP model across the Emo-DB, RAVDESS, and IEMOCAP datasets. Two spectrogram representations, CQT and Mel-STFT, are individually combined with two transformer architectures, MaxViT and MViTv2, to assess their standalone effectiveness.

Firstly, employing CQT with MaxViT yields an accuracy of 85.85% on Emo-DB, 77.54% on RAVDESS, and 62.49% on IEMOCAP. Similarly, Mel-STFT with MViTv2 produces competitive results with accuracy rates of 88.68%, 77.89%, and 62.85% across the three datasets. These specific combinations are chosen based on their initial promise and individual strengths.

Furthermore, combining both CQT + MaxViT and Mel-STFT + MViTv2 leads to a notable improvement, achieving accuracy rates of 91.51%, 84.91%, and 67.85%. This amalgamation leverages the complementary features of CQT and Mel-STFT, enhancing the model's ability to capture diverse spectral characteristics present in speech signals.

Finally, the addition of the MLP layer to the amalgamated configuration (CQT + MaxViT + Mel-STFT + MViTv2 + MLP) results in the best performance, reaching accuracy rates of 95.28%, 89.12%, and 68.39% across the three datasets. The MLP introduces non-linearity and further refines the model's representation capabilities, demonstrating its crucial role in achieving optimal performance in the context of speech emotion recognition. The stepwise progression in accuracy highlights the cumulative improvement obtained by incorporating diverse components, underscoring the effectiveness of the proposed MaxMViT-MLP model.

C. COMPARISON RESULTS WITH EXISTING METHODS

Table 9 presents the performance of various methods for speech emotion recognition on the Emo-DB. Traditional

TABLE 8. Ablation study across the Emo-DB, RAVDESS, and IEMOCAP datasets.

Configuration	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
CQT + MaxViT	85.85	77.54	62.49
CQT + MViTv2	78.30	73.33	61.76
Mel-STFT + MaxViT	81.13	75.79	62.13
Mel-STFT + MViTv2	88.68	77.89	62.85
CQT + MaxViT + Mel-STFT + MViTv2	91.51	84.91	67.85
CQT + MaxViT + Mel-STFT + MViTv2 + MLP	**95.28**	**89.12**	**68.39**

TABLE 9. Comparison results on the Emo-DB.

Methods	Accuracy (%)
CapCNN [1]	82.90
2D-CNN [2]	88.00
CNN-BiLSTM [3]	86.87
MFMC with SVM [4]	81.50
CNN [5]	76.40
F-ScatNet with SVM [6]	74.59
ScatNet with SVM [6]	74.40
MFCC with SVM [6]	58.39
INCA with SVM [7]	90.09
i-vector with SVM [8]	85.75
i-vector with PLDA [8]	82.84
i-vector with RF [8]	82.10
Hybrid LSTM [9]	85.55
CNN-VGG16 [10]	92.79
ResBLSTMA [11]	90.57
ResBLSTM [11]	79.25
Conv2D-LSTM [12]	65.69
MFMC with DNN [15]	84.72
MFMC with CNN [15]	82.41
MFCC with DNN [15]	82.24
MFCC with CNN [15]	81.31
Dual-TBNet [18]	84.10
Emo-LGBM [25]	84.91
DNN-SVM [26]	79.86
MaxMViT-MLP (Proposed)	**95.28**

methods often rely on handcrafted feature extraction and machine learning models. For instance, MFCC with SVM attains a modest accuracy of 58.39%, indicating the limitations of traditional feature-based methods in this task. i-vector techniques, such as i-vector with SVM (85.75%) and i-vector with PLDA (82.84%), display more competitive results, highlighting the importance of feature engineering in achieving accurate emotion recognition. However, these traditional methods are outperformed by the deep learning models in this evaluation.

Deep learning methods, including models like CNN-VGG16 and Hybrid LSTM, demonstrate their effectiveness in capturing emotional cues from audio data. CNN-VGG16 achieves an impressive accuracy of 92.79%, emphasizing the power of deep neural networks in extracting relevant emotional features. Moreover, the proposed MaxMViT-MLP model emerges as the top performer, setting a new benchmark with an accuracy of 95.28%.

Moving to the RAVDESS dataset, a comparison of results with existing methods is presented in Table 10. Traditional methods, as evidenced by the results, yield a mixed bag of outcomes. MFCC with SVM delivers an accuracy of 36.74%, underscoring the complexities encountered when applying traditional feature-based approaches to this dataset. Similarly, F-ScatNet with SVM (51.81%) and ScatNet with SVM (50.00%) produce suboptimal results, underscoring the limitations of classical signal processing techniques when dealing with RAVDESS data.

In contrast, deep learning models exhibit a more robust performance in addressing the intricacies of the RAVDESS dataset. Models like ResBLSTMA (84.50%) and ResBLSTM (85.41%) showcase strong performance, highlighting the effectiveness of recurrent neural networks in this specific context. Furthermore, the CNN model (87.08%) and MFMC with DNN (76.72%) yield competitive results, emphasizing the adaptability of deep learning architectures in the domain of emotion recognition when dealing with the RAVDESS dataset.

The proposed MaxMViT-MLP method maintains its exceptional performance, achieving an accuracy of 89.12% on the RAVDESS dataset. This reinforces the model's capacity to generalize effectively across different datasets and underscores its efficacy as a versatile solution for emotion recognition, regardless of the distinctive characteristics of the dataset at hand.

The performance of various emotion recognition methods on the IEMOCAP dataset is presented in Table 11. In the context of the IEMOCAP dataset, traditional approaches, such as MFCC with SVM (55.54%), ScatNet with SVM (60.41%), F-ScatNet with SVM (61.55%), Emo-LGBM (62.94%), i-vector with RF (63.57%), i-vector with PLDA (64.52%), and i-vector with SVM (65.78%), exhibit moderate accuracy. These results highlight the challenges of utilizing traditional feature-based approaches for emotion recognition on the IEMOCAP dataset, which contains complex and dynamic emotional expressions. Likewise, the deep learning models, such as Fusion features with MHA-CRNN (66.44%) and Dual-TBNet (64.80%), demonstrate moderate accuracy. Fusion features with MHA-CRNN method achieves the highest accuracy among the existing methods, emphasizing the strength of combining multiple modalities and context-aware modeling. The proposed MaxMViT-MLP model stands out as the top-performing method, setting a new benchmark with an accuracy of 68.39%.

TABLE 10. Comparison results on the RAVDESS dataset.

Methods	Accuracy (%)
MFMC with SVM [4]	64.31
CNN [5]	70.80
F-ScatNet with SVM [6]	51.81
ScatNet with SVM [6]	50.00
MFCC with SVM [6]	36.74
INCA with SVM [7]	87.42
Hybrid LSTM [9]	75.62
ResBLSTMA [11]	84.50
ResBLSTM [11]	85.41
Conv2D-LSTM [12]	46.69
MLP [13]	66.67
CNN [14]	87.08
MFMC with DNN [15]	76.72
MFMC with CNN [15]	72.90
MFCC with DNN [15]	73.26
MFCC with CNN [15]	68.54
Emo-LGBM [25]	67.72
DNN-SVM [26]	52.24
MaxMViT-MLP (Proposed)	**89.12**

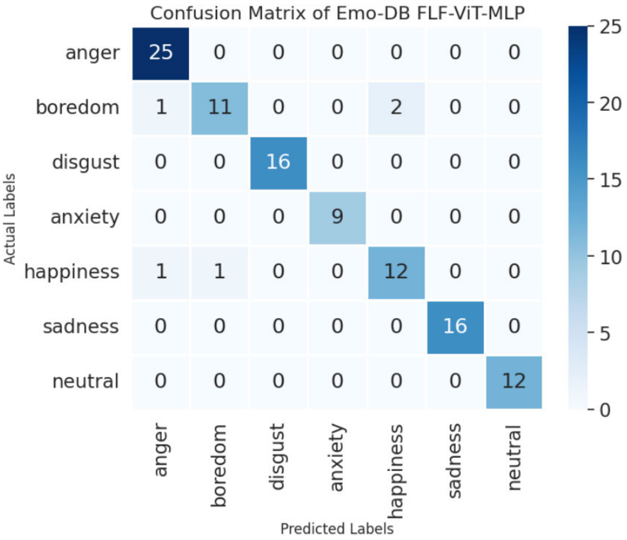
TABLE 11. Comparison results on the IEMOCAP dataset.

Methods	Accuracy (%)
F-ScatNet with SVM [6]	61.55
ScatNet with SVM [6]	60.41
MFCC with SVM [6]	55.54
i-vector with SVM [8]	65.78
i-vector with PLDA [8]	64.52
i-vector with RF [8]	63.57
Conv2D-LSTM [12]	54.83
MLP [13]	57.09
syl-SVM [16]	62.90
Fusion features with MHA-CRNN [17]	66.44
Dual-TBNet [18]	64.80
Emo-LGBM [25]	62.94
MaxMViT-MLP (Proposed)	**68.39**

The proposed MaxMViT-MLP model harnesses the strengths of two audio feature representation techniques: CQT and Mel-STFT. CQT excels in capturing tonal characteristics through its logarithmic frequency binning, closely mimicking human perception of pitch and timbre. The constant-Q resolution of CQT provides detailed information in the low-frequency range, making it well-suited for tasks that demand harmonic analysis. In contrast, Mel-STFT employs a linear frequency scale, enabling the capture of broader spectral features that align with human auditory interpretation. This approach yields rich time-frequency representations, particularly valuable in tasks involving transient events, such as speech recognition and environmental sound classification. By combining CQT and Mel-STFT, the proposed model assembles a comprehensive set of features for audio analysis, capitalizing on their complementary abilities to capture diverse aspects of audio signals.

As for the feature extraction, the MaxMViT-MLP model employs two Vision Transformers: MaxViT and MViTv2. The hierarchical architecture of MaxViT effectively captures multiscale information, utilizing advanced features like the block attention module to foster contextual understanding, making it proficient in recognizing intricate audio patterns and relationships within CQT spectrograms. In contrast, MViTv2 provides a range of multiscale features with different channel-resolution scales. Its incorporation of relative positional embeddings enhances its understanding of spatial relationships, rendering it a suitable choice for the analysis of Mel-STFT spectrograms. The subsequent Multilayer Perceptron (MLP) further enhances the model's representation learning capabilities.

Figure 7 presents the confusion matrix for the MaxMViT-MLP method applied to the Emo-DB dataset. Notably, the anger, disgust, and neutral classes achieve impeccable classification accuracy, while occasional misclassifications occur between the boredom and happiness classes. Figure 8 displays the confusion matrix for the MaxMViT-MLP method on the RAVDESS dataset. In RAVDESS, high misclassification rates are evident in the disgust, sadness, and fear classes. These challenges arise from the intra-class acoustic similarities and variability in speech emotion expressions among different speakers.



	anger	boredom	disgust	anxiety	happiness	sadness	neutral
anger	25	0	0	0	0	0	0
boredom	1	11	0	0	2	0	0
disgust	0	0	16	0	0	0	0
anxiety	0	0	0	9	0	0	0
happiness	1	1	0	0	12	0	0
sadness	0	0	0	0	0	16	0
neutral	0	0	0	0	0	0	12

FIGURE 7. Confusion matrix of the Emo-DB.

For the IEMOCAP dataset, the confusion matrix in Figure 9 introduces additional complexities due to the presence of dual speakers in conversations. The analysis of the confusion matrix reveals that the happiness and sadness classes exhibit the highest misclassification rates. This may

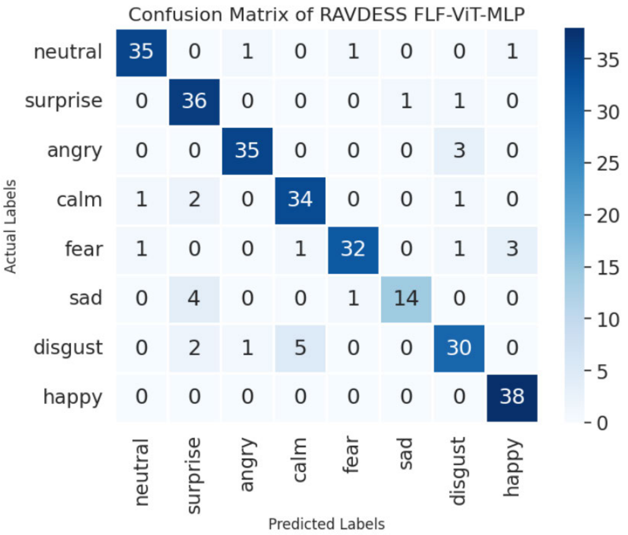


FIGURE 8. Confusion matrix of the RAVDESS dataset.

be attributed to instances where individuals express multiple emotions simultaneously or swiftly transition between emotions, making it challenging for the model to assign a single, accurate label.

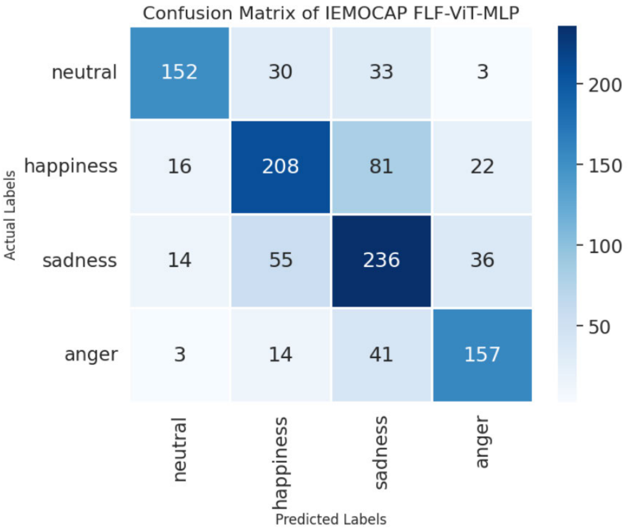


FIGURE 9. Confusion matrix of the IEMOCAP dataset.

V. CONCLUSION

This paper introduces a dual-path speech emotion recognition model, referred to as the "MaxMViT-MLP", capitalizing on two potent audio feature representations, CQT and Mel-STFT. CQT's proficiency in capturing tonal characteristics and its suitability for harmonic analysis complements Mel-STFT's ability to record broader spectral features, ideal for transient event analysis. The fusion of these representations results in a versatile feature set, contributing significantly to the model's performance.

The adoption of MaxViT and MViTv2, further enhances the model's capacity to process CQT and Mel-STFT spectrograms. MaxViT effectively captures multiscale information and intricate patterns within audio data, while MViTv2 excels in handling multiscale features with different channel-resolution scales. This collaborative approach between the models plays a pivotal role in the model's robustness and adaptability. The experimental results across three datasets: Emo-DB, RAVDESS, and IEMOCAP, affirm the outstanding performance of the MaxMViT-MLP model, achieving accuracy rates of 95.28% on Emo-DB, 89.12% on RAVDESS, and 68.39% on IEMOCAP. The harmonious blend of CQT and Mel-STFT, coupled with the strengths of MaxViT and MViTv2, underpins the model's remarkable results.

Looking ahead, future work could explore alternative transformer-based models that might offer valuable insights to enhance performance in speech emotion recognition. Continuous efforts in refining and extending the proposed MaxMViT-MLP model will contribute to advancing the field of speech emotion recognition and its broader applications. Additionally, investigating the effectiveness of the proposed model in real-world scenarios and diverse cultural contexts would contribute to its practical applicability.

REFERENCES

[1] X.-C. Wen, K.-H. Liu, W.-M. Zhang, and K. Jiang, "The application of capsule neural network based CNN for speech emotion recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 9356–9362.

[2] A. Mujaddidurrahman, F. Ernawan, A. Wibowo, E. A. Sarwoko, A. Sugiharto, and M. D. R. Wahyudi, "Speech emotion recognition using 2D-CNN with data augmentation," in *Proc. Int. Conf. Softw. Eng. Comput. Syst. 4th Int. Conf. Comput. Sci. Inf. Manage. (ICSECS-ICOCSIM)*, Aug. 2021, pp. 685–689.

[3] J. He and L. Ren, "Speech emotion recognition using XGBoost and CNN BLSTM with attention," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI)*, Oct. 2021, pp. 154–159.

[4] J. Ancilin and A. Milton, "Improved speech emotion recognition with mel frequency magnitude coefficient," *Appl. Acoust.*, vol. 179, Aug. 2021, Art. no. 108046.

[5] M. H. Pham, F. M. Noori, and J. Torresen, "Emotion recognition using speech data with convolutional neural network," in *Proc. IEEE 2nd Int. Conf. Signal, Control Commun. (SCC)*, Dec. 2021, pp. 182–187.

[6] P. Singh, G. Saha, and M. Sahidullah, "Deep scattering network for speech emotion recognition," 2021, *arXiv:2105.04806*.

[7] T. Tuncer, S. Dogan, and U. R. Acharya, "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Knowl.-Based Syst.*, vol. 211, Jan. 2021, Art. no. 106547.

[8] R. Thirumuru, K. Gurugubelli, and A. K. Vuppala, "Novel feature representation using single frequency filtering and nonlinear energy operator for speech emotion recognition," *Digit. Signal Process.*, vol. 120, Jan. 2022, Art. no. 103293.

[9] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Hybrid LSTM-transformer model for emotion recognition from speech audio files," *IEEE Access*, vol. 10, pp. 36018–36027, 2022.

[10] D. H. Rudd, H. Huo, and G. Xu, "Leveraged mel spectrograms using harmonic and percussive components in speech emotion recognition," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2022, pp. 392–404.

[11] S. Kakuba and D. S. Han, "Residual bidirectional LSTM with multi-head attention for speech emotion recognition," in *Proc. Korea Commun. Assoc. Summer Gen. Acad. Conf.*, 2022, pp. 1419–1421.

[12] P. Singh, S. Waldekar, M. Sahidullah, and G. Saha, "Analysis of constant-Q filterbank based representations for speech emotion recognition," *Digit. Signal Process.*, vol. 130, Oct. 2022, Art. no. 103712.

[13] L. Vu, R. C.-W. Phan, L. W. Han, and D. Phung, "Improved speech emotion recognition based on music-related audio features," in *Proc. 30th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2022, pp. 120–124.

[14] S. Sekkate, M. Khalil, and A. Adib, "A statistical feature extraction for deep speech emotion recognition in a bilingual scenario," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 11443–11460, Mar. 2023.

[15] S. P. Mishra, P. Warule, and S. Deb, "Deep learning based emotion classification using mel frequency magnitude coefficient," in *Proc. 1st Int. Conf. Innov. High Speed Commun. Signal Process. (IHCSP)*, Mar. 2023, pp. 93–98.

[16] A. Rehman, Z.-T. Liu, M. Wu, W.-H. Cao, and C.-S. Jiang, "Speech emotion recognition based on syllable-level feature extraction," *Appl. Acoust.*, vol. 211, Aug. 2023, Art. no. 109444.

[17] Z. Tu, B. Liu, W. Zhao, R. Yan, and Y. Zou, "A feature fusion model with data augmentation for speech emotion recognition," *Appl. Sci.*, vol. 13, no. 7, p. 4124, Mar. 2023.

[18] Z. Liu, X. Kang, and F. Ren, "Dual-TBNet: Improving the robustness of speech features via dual-transformer-BiLSTM for speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2193–2203, 2023.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 1459–1462.

[20] A. T. Liu, S.-W. Li, and H.-Y. Lee, "TERA: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2351–2366, 2021.

[21] P. Chi, P. Chung, T. Wu, C. Hsieh, Y. Chen, S. Li, and H. Lee, "Audio Albert: A lite BERT for self-supervised learning of audio representation," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 344–350.

[22] A. H. Liu, Y.-A. Chung, and J. Glass, "Non-autoregressive predictive coding for learning speech representations from local dependencies," 2020, *arXiv:2011.00406*.

[23] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," 2019, *arXiv:1904.05862*.

[24] A. Baevski, S. Schneider, and M. Auli, "Vq-wav2vec: Self-supervised learning of discrete speech representations," 2019, *arXiv:1910.05453*.

[25] 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.

[26] P. Singh, M. Sahidullah, and G. Saha, "Modulation spectral features for speech emotion recognition using deep neural networks," *Speech Commun.*, vol. 146, pp. 53–69, Jan. 2023.

[27] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-axis vision transformer," in *Proc. 17th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 459–479.

[28] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2021, pp. 3965–3977.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[30] 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. 6824–6835.

[31] 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. 4804–4814.

[32] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Sep. 2005, pp. 1517–1520.

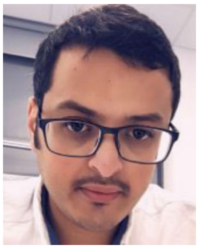
[33] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.

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 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 (abnormal behavior detection and gait recognition). She is currently an Associate Professor 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. She holds the status of a Certified Professional Technologist. She has been a member of the International Association of Engineers and serves as an Outcome-Based Education Consultant and Trainer.

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 an 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 Master of Engineering Science (M.Eng.Sc.) and Ph.D. (I.T.) degrees from Multimedia University. He is currently an Associate Professor with the Faculty of Information Science and Technology, Multimedia University. His research 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 conferences and journal publications. His research interests include aspects of pattern recognition, deep learning, and machine intelligence and their applications to real-world problems.