Received 17 May 2023, accepted 19 June 2023, date of publication 22 June 2023, date of current version 27 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3288526

RESEARCH ARTICLE

SCQT-MaxViT: Speech Emotion Recognition With Constant-Q Transform and Multi-Axis Vision Transformer

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 TAKEKI MUKAIDA[3]

[1]Faculty of Information Science and Technology, Multimedia University, Malacca 75450, Malaysia
[2]Faculty of Engineering and Technology, Multimedia University, Malacca 75450, Malaysia
[3]School of Informatics and Engineering, University of Electro-Communications, Chofu, Tokyo 182-8585, Japan

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

This work was supported in part by the Fundamental Research Grant Scheme of the Ministry of Higher Education under Award FRGS/1/2021/ICT02/MMU/02/4, and in part by the Telekom Malaysia Research and Development under Grant RDTC/231084.

ABSTRACT Speech emotion recognition presents a significant challenge within the field of affective computing, requiring the analysis and detection of emotions conveyed through speech signals. However, existing approaches often rely on traditional signal processing techniques and handcrafted features, which may not effectively capture the nuanced aspects of emotional expression. In this paper, an approach named "SCQT-MaxViT" is proposed for speech emotion recognition, combining signal processing, computer vision, and deep learning techniques. The method utilizes the Constant-Q Transform (CQT) to convert speech waveforms into spectrograms, providing high-frequency resolution and enabling the model to capture intricate emotional details. Additionally, the Multi-axis Vision Transformer (MaxViT) is employed for further representation learning and classification of the CQT spectrograms. MaxViT incorporates a multi-axis self-attention mechanism, facilitating both local and global interactions within the network and enhancing the ability of the model to learn meaningful features. Furthermore, the dataset is augmented using random time masking techniques to enhance the generalization capabilities. Achieving accuracies of 88.68% on the Emo-DB dataset, 77.54% on the RAVDESS dataset, and 62.49% on the IEMOCAP dataset, the proposed SCQT-MaxViT method exhibits promising performance in capturing and recognizing emotions in speech signals.

INDEX TERMS Speech, speech emotion, speech emotion recognition, spectrogram, constant-Q transform, vision transformer, multi-axis vision transformer, Emo-DB, RAVDESS, IEMOCAP.

I. INTRODUCTION

Speech emotion recognition is an interdisciplinary field that aims to identify and classify emotions in spoken language. Speech emotion recognition plays a crucial role in various applications, such as human-computer interaction, mental health monitoring, customer service, and virtual assistants. The growing interest in speech emotion recognition stems from the increasing demand for intelligent systems capable of understanding and adapting to human emotions, thus providing more natural and intuitive user experiences.

The rapid advancements in Artificial Intelligence and machine learning have garnered significant attention for speech emotion recognition from both researchers and practitioners. Numerous approaches have been proposed, ranging from traditional machine learning techniques like Support Vector Machines and Hidden Markov Models to more recent deep learning methods, such as Convolutional Neural Networks and Recurrent Neural Networks. These approaches depend on extracting relevant acoustic features, including

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

pitch, energy, and spectral characteristics, as well as employing effective classification techniques.

Despite substantial progress in speech emotion recognition, several challenges remain unaddressed. One key challenge is the variability and ambiguity of emotions, which often complicate the definition and annotation of emotional states in speech data. Moreover, factors such as speaker variability, language, and cultural differences heavily influence the performance of speech emotion recognition systems. Consequently, there is an ongoing need for more robust and generalizable models to tackle these challenges.

This paper presents an approach that leverages the Constant-Q Transform (CQT) to convert speech signals into a time-frequency representation, effectively capturing the non-stationary characteristics of speech. The CQT spectrogram offers several benefits, such as enhanced frequency resolution at lower frequencies and improved time resolution at higher frequencies, making it an ideal choice for speech processing tasks. To enhance the model's robustness and generalization capabilities, time masking is employed as a data augmentation technique. Time masking operates directly on the CQT spectrogram by applying masking to segments of the spectrogram. This augmentation method enables the model to learn more diverse and invariant features from the input data, thereby improving its performance on unseen data. Finally, a Multi-Axis Vision Transformer (MaxViT) is utilized for representation learning and classification. MaxViT combines the strengths of both the Vision Transformer and the multi-axis attention mechanism, allowing the model to capture both local and global contextual information from the CQT spectrogram. This results in a more expressive representation, leading to superior classification performance and generalization in the speech emotion recognition task. This paper presents the following key contributions:

- Representation of the speech waveforms as CQT spectrograms: The CQT spectrograms offer high-frequency resolution, facilitating the accurate representation of various frequency components present in speech waveforms. This capability enables the model to capture fine-grained details related to emotional expression, including variations in pitch and intonation.
- Augmentation of the diversity of the speech dataset through time masking: To mitigate overfitting and enhance the generalization capabilities of the model, the speech dataset is augmented using time masking. This technique involves randomly masking segments of the CQT spectrogram, thereby introducing variations and enhancing the model's ability to handle diverse audio signals.
- Utilization of MaxViT for the classification of CQT spectrograms: The MaxViT model is employed for the classification of CQT spectrograms in speech emotion recognition. MaxViT incorporates the blocked multi-axis self-attention, which enables both global and local interactions within the network, reducing computational complexity while preserving non-local information. The Max-SA module decomposes fully dense attention into block attention and grid attention, providing efficient and effective attention across the spatial dimensions of CQT spectrograms.

II. RELATED WORKS

The existing works in speech emotion recognition can be broadly categorized into: traditional machine learning models and deep learning models.

A. TRADITIONAL MACHINE LEARNING

Singh et al. [1] proposed a support vector machine (SVM) model for speech emotion recognition tasks with acoustic features. The researchers used two classifiers, SVM and recurrent neural network, to evaluate the performance of spectral features and prosodic features on the Emo-DB and RAVDESS datasets. They found that the combination of both feature representations with the SVM classifier achieved better results than the RNN classifier, with an accuracy of 86.36% and 64.15% on the Emo-DB and the RAVDESS datasets, respectively.

In the study by Liu et al. [2], a novel approach was presented to improve the accuracy of speech emotion recognition by combining formant characteristics feature extraction and phoneme type convergence. They extracted formant characteristics from speech data and clustered them into different phoneme types using a k-means classifier. The resulting phoneme types were then used to train random forest, k-nearest neighbors, and multi-layer perceptron (MLP) models for classification. The MLP classifier achieved the highest accuracy of 72.91% on the Emo-DB dataset, followed by the same MLP classifier on the RAVDESS dataset with 61.02% accuracy. The RF classifier achieved the best accuracy of 62.01% on the IEMOCAP dataset.

Ancilin and Milton [3] found that the mel frequency magnitude coefficient (MFMC) is a superior feature representation compared to other spectral features for speech emotion recognition. They extracted MFMC features from speech data by using the magnitude of Fourier transform, excluding the discrete cosine transformation (DCT) applied in MFCC. The proposed SVM method achieved an accuracy of 81.50% and 75.63% on the Emo-DB and RAVDESS datasets, respectively, which outperformed traditional MFCC methods.

Seknedy and Fawzi [4] conducted a comparative study of four machine learning classifiers for speech emotion recognition, using three different feature sets: the INTERSPEECH 2009 Emotion Challenge feature set (IS09), time-domain features, and frequency-domain features. They found that the combination of the second feature set with SVM classifier achieved the highest accuracy of 85.97% on the Emo-DB dataset. On the RAVDESS dataset, the combination of IS09 with SVM classifier resulted in the best accuracy of 70.56%, followed closely by the second feature set with SVM classifier, which achieved an accuracy of 70.42%.

Parra-Gallego and Orozco-Arroyave [5] proposed a soft-margin SVM with a Gaussian kernel for speech emotion recognition, using different feature approaches. These features included i-vectors, x-vectors, the I2010PC feature set, and phonation (pho), articulation (art), and prosody (pro) features. I-vectors represent low-dimensional vector representations also known as identity vectors. X-vectors represent embedding features obtained from deep neural networks, while the I2010PC feature set comprises 38 low-level descriptor representations. The phonation features include temporal changes in frequency (jitter), amplitude changes in the signal (shimmer), amplitude perturbation quotient (APQ), and pitch perturbation quotient (PPQ). The articulation features encode relevant information by representing the transition between voiced and unvoiced segments, while the prosody features represent prosodic features such as pitch, voice quality, intonation, accentuation, phrases, and rhythm. Based on the results, the combination of features (I2010PC, x-vector, art, pro, pho) achieved the highest accuracy of 80.70% and 63.80% on the Emo-DB and RAVDESS datasets, respectively. The best combination of features (I2010PC, x-vector) achieved an accuracy of 58.90% on the IEMOCAP dataset.

Singh et al. [6] proposed an approach for speech emotion recognition using a deep neural network with support vector machine (DNN-SVM) model. The proposed model employed the constant-Q transform based modulation spectral features (CQT-MSF) to process the speech data. First, the speech data was converted into CQT spectrogram, and then the CQT spectrogram was process to extract the temporal modulations. The DNN model was used to extract temporal feature representations, and the SVM model was used to classify emotions. The proposed DNN-SVM model with CQT-MSF was evaluated on two different datasets, Emo-DB and RAVDESS. The results showed remarkable accuracy of 79.86% and 52.24% on the Emo-DB and RAVDESS datasets, respectively.

In their recent study, Ong et al. [7] conducted speech emotion recognition by leveraging both frequency and temporal domain features. Their approach involved applying data augmentation techniques, specifically pitch shifting and time stretching, to the audio waveforms. Subsequently, the authors extracted seven distinct features from the augmented waveforms, including MFCC, Mel Spectrogram, Wavelet Transform, Kurtosis, Root Mean Square, Chroma, and Zero-Crossing Rate. These features were then fed into a LightGBM classifier for classification purposes. The experimental findings revealed that the proposed method achieved an accuracy of 84.91% on the Emo-DB dataset and 67.72% on the RAVDESS dataset.

B. DEEP LEARNING

Zhang et al. [8] proposed a discriminant temporal pyramid matching (DTPM) strategy with deep CNNs. The DTPM was designed to identify the most discriminative utterance-level representation from the Mel-frequency cepstral coefficients (MFCC) features. Experimental results on the Emo-DB dataset showed that the DCNN-DTPM method achieved an accuracy of 87.31%, demonstrating its effectiveness in speech emotion recognition.

Guo et al. [9] presented an improved speech emotion recognition approach by using a combination of complementary features and kernel extreme learning machine (KELM). The feature extraction involved both CNN and deep neural network (DNN) models to extract a combined set of features, namely MFCC, pitch, and voice quality representations from speech data. The output of these complementary features was then fed into the KELM for classification. To evaluate the proposed KELM method, the Emo-DB and IEMOCAP datasets were used. The results indicated that the proposed KELM approach achieved an accuracy of 84.49% and 57.10% on the Emo-DB and IEMOCAP datasets, respectively.

Jiang et al. [10] introduced a parallelized convolutional recurrent neural network (PCRN) to utilize the learning ability of neural networks. The PCRN is composed of a CNN and an LSTM, which split and process speech waveforms into two distinct feature representations in parallel. The CNN model learns the time and frequency representation, while the LSTM model learns the temporal features from the Mel-spectrograms features. In the end, the final output is obtained by combining the outputs of all the parallel branches. The proposed PCRN achieved 86.44% accuracy on the Emo-DB dataset.

In their study, Chatziagapi et al. [11] compared different data augmentation methods, including pitch shifting, time stretching, and GANs, to augment speech data for emotion recognition. The researchers evaluated the augmented data using the VGG19 model on the IEMOCAP dataset and found that the GAN-based data augmentation method outperformed the other methods, achieving a 54.60% accuracy.

Chauhan et al. [12] proposed a simple CNN architecture for speech emotion recognition, which takes log-mel spectrograms as input features. The model includes multiple convolutional layers, followed by pooling and fully connected layers. Each convolutional layer includes 2D convolutional layers, batch normalization, activation, and max-pooling layers. The proposed CNN model achieved accuracies of 72.02% and 59.33% on the Emo-DB and IEMOCAP datasets, respectively.

In a similar approach, Neumann and Vu [13] presented an unsupervised representation learning method to enhance speech emotion recognition performance on unlabeled speech data. The authors extracted 26 log mel-filterbanks as input features and applied the auDeep toolkit to learn additional feature vectors from the unlabeled speech data. The unsupervised representation was then fed into an attention convolutional neural network (ACNN) to learn the unlabeled speech data representation. The proposed method achieved an accuracy of 59.54% on the IEMOCAP dataset.

Seo and Kim [14] proposed a Visual Attention-based Convolutional Neural Network (VACNN) for speech emotion recognition, which uses the Bag of Visual Words (BoVW) technique to extract features from the log-mel spectrogram.

The VACNN model incorporates a visual attention mechanism to learn local and global feature representations by constructing a frequency histogram of visual words. This enables the model to capture both auditory and visual cues, resulting in improved accuracy in emotion recognition. The proposed method was evaluated on two datasets, Emo-DB and RAVDESS, achieving accuracies of 79.44% and 74.31%, respectively.

Yao et al. [15] employed a multi-task learning-based approach for speech emotion recognition, which combines three classifiers based on DNN, CNN, and RNN architectures. Each classifier is trained independently on mel-spectrogram features to recognize emotions, with the DNN model learning high-level spectral features, the CNN model learning mel-spectrogram features, and the RNN model learning low-level descriptors. The outputs of the classifiers are fused using a weighted sum to obtain the final emotion classification. The proposed method was evaluated on the IEMOCAP dataset and achieved an accuracy of 58.30%.

Singh et al. [16] presented an approach for speech emotion recognition using multidimensional convolutional neural networks with different feature representations. Three features are used namely, MFCC, Chroma MFCC, and Chroma. The performance of one-dimensional convolutional neural networks (1D-CNN) and two-dimensional convolutional neural networks (2D-CNN) with and without data augmentation were compared to evaluate the proposed approach. The data augmentation technique involved inserting noise into the speech data. The experiment reported that the 2D-CNN model with augmented log-mel spectrogram achieved an accuracy of 63.00%, while the 2D-CNN model without augmented MFCC features achieved 64.00% accuracy on the RAVDESS dataset.

Singh et al. [17] proposed a deep learning model that combines a 2D-CNN and a LSTM network enhanced with a self-attention mechanism. The proposed model leveraged MFCC features as input representations from speech data. The researchers evaluated their proposed model on the RAVDESS dataset, where the CNN-2D and LSTM with self-attention approach achieved an accuracy of 74.44%.

III. SPEECH EMOTION RECOGNITION WITH CONSTANT-Q TRANSFORM AND MULTI-AXIS VISION TRANSFORMER

The proposed method for speech emotion recognition, called SCQT-MaxViT, integrates the spectrogram of the constant-Q transform (CQT) with the Multi-Axis Vision Transformer (MaxViT). The method comprises three fundamental steps: signal representation, data augmentation, and classification.

In the signal representation step, the speech signals are transformed into CQT spectrograms. This representation captures the frequency content of the signals and provides valuable insights into the underlying emotions. The CQT spectrograms serve as the input data for subsequent processing. The CQT spectrograms are then divided into separate training and testing sets, enabling the model to learn from a labeled dataset and evaluate its performance on unseen data.

To enhance the quantity and diversity of the training data, the training set undergoes data augmentation. This augmentation process involves applying time masking transformations to the CQT spectrograms. Time masking introduces variations in the temporal structure of the signals, enriching the training dataset and improving the model's ability to generalize to different speech samples.

The augmented CQT spectrograms are then utilized to train the MaxViT model, which employs a multi-axis attention mechanism to learn powerful representations from the input data. This representation learning phase enables the model to capture complex patterns and discriminative features related to different emotional states.

Finally, the performance evaluation is conducted on the trained MaxViT model using the CQT spectrograms from the testing set. This evaluation assesses the model's ability to accurately classify and recognize emotions present in the speech signals. The workflow of the proposed SCQT-MaxViT method is illustrated in Figure 1.

A. CONSTANT-Q TRANSFORM

The Constant-Q Transform (CQT) is a frequency-domain analysis technique widely used in audio signal processing. Unlike the traditional Fourier transform or the Short-Time Fourier Transform (STFT), which use a linear frequency scale, the CQT employs a logarithmic frequency scale that closely approximates the nonlinear frequency response of the human auditory system.

The CQT is computed by convolving the input signal with a set of complex exponential functions that are equally spaced on a logarithmic frequency scale, with the Q-factor (i.e., the ratio of the center frequency to bandwidth) being kept constant for all bins. This results in a representation of the signal in the time-frequency domain, where each bin corresponds to a fixed Q value and a different center frequency. The CQT can be expressed mathematically as follows:

$$S = \sum_{n=0}^{N-1} x(n) \cdot g^{*}(t - n\Delta t) e^{-2\pi ikn/N} \tag{1}$$

where $x(n)$ is the input signal, $g^{*}(t - n\Delta t)$ is the complex conjugate of the CQT kernel function, k is the frequency bin index, N is the total number of samples, t is the time variable, and Δt is the time resolution of the kernel function. The CQT kernel function is defined as:

$$g(t) = \frac{1}{\sqrt{Q}} w\left(\frac{t}{Q}\right) e^{2\pi if_0 t} \tag{2}$$

where f_0 is the center frequency of the kernel, $w(t)$ is a window function that localizes the kernel in time, and Q is the Q-factor of the kernel.

The CQT produces a time-frequency representation of the signal in the form of a spectrogram, where the amplitude of each bin represents the energy or power of the signal at

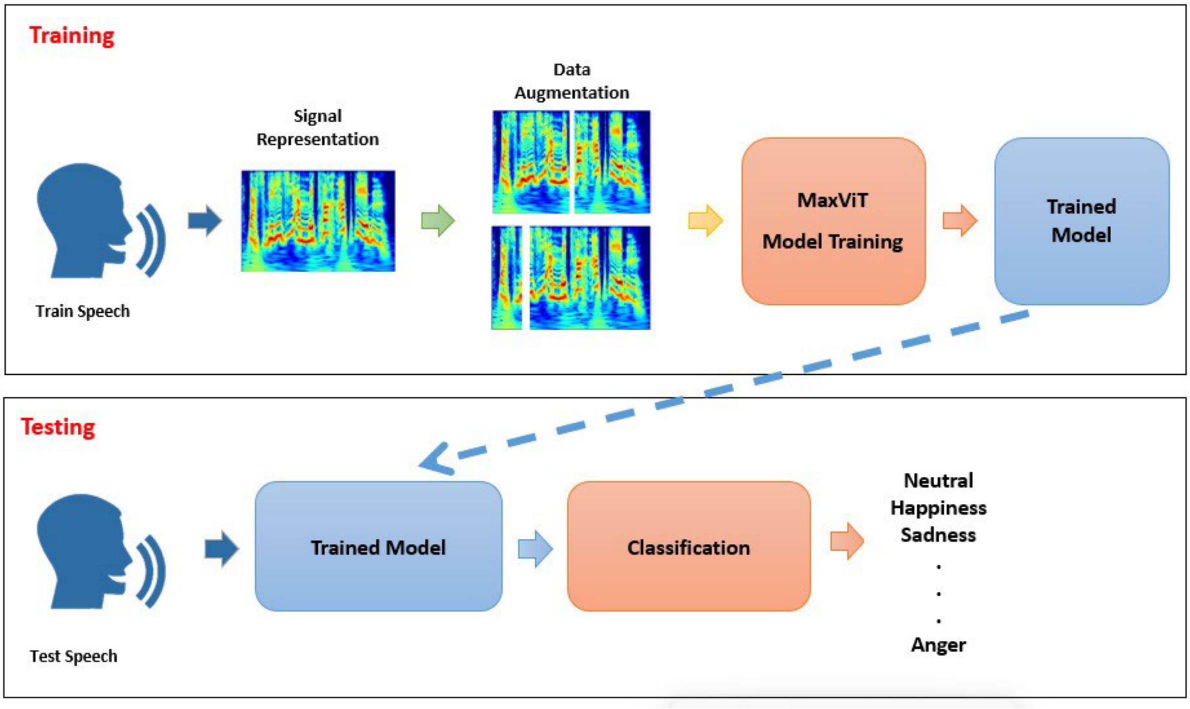


FIGURE 1. Workflow of SCQT-MaxViT.

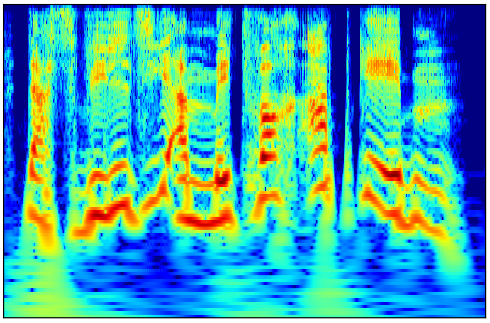


FIGURE 2. Sample of a Constant-Q Transform spectrogram.

that particular frequency and time. Since the CQT uses a logarithmic frequency scale, it provides higher resolution at lower frequencies and lower resolution at higher frequencies, which is well-suited for analyzing musical signals that have a pitch structure. Moreover, the CQT can capture the harmonic structure of the signal even when the fundamental frequency varies over time, making it more robust to pitch variations than the STFT. Not only that, the CQT also provides better resolution and robustness than traditional Fourier-based techniques and can be computed efficiently using fast algorithms such as the FFT. Figure 2 shows an example of a CQT spectrogram.

B. DATA AUGMENTATION

Data augmentation techniques are widely employed to generate additional training samples, thereby enhancing the robustness and generalization capabilities of models. In this study, time masking is performed on the CQT spectrograms. Time masking involves masking (i.e., setting to zero) a continuous time segment of the spectrogram, effectively removing some of the temporal information. The time masking operation can be represented as follows:

Let S be the original spectrogram of size $F \times T$, where F is the number of frequency bins and T is the number of time steps.

1) Select a random time step t, where $0 \leq t < T$.
2) Select a random time mask width τ, where $0 \leq \tau \leq T$ and $\tau \leq T_{\max}$, with $T_{\max}$ being the maximum allowed time mask width.
3) Create a time-masked spectrogram S' by setting the values in the time range $[t, t+\tau)$ to zero.

Mathematically, this can be represented as:

$$S'(f, t') = \begin{cases} 0, & \text{if } t \leq t' < t+\tau \\ S(f, t'), & \text{otherwise} \end{cases} \tag{3}$$

Here, $S'(f, t')$ represents the value in the time-masked spectrogram at frequency bin f and time step t', and $S(f, t')$ is the value in the original spectrogram at frequency bin f and time step t'.

C. MULTI-AXIS VISION TRANSFORMER

This study employs the Multi-Axis Vision Transformer (MaxViT) [18] for further representation learning and classification. MaxViT incorporates the blocked multi-axis self-attention (Max-SA) module, which introduces a novel

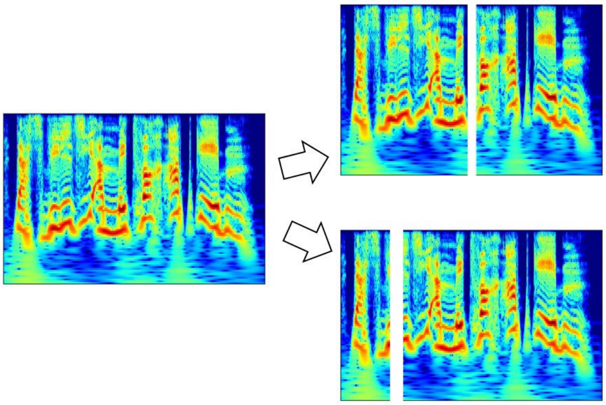


FIGURE 3. Sample time masking with random segments and noise factors.

attention mechanism for enabling global and local interactions within the network while minimizing computational complexity. To achieve this, the Max-SA module decomposes fully dense attention mechanisms into two sparse forms: block attention and grid attention, inspired by previous sparse approaches. This decomposition effectively converts the quadratic complexity of vanilla attention to linear complexity without compromising non-locality. The Max-SA module utilizes self-attention to enable spatial mixing across the entire spatial or sequence locations based on content-dependent weights derived from normalized pairwise similarity. The pre-normalized relative self-attention, known for incorporating a learned bias, is utilized in this study, consistently outperforming the original attention mechanism in various vision tasks.

To address the computational challenges associated with applying attention across the entire space, Max-SA introduces a multi-axis approach encompassing block attention and grid attention. These two types of attention are sequentially stacked within a single block to facilitate local and global interactions. In the block attention, the input feature map, denoted as $X \in \mathbb{R}^{H\times W\times C}$, undergoes a partitioning technique to enable attention mechanisms. Rather than directly applying attention to the flattened spatial dimension HW, the feature map is reshaped into a tensor of shape $(\frac{H}{P} \times \frac{W}{P}, P \times P, C)$, resulting in non-overlapping windows within the feature maps, each with a size of $P \times P$.

Furthermore, the grid attention mechanism is employed by reshaping the tensor into $(G \times G, \frac{H}{G} \times \frac{W}{G}, C)$, gridding the feature maps into $G \times G$ partitions. By utilizing fixed window and grid sizes ($P = G = 7$), the computational load is evenly distributed between local and global operations, both exhibiting linear complexity relative to the spatial size or sequence length.

The Max-SA module seamlessly integrates into the MaxViT block, which incorporates additional components such as LayerNorm, Feedforward networks (FFNs), skip-connections, and a Mobile Inverted Residual Bottleneck Convolution (MBConv) block with a squeeze-and-excitation (SE) module. The MBConv block enhances the network's generalization and trainability, while the depthwise convolutions serve as conditional position encoding (CPE), eliminating the need for explicit positional encoding layers. The hierarchical design of the MaxViT block involves stacking alternating layers of Max-SA with MBConv, providing global and local receptive fields throughout the entire network.

TABLE 1. MaxViT-T configurations.

Stage	Size	MaxViT-T
S0: Conv-stem	1/2	B=2 C=64
S1: MaxViT-Block	1/4	B=2 C=64
S2: MaxViT-Block	1/8	B=2 C=128
S3: MaxViT-Block	1/16	B=5 C=256
S4: MaxViT-Block	1/32	B=2 C=512

The architecture of the MaxViT model is visualized in Figure 4. Following the conventional practices of CNN, a hierarchical backbone is employed, which includes a downsampling step in the initial stem stage (S0) using two convolutional layers with a kernel size of 3 × 3 (Conv3 × 3). The network body consists of four stages (S1-S4), where each subsequent stage has half the resolution of the previous one and a doubled number of channels in the hidden dimension. Identical MaxViT blocks are utilized throughout the entire backbone.

The downsampling operation is applied in the Depthwise Conv3 × 3 layer of the first MobileNetV3-like Convolutional (MBConv) block within each stage. The inverted bottleneck and squeeze-excitation (SE) mechanisms have expansion rates of 4 and shrink rates of 0.25, respectively. The attention head size is set to 32 for all attention blocks. Model scaling is achieved by increasing the number of blocks per stage (B) and the channel dimension (C). In this study, the MaxViT-T variant is employed, and the specific configurations are presented in Table 1. The "Size" column indicates the downsampling ratio of each stage.

MaxViT incorporates both global and local receptive fields throughout the entire network. By leveraging self-attention mechanisms, MaxViT enables the model to capture long-range dependencies and spatial correlations in the CQT spectrograms. This allows for a comprehensive understanding of the spectrogram features at different scales, enhancing the model's ability to recognize subtle patterns and variations associated with different emotions.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section provides an analysis of the key aspects in speech emotion recognition research. It includes an overview of the datasets utilized, an evaluation of various spectrograms as feature representations, a discussion on the classifiers employed, an examination of data augmentation techniques, and a comparison of the proposed approach with existing works in the field.

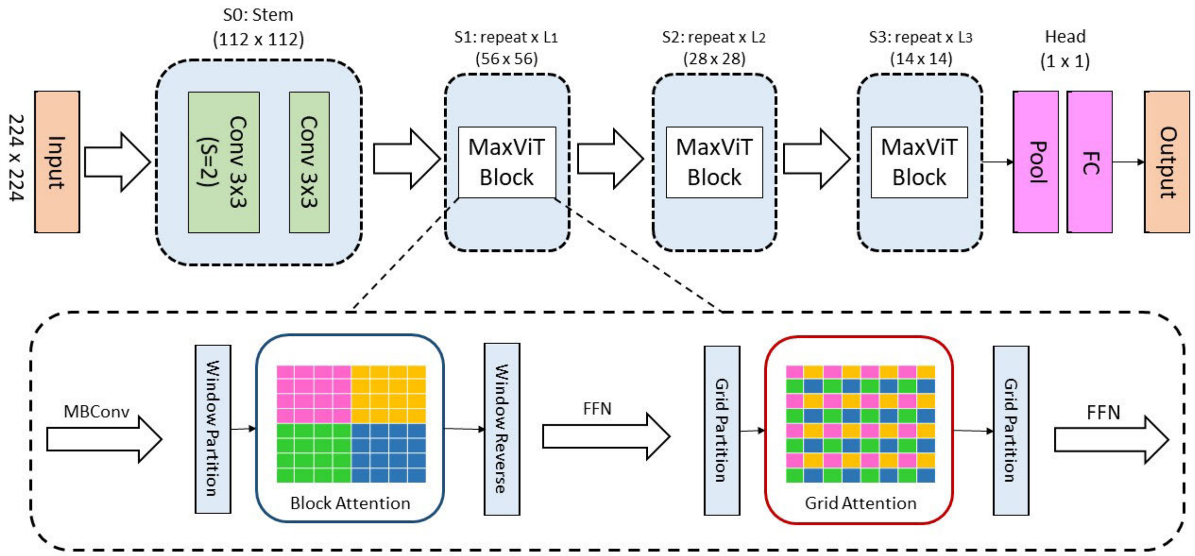


FIGURE 4. Architecture of the MaxViT model.

A. DATASETS

The proposed SCQT-MaxViT method is evaluated on three speech emotion datasets, namely the Berlin Database of Emotional Speech (Emo-DB), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Interactive Emotional Dyadic Motion Capture (IEMOCAP). To ensure a fair comparison with previous research, each dataset is split into an 80% training set and a 20% testing set.

The Berlin Database of Emotional Speech (Emo-DB) [19] is a German language speech emotion recognition dataset that includes 535 samples from five male and five female professional speakers. The dataset contains 535 samples with 7 emotions: 127 anger, 81 boredom, 79 neutral, 71 happiness, 69 anxiety, 62 sadness, and 46 disgust audio samples. Each utterance is labeled with the corresponding emotion and stored as a separate audio file in WAV format.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [20] consists of 1440 samples with eight emotions: 96 neutral, 192 calm, 192 happy, 192 sad, 192 angry, 192 fearful, 192 disgust, and 192 surprised audio samples. The dataset was recorded in English by twelve male and twelve female professional actors who produced two versions of each emotion, one with speech and one with singing, resulting in a total of 1440 recordings. The audio recordings are stored in the WAV format.

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [21] dataset includes 5507 samples produced by five male and five female actors in various conversational tasks. The actors were asked to converse with each other and to engage in tasks such as telling stories or solving problems. To have a fair comparison with existing works, this study considered four emotions: 1704 neutral, 1636 happiness, 1090 anger, and 1077 sadness audio samples.

B. EXPERIMENTAL RESULTS OF DIFFERENT SPECTROGRAMS

The performance analysis of different spectrogram types in conjunction with MaxViT for speech emotion recognition is presented. The speech signals are initially sampled at a frequency of 44.1kHz and subsequently transformed into CQT spectrograms. To conform to the input size requirements of the MaxViT model, the CQT spectrograms are appropriately resized to a resolution of 224 × 224 pixels.

Experimental results comparing three spectrogram types (CQT, Linear-STFT, and MFCC) combined with MaxViT are summarized in Table 2. Notably, the accuracy values consistently indicate that CQT spectrograms outperform the other two types. In terms of accuracy, CQT+MaxViT yielded the highest scores of 86.79%, 76.14%, and 62.31% for the Emo-DB, RAVDESS, and IEMOCAP datasets, respectively.

CQT spectrograms are well-regarded for their ability to provide a more comprehensive representation of audio signals in the time-frequency domain, enabling them to capture intricate details in human speech. Unlike Linear-STFT, CQT adopts a logarithmic frequency scale, which closely aligns with the characteristics of the human auditory system. Furthermore, CQT spectrograms possess enhanced resolution at lower frequencies, which is crucial for accurately detecting the fundamental frequency of human speech. Conversely, MFCC leverages a mel-scale filterbank to extract features, potentially leading to the loss of valuable information embedded within the audio signals.

C. EXPERIMENTAL RESULTS OF DIFFERENT CLASSIFIERS

Table 3 provides the experimental results of different classifiers with the MaxViT model. The evaluated methods involve combining the CQT with various classifiers,

TABLE 2. Experimental results of different spectrograms with MaxViT.

Methods	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
Linear-STFT + MaxViT	73.58	63.86	55.68
MFCC + MaxViT	79.25	74.74	56.49
CQT + MaxViT	**86.79**	**76.14**	**62.31**

TABLE 3. Experimental results of different classifiers.

Methods	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
CQT + RF	49.10	41.00	49.80
CQT + KNN	58.50	46.00	41.30
CQT + SVM	70.80	63.00	46.60
CQT + MLP	51.90	31.90	30.00
CQT + VGG19	85.85	73.33	58.49
CQT + CNN	84.91	74.39	59.95
CQT + CoAtNet	83.96	71.58	59.04
CQT + ViT	83.02	66.67	59.40
CQT + MaxViT	**86.79**	**76.14**	**62.31**

including Random Forest, K-Nearest Neighbors (KNN), SVM, MLP, VGG19, CNN, CoAtNet [22], Vision Transformer (ViT) [23], and MaxViT. Accuracy percentages for three datasets, namely Emo-DB, RAVDESS, and IEMOCAP, are reported.

The CQT + MaxViT method demonstrates the highest accuracy across all three datasets, with values of 86.79% for Emo-DB, 76.14% for RAVDESS, and 62.31% for IEMOCAP. This significant performance superiority over other methods highlights the effectiveness of the MaxViT model in capturing intricate relationships within the audio data. By leveraging multi-axis self-attention mechanisms, MaxViT effectively captures long-range dependencies and spatial correlations within the CQT spectrograms. This allows the model to understand the spectrogram features at different scales and recognize subtle patterns and variations associated with different emotions. The combination of MaxViT with CQT spectrograms provides a powerful framework for speech emotion classification, enabling the model to achieve high accuracy by effectively extracting meaningful representations from the audio data.

Apart from that, several other classifiers, including CQT + VGG19, CQT + CNN, CQT + CoAtNet, and CQT + ViT, achieve relatively high accuracy percentages. These models exhibit accuracy values ranging from 58.49% to 85.85% for Emo-DB, 59.95% to 74.39% for RAVDESS, and 66.67% to 73.33% for IEMOCAP. These outcomes suggest the efficacy of pre-trained models, especially those trained on visual recognition tasks, in extracting meaningful representations from audio data.

TABLE 4. Experimental results MaxViT with and without data augmentation.

Methods	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
MaxViT without Data Augmentation	86.79	76.14	62.31
MaxViT with Data Augmentation	**88.68**	**77.54**	**62.49**

D. EXPERIMENTAL RESULTS OF DATA AUGMENTATION

This section presents the experimental results of the proposed SCQT-MaxViT in the context of data augmentation. The obtained results are shown in Table 4. The findings demonstrate that performing data augmentation significantly enhances the performance of SCQT-MaxViT, as compared to SCQT-MaxViT without data augmentation, across all three datasets. Particularly noteworthy improvements are observed in Emo-DB and RAVDESS, where statistically significant enhancements are observed. Specifically, the accuracy in Emo-DB increases from 86.79% to 88.68%, while in RAVDESS, it increases from 76.14% to 77.54%. Although the improvement in IEMOCAP is marginal, there is a slight increase from 62.31% to 62.49%.

These results underscore the efficacy of data augmentation in bolstering the performance of SCQT-MaxViT in speech-based emotion recognition tasks. Data augmentation involves the application of random time masking to the input spectrograms during training. This technique aids the model in developing improved generalization capabilities, enabling it to effectively handle unseen data and mitigate overfitting issues.

E. COMPARISON RESULTS WITH THE EXISTING WORKS

Table 5 compares the proposed SCQT-MaxViT method with existing methods on three emotion speech datasets: Emo-DB, RAVDESS, and IEMOCAP. The experimental result shows that the proposed method outperforms the existing methods on all three datasets. The dash "-" in the table represents that the dataset was not used in the existing works. On the Emo-DB dataset, existing methods achieved accuracy in the range of 60.05% to 87.31%, while the proposed SCQT-MaxViT method achieved a significantly higher accuracy of 88.68%. Similarly, on the RAVDESS dataset, the best existing method MFCC with CNN [3] achieved an accuracy of 75.63%, which is 1.91% lower than the proposed SCQT-MaxViT method. However, all methods performed relatively poorly on the IEMOCAP dataset, likely due to it contains spontaneous, unscripted and sometimes overlapping conversations between actors. Existing methods achieved an accuracy of 54.60% to 62.01%, while the proposed SCQT-MaxViT method achieved a higher accuracy of 62.49%.

The remarkable improvement in performance confirms the effectiveness of the proposed SCQT-MaxViT method in

TABLE 5. Comparative results on Emo-DB, RAVDESS, IEMOCAP dataset.

Methods	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
Support Vector Machine [1]	86.36	64.15	-
Formant Characteristics Features with RF [2]	71.05	49.20	62.01
Formant Characteristics Features with KNN [2]	60.05	43.24	59.28
Formant Characteristics Features with MLP [2]	72.91	61.02	61.91
MFMC with Convolutional Neural Network [3]	81.50	75.63	-
INTERSPEECH 2009 Feature set with LR [4]	80.75	62.64	-
Feature set 2 with MLP [4]	84.86	68.06	-
Feature set 2 with SVM [4]	85.97	70.42	-
Feature set 2 with RF [4]	74.01	62.97	-
SVM with Gaussian kernel [5]	80.70	63.80	58.90
Deep Neural Network with Support Vector Machine [6]	79.86	52.24	-
Temporal and Frequency Features with LGBM [7]	84.91	67.72	-
Deep CNN with Discriminant Temporal Pyramid Matching [8]	87.31	-	-
Complementary Features with KELM [9]	84.49	-	57.10
Parallelized Convolutional Recurrent Neural Network [10]	86.44	-	-
Visual Geometry Group-19 [11]	-	-	54.60
Convolutional Neural Network [12]	72.02	-	59.33
Attention Convolutional Neural Network [13]	-	-	59.54
Visual Attention-based Convolutional Neural Network [14]	79.44	74.31	-
Fusion Classifiers [15]	-	-	58.30
2D-CNN [16]	-	64.00	-
CNN-2D & LSTM with self-attention [17]	-	74.44	-
CNN-2D [17]	-	73.70	-
CNN-2D with LSTM [17]	-	70.37	-
SCQT-MaxViT (Proposed)	**88.68**	**77.54**	**62.49**

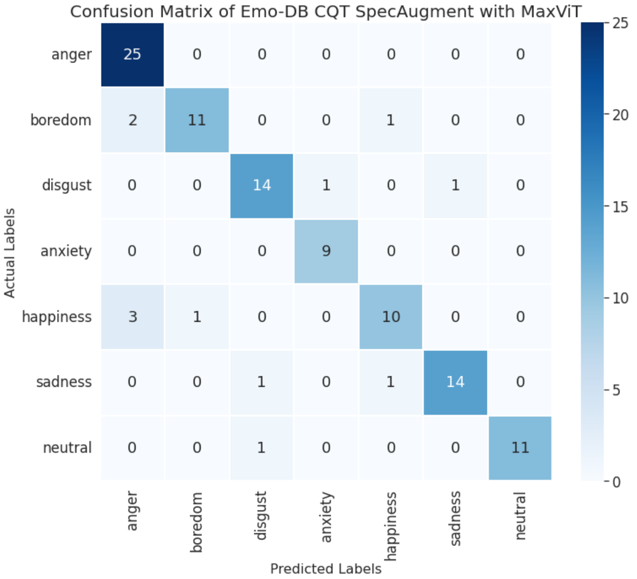


FIGURE 5. Confusion matrix of the Emo-DB.

speech emotion recognition. The use of CQT, with its logarithmically spaced frequency bins, provides better frequency resolution at lower frequencies and allows for accurate identification of relevant features in speech signals. Additionally, the logarithmic frequency spacing of the CQT makes it less sensitive to pitch variations, which are common in emotional speech, resulting in the extraction of more consistent and reliable features for emotion recognition tasks. The variable window size of the CQT also provides better temporal resolution for higher frequencies.

The application of data augmentation increases the amount of training data by applying random transformations to the original spectrograms. This improves the model's generalization ability and makes it more robust to variations in emotional speech patterns. The use of time masking randomly removes a range of segments from the spectrogram, forcing the model to rely on other segments for emotion recognition, improving its ability to recognize emotions even in the presence of noise or other distortions.

MaxViT can extract features from different levels of abstraction in the input data, enabling the model to capture both local and global patterns crucial for identifying emotional cues in speech signals. The multi-axis attention mechanism allows the model to focus on different spatial positions and scales in the CQT spectrograms, leading to a better understanding of the relationships between different frequency components and their temporal variations, thereby improving emotion recognition.

Figure 5 presents the confusion matrix of the proposed SCQT-MaxViT method on the Emo-DB dataset, indicating the model's performance in classifying different emotions. As seen in the confusion matrix, misclassifications are typically observed in classes with relatively fewer samples. The results suggest that happiness and boredom are the emotions that are most frequently misclassified, often mistaken for anger. The possible reason behind this could be that these emotions share certain acoustic characteristics with anger, leading to confusion in the classification.

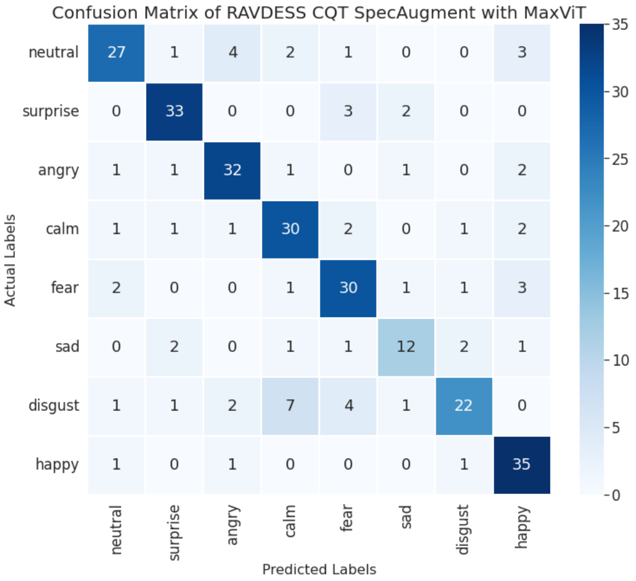


FIGURE 6. Confusion matrix of the RAVDESS dataset.

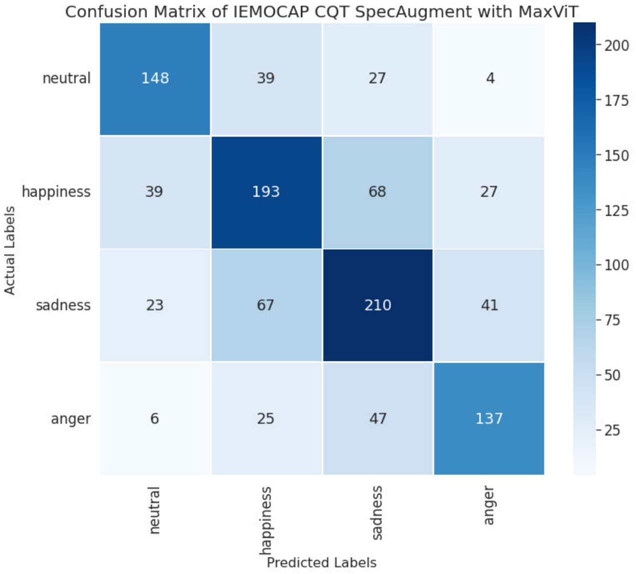


FIGURE 7. Confusion matrix of the IEMOCAP dataset.

The confusion matrix depicted in Figure 6 provides insights into the performance of the proposed SCQT-MaxViT model on the RAVDESS dataset. Notably, certain classes exhibit a higher degree of confusion with others, indicating inherent challenges in accurately distinguishing between them. Specifically, the "disgust" class frequently overlaps with "fear" or "calm" in terms of misclassification, while the "neutral" class is often mistaken for "angry" or "happy". These misclassifications could be attributed to shared acoustic characteristics between the audio waveforms associated with these emotions. However, individual differences in emotional expression may also contribute to misclassification, as individuals may exhibit unique variations in expressing the same emotion through their audio signals.

The confusion matrix presented in Figure 7 provides an overview of the performance of the proposed SCQT-MaxViT model on the demanding IEMOCAP dataset. The majority of misclassifications are observed within the happiness and sadness classes, highlighting the difficulty in accurately distinguishing between these two emotions. Notably, the IEMOCAP dataset poses considerable challenges for emotion recognition models due to its unique characteristics. The dataset encompasses a substantial number of samples, and the speech recordings involve multiple speakers simultaneously, complicating the isolation of individual speaker characteristics. Furthermore, the brevity of some speech segments presents a challenge in accurately capturing the underlying emotion contained within the audio waveforms.

V. CONCLUSION

This paper presents a novel approach, SCQT-MaxViT, which enhances speech emotion recognition by combining the Constant-Q Transformer (CQT) spectrogram and the Multi-Axis Vision Transformer (MaxViT). The proposed method involves a series of stages to extract and process features from audio samples. Initially, the audio signals are transformed into visual representations using the CQT spectrogram, which captures the frequency content of the speech. To improve the diversity and robustness of the training data, a data augmentation technique called time masking is employed, introducing random variations in the spectrogram by masking specific time bands. This augmentation strategy effectively reduces overfitting and enables the model to generalize better to unseen data. The augmented CQT spectrograms are then fed into the Multi-Axis Vision Transformer for speech emotion recognition, allowing both global and local interactions within the network. The SCQT-MaxViT approach achieves impressive accuracy rates of 88.68%, 77.54%, and 62.49% on the Emo-DB, RAVDESS, and IEMOCAP datasets, respectively, demonstrating its effectiveness as a reliable solution for speech emotion recognition tasks. Despite the remarkable performance exhibited by the proposed SCQT-MaxViT model on various datasets, it is worth noting that the IEMOCAP dataset presents unique challenges due to its dyadic nature, involving interactions between two individuals. The complexity of capturing and interpreting emotional cues in dialogue scenarios requires further attention in future research.

REFERENCES

[1] R. Singh, H. Puri, N. Aggarwal, and V. Gupta, "An efficient language-independent acoustic emotion classification system," *Arabian J. Sci. Eng.*, vol. 45, no. 4, pp. 3111–3121, Apr. 2020.

[2] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, and M. Hao, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence," *Inf. Sci.*, vol. 563, pp. 309–325, Jul. 2021, doi: 10.1016/j.ins.2021.02.016.

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

[4] M. E. Seknedy and S. Fawzi, ''Speech emotion recognition system for human interaction applications,'' in *Proc. 10th Int. Conf. Intell. Comput. Inf. Syst. (ICICIS)*, Dec. 2021, pp. 361–368.

[5] L. F. Parra-Gallego and J. R. Orozco-Arroyave, ''Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments,'' *Digit. Signal Process.*, vol. 120, Jan. 2022, Art. no. 103286.

[6] 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.

[7] 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.*, vol. 13, no. 4, p. 4020, Aug. 2023.

[8] S. Zhang, S. Zhang, T. Huang, and W. Gao, ''Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching,'' *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, Jun. 2018.

[9] L. Guo, L. Wang, J. Dang, Z. Liu, and H. Guan, ''Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine,'' *IEEE Access*, vol. 7, pp. 75798–75809, 2019.

[10] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, ''Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition,'' *IEEE Access*, vol. 7, pp. 90368–90377, 2019.

[11] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, ''Data augmentation using GANs for speech emotion recognition,'' in *Proc. Interspeech*, Sep. 2019, pp. 171–175.

[12] P. M. Larisa and R. Tapu, ''Speech emotion recognition using 1D/2D convolutional neural networks,'' in *Proc. Int. Symp. Electron. Telecommun. (ISETC)*, Nov. 2022, pp. 1–4.

[13] M. Neumann and N. T. Vu, ''Improving speech emotion recognition with unsupervised representation learning on unlabeled speech,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7390–7394.

[14] M. Seo and M. Kim, ''Fusing visual attention CNN and bag of visual words for cross-corpus speech emotion recognition,'' *Sensors*, vol. 20, no. 19, p. 5559, Sep. 2020.

[15] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, ''Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN,'' *Speech Commun.*, vol. 120, pp. 11–19, Jun. 2020.

[16] S. P. Singh, S. Kumar, S. Verma, and I. Kaur, ''Hybrid approach for human emotion recognition from speech,'' in *Proc. 4th Int. Conf. Adv. Comput., Commun. Control Netw. (ICAC3N)*, Dec. 2022, pp. 1282–1285.

[17] J. Singh, L. B. Saheer, and O. Faust, ''Speech emotion recognition using attention model,'' *Int. J. Environ. Res. Public Health*, vol. 20, no. 6, p. 5140, Mar. 2023.

[18] 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.

[19] 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, doi: 10.21437/interspeech.2005-446.

[20] 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.

[21] 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.

[22] Z. Dai, H. Liu, Q. V. Le, and M. Tan, ''CoatNet: Marrying convolution and attention for all data sizes,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 3965–3977.

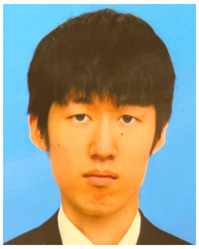
[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ''An image is worth 16×16 words: Transformers for image recognition at scale,'' 2020, *arXiv:2010.11929*.

KAH LIANG ONG received the bachelor's degree (Hons.) in information technology and in 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 Master of Science and Ph.D. degrees in information technology and in abnormal behavior detection and gait recognition. She is currently a Senior Lecturer with the Faculty of Information Science and Technology, Multimedia University, Malaysia. She has been a certified professional technologist, since 2018, a member of the International Association of Engineers, since 2020, and the Outcome-Based Education Consultant and a Trainer. 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 an emphasis on detection and estimation theory and their applications.

KIAN MING LIM (Senior Member, IEEE) received the B.IT. (Hons.) degree 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 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.



TAKEKI MUKAIDA received the Associate of Science degree in engineering from the Gifu College, National Institute of Technology, in March 2022. He is currently pursuing the Bachelor of Science degree in informatics with the University of Electro-Communications.