applied sciences

MDPI

Article

Multi-Level Attention-Based Categorical Emotion Recognition Using Modulation-Filtered Cochleagram

Zhichao Peng [1,*], Wenhua He [1], Yongwei Li [2], Yegang Du [3] and Jianwu Dang [4,5,*]

[1] Information School, Hunan University of Humanities, Science and Technology, Loudi 417000, China; hehuald@163.com
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100045, China; yongwei.li@nlpr.ia.ac.cn
[3] Future Robotics Organization, Waseda University, Tokyo 169-8050, Japan; yg.du@aoni.waseda.jp
[4] College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[5] Pengcheng Laboratory, Shenzhen 518055, China
* Correspondence: zcpeng@tju.edu.cn (Z.P.); jdang@jaist.ac.jp (J.D.)

Abstract: Speech emotion recognition is a critical component for achieving natural human–robot interaction. The modulation-filtered cochleagram is a feature based on auditory modulation perception, which contains multi-dimensional spectral–temporal modulation representation. In this study, we propose an emotion recognition framework that utilizes a multi-level attention network to extract high-level emotional feature representations from the modulation-filtered cochleagram. Our approach utilizes channel-level attention and spatial-level attention modules to generate emotional saliency maps of channel and spatial feature representations, capturing significant emotional channel and feature space from the 3D convolution feature maps, respectively. Furthermore, we employ a temporal-level attention module to capture significant emotional regions from the concatenated feature sequence of the emotional saliency maps. Our experiments on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset demonstrate that the modulation-filtered cochleagram significantly improves the prediction performance of categorical emotion compared to other evaluated features. Moreover, our emotion recognition framework achieves comparable unweighted accuracy of 71% in categorical emotion recognition by comparing with several existing approaches. In summary, our study demonstrates the effectiveness of the modulation-filtered cochleagram in speech emotion recognition, and our proposed multi-level attention framework provides a promising direction for future research in this field.

Keywords: categorical emotion recognition; auditory signal processing; modulation-filtered cochleagram; multi-level attention

check for updates

Citation: Peng, Z.; He, W.; Li, Y.; Du, Y.; Dang, J. Multi-Level Attention-Based Categorical Emotion Recognition Using Modulation-Filtered Cochleagram. *Appl. Sci.* **2023**, *13*, 6749. https://doi.org/10.3390/app13116749

Academic Editor: Douglas O'Shaughnessy

Received: 17 April 2023
Revised: 29 May 2023
Accepted: 30 May 2023
Published: 1 June 2023

1. Introduction

The Internet of Everything (IoE) presents a plethora of opportunities for human–robot interaction (HRI), and incorporating emotion information can significantly improve the robot's ability to comprehend human intentions during HRI. Emotion information can be obtained through various means, such as speech, facial expressions, gestures, and electroencephalography (EEG) [1]. Among these, speech is the most natural and convenient communication mode between humans and robots. Therefore, speech emotion recognition (SER) holds immense potential for diverse applications in HRI, such as intelligent driving, service robotics, online education, telemedicine, and criminal investigations [2].

The extraction of emotional features is one of the key technologies in SER. The commonly used emotional features mainly include: hand-crafted low-level descriptor (LLD) and its high-level statistical features (HSF) [3], Mel filterbank features [4], spectrogam [5,6], etc. However, researchers have not identified the best speech features for SER and still explore the effective features that can represent emotional states [7]. Humans can easily perceive emotional information and its changes through the auditory system. Sounds reach the auditory cortex after passing through several auditory signal processing stages, which then perceives differences in intensity and tone to produce varying psychological responses. Therefore, identifying emotions from the perspective of auditory perception could be an effective approach. However, the human auditory system is highly complex, and the mechanisms of signal processing are not yet fully understood. To address this, researchers have developed functional models of the auditory system based on its characteristics, such as the models of the cochlear basilar membrane, the inner hair cell, the nerve conduction, and the auditory center. These models are mainly applied in a cochlear implant, hearing aid, sound source positioning, speech enhancement [8], and other areas, yet limited studies have explored auditory perception and understanding. Psychoacoustic research reveals that speech signals are decomposed into spectral–temporal components in the cochlea and are subject to spectral–temporal modulation through the auditory pathway, generating a modulation spectrum [9]. This modulation spectrum plays an essential role in speech perception and understanding [10,11]. Several studies have used statistical functions on the modulation spectrum to obtain modulation spectral features (MSF) for SER tasks [12]. Avila et al. [13] proposed a feature-pooling scheme for dimensional emotion recognition using a combined MSF and 3D modulation spectrum representation. Recent research shows that cochleagrams have more advantages in SER than features such as spectrograms [14,15]. Peng et al. [16] proposed a modulation-filtered cochleagram (MCG) feature to extract high-level auditory representations for dimensional emotion recognition. The experimental results showed excellent performance in terms of arousal and valence prediction, but the effectiveness of this feature in categorical emotion recognition requires further improvement.

In order to extract high-level feature representations from speech features, deep learning methods, such as convolutional neural network (CNNs), recurrent neural network (RNNs), and transformers, are mainly used for the SER task [17,18]. CNNs are often used to extract high-level speech feature representation due to their scale and rotation invariance [19]. RNNs, including long short-term memory (LSTM) [20], are commonly employed to capture sequence dependencies owing to their ability to handle long-term dependencies in the speech sequence [21]. Recently, attention mechanisms have been incorporated into deep learning methods to automatically capture salient emotion features in speech sequences. Neumann et al. [4] proposed an attentive CNN (ACNN) based on the attention model to identify emotions from the log-Mel filterbank features. Mirsamadi et al. [22] introduced an attentive RNN (ARNN) model to recognize emotions from frame-level LLDs, with local attention as a weighted pooling method. Peng et al. [23] proposed an attention-based sliding recurrent neural network (ASRNN), which mimics the auditory attention to effectively model the auditory representation sequence and capture salient emotion regions. Moreover, the transformer employs a self-attention mechanism in conjunction with RNN-based encoder–decoder architecture to track the context relations in the sequence data. Chen et al. [8] introduced a Key-Sparse Transformer, which dynamically judges the importance of each frame in the speech signal, so as to help the model pay attention to the emotionally related fragments as much as possible.

Some novel attention models, such as channel attention and spatial attention, are proposed for image recognition and behavior detection. Channel attention focuses on determining the importance of different channels and has been employed in models, such as SE-Net [24], SK-Net [25], and ECA-Net [26]. Spatial attention transforms information into another space through a spatial conversion module to retain key information. Notable examples of spatial attention models include A2-Net [27], DANet [28], and convolutional block attention module (CBAM) [29]. In addition, researchers have developed multi-level attention models that operate in different dimensions. Ma et al. [30] introduced TripleNet, a model that employs a hierarchical representation module to construct representations of context, reply, and query in multi-turn dialogues. The triple attention mechanism is utilized to update these representations. Liu et al. [31] proposed TANet for speech dialogue and object detection, which considers triple attention at the channel, point, and voxel. Jiang et al. [32] proposed a convolutional-recurrent neural network with multiple attention mechanisms for SER. This method employed the multiple attention layer to calculate the weights for different frames and features and a self-attention layer to calculate the weights from Mel-spectrum features. Liu et al. [33] proposed a novel multi-level attention network comprising a multiscale low-level feature extractor and a multi-unit attention module for SER. Zou et al. [34] proposed an end-to-end speech emotion recognition system using multi-level acoustic information with a newly designed co-attention module. These methods leverage multiple attention models to extract different channel and spatial attention maps from LLDs, spectrograms, and waveforms and subsequently fuse these attention maps to recognize emotions. However, they do not consider the capture of significant emotional regions in speech sequences using temporal attention.

The MCG features include multi-dimensional spectral–temporal feature representations of acoustic, modulation, and temporal attributes, considering that channel attention and spatial attention are efficient ways to extract high-level features from high-dimensional space, and temporal attention can effectively capture the significant emotional parts in speech feature sequences. To address the efficiency problem of high-dimensional feature extraction in MCG features, this paper proposes a categorical emotion recognition method that employs a multi-level attention network to extract salient information from modulation-filtered cochleagram features. Firstly, a 3D CNN is used to extract high-level auditory feature representation from modulation-filtered cochleagram. Then, the channel-level attention module is used to capture the dependence of the channel structure from the 3D convolution feature map; the spatial-level attention module is used to capture the dependence of the spectral–temporal spatial structure of the feature representation. Finally, a temporal-level attention module is used to capture the significant emotional regions from the concatenated feature sequence derived from the channel and spatial attention map.

The major contributions of this study are as follows:

- Using the same convolutional recurrent neural network, the MCG features perform better than other evaluation features in categorical emotion recognition.
- The multi-level attention network is proposed for improving emotion recognition performance, in which channel-level and spatial-level attention modules obtain fused features from MCG features, and temporal-level attention further captures significant emotional regions from fused feature sequences.
- The proposed method is evaluated on the Interactive Emotional Dyadic Motion Capture database (IEMOCAP). It obtains an unweighted accuracy of 71% and an F1 score of 69.2%, showing the effectiveness of our approach.

The remainder of this paper is organized as follows. In Section 2, we describe the modulation-filtered cochleagram feature. In Section 3, we describe the proposed emotional recognition framework with a multi-level attention module. The experiments and results are presented in Section 4. Finally, the paper is concluded in Section 5.

2. Modulation-Filtered Cochleagram

In this section, we introduce modulation-filtered cochleagram features from spectral–temporal modulation representation.

2.1. Modulation-Filtered Cochleagram Features

The modulation-filtered cochleagram feature is used to capture the temporal modulation cues from emotional speech and achieves significant effects in dimensional emotion prediction. In this study, we explore the potential of employing the modulation-filtered cochleagram features for categorical emotion recognition. The emotional speech signal $s(t)$ is first filtered using a bank of Gammatone cochlea filters. Then, the temporal envelope of the subchannel signal is extracted using Hilbert transform. Furthermore, the m th modulation filter in the nth channel envelope signal is used to obtain the spectral–temporal modulation signal $s_{mu}(n,m,i)$, which is defined as:

$$s_{mu}(n,m,i) = \mathrm{w}(t_w) \cdot s_m(n,m,(i-1) \cdot Len_s + t_w) \tag{1}$$

where $\mathrm{w}(t_w)$ is the window function, t_w is the time window size, and Len_s is the frame shift. $s_{mu}(n,m,i)$ refers to the m th modulation channel and the n th cochlea acoustic channel of the i th modulation unit, and a total of $n * m$ channel signals are generated, where $1 \le i \le L$, L is equal to $Len_\mathrm{t}/Len_\mathrm{s}$, and Len_t is the total length of the speech signal $s(t)$. $\mathrm{s_m}(n,m,(i-1) \cdot Len_s + t_w)$ is the spectral–temporal modulation signal of the n subchannel and the m subchannel of the i modulation unit. $s_{mu}(n,m,i)$ represents the m modulation subchannel in the n acoustic subchannel. The calculation formula is as follows:

$$\mathrm{s_m}(n,m,t) = \mathrm{m_f}(m,t) * \mathrm{s_e}(n,t), 1 \le \mathrm{m} \le \mathrm{M} \tag{2}$$

where $\mathrm{m_f}(m,t)$ is the pulse response of the modulation filterbank, M is the number of channels in the modulation filterbank, and $\mathrm{s_e}(n,t)$ is calculated by $\mathrm{s_g}(n,t)$ as the size of the complex resolution signal $\hat{\mathrm{s_g}}(n,t) = \mathrm{s_g}(n,t) + \mathrm{j}\mathcal{H}\,\{\mathrm{s_g}(n,t)\}$. $\mathcal{H}\,\{\cdot\}$ represents the Hilbert transformation. Therefore, $\mathrm{s_e}(n,t)$ is calculated as follows:

$$\mathrm{s_e}(n,t) = \left|\hat{\mathrm{s_g}}(n,t)\right| = \sqrt{\mathrm{s_g}(n,t)^2 + \mathcal{H}\{\mathrm{s_g}(n,t)\}^2} \tag{3}$$

The $\mathrm{s_g}(n,t)$ represents the speech signal $\mathrm{s}(t)$ of the nth channel of the speech signal processed by the auditory filter, using the following formula:

$$\mathrm{s_g}(n,t) = \mathrm{g_t}(n,t) * \mathrm{s}(t), 1 \le \mathrm{n} \le \mathrm{N} \tag{4}$$

where $g_\mathrm{t}(n,t)$ represents the pulse response of the nth channel of the filterbank, $*$ represents the convolution operation, t is the number of samples in the time domain, and N is the number of channels in the auditory filterbank. The Gammatone filterbank is used to simulate the motion of the cochlear basilar membrane, and its pulse response is the product of the Gamma distribution and the cosine signal:

$$g_\mathrm{t}(n,t) = \mathrm{A}t^{n_\mathrm{f}-1}\exp\Big(-2\pi w_f ERB_N(f_\mathrm{n})t\Big)\cos\Big(2\pi f_\mathrm{n}\,t + \varphi\Big) \tag{5}$$

where A, n_f and w_f are the amplitude, order, and bandwidth of the filter, $\mathrm{A}t^{n_\mathrm{f}-1}\exp\Big(-2\pi w_f ERB_N(f_\mathrm{n})t\Big)$ is the amplitude term of the Gamma distribution representation, f_n is the central frequency of the nth channel of the filter, and $ERB_N(f_\mathrm{n})$ is the equivalent rectangular bandwidth of f_n, which is a psychoacoustic measure of the width of the auditory filter at each point along the cochlea. The calculation formula is provided as follows:

$$ERB_n(f_n) = \frac{f_n}{Q_{ear}} + B_{min} \tag{6}$$

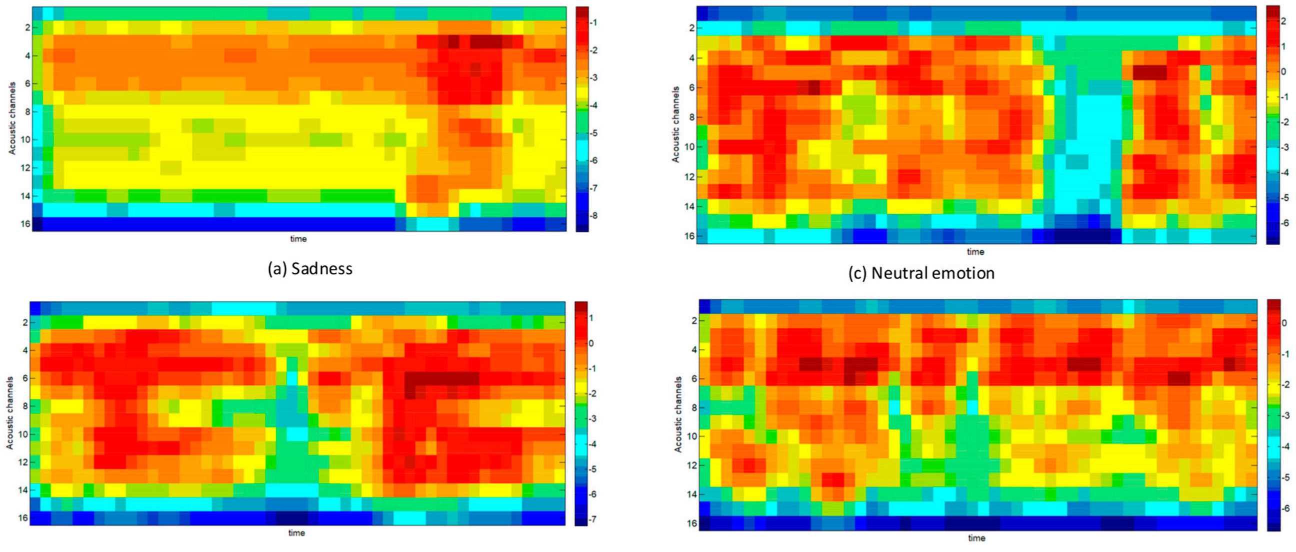
where f_n is the central frequency of the n th filter, $\frac{f_n}{Q_{ear}}$ is the quality factor, which approximates the filtering quality of the high-frequency band, and B_{min} is the minimum bandwidth, representing the approximation of the filtering quality of the low-frequency band. Q_{ear} and B_{min} generally adopt the values proposed in the literature [35], with 9.26449 and 24.7, respectively.

$MCG(c,i)$ results from the convolution operation of each modulation unit:

$$MCG(c,i) = \sum_{i=0}^{L-1} s_{mu}(c,i) * s_{mu}(c,i). \tag{7}$$

2.2. MCG Feature Representation of Different Emotions

The MCG feature exhibits varying weights of emotions across different channels, with a particular focus on low-modulation frequency channels around 4 Hz. Within these channels, neutral emotion and sadness tend to express themselves at lower modulation frequencies, while anger and happiness display contrasting patterns [11]. Figure 1 shows examples of the MCG feature of the first modulation channel in different emotional speech from the IEMOCAP dataset [36]. The x-axis represents the speech sequence, and the y-axis represents the number of acoustic channels n (n = 16). Figure 1a–d show the modulation-filtered cochleagram of sadness, anger, neutral emotion, and happiness, respectively. By observing these panels, we can discern that different emotions manifest in distinct acoustic channels, suggesting potential discrimination based on MCG features. In the cochleagram, the energy associated with sadness primarily concentrates in the slower acoustic channel, while the energy related to anger and happiness primarily concentrates in the higher acoustic channel. However, in comparison to happiness, the energy distribution of anger is relatively concentrated in higher acoustic channels. This shows that different emotions characterized by the acoustic channels are significantly different in the MCG features. We can capture the distinctive characteristics of different emotions from the MCG features.



(a) Sadness

(c) Neutral emotion

(b) Anger

(d) Happiness

Figure 1. Modulation-filtered cochleagram feature representation of different emotions.

3. Emotional Recognition Model

In this section, we introduce a multi-level attention-based emotion recognition model using the modulation-filtered cochleagrams.

3.1. Overview of the Emotion Recognition Model

The proposed emotion recognition model is shown in Figure 2. Firstly, MCG features are extracted through auditory signal processing of the speech signal and fed into the 3D convolution to obtain the high-level feature representation F_{3D}, with a shape of W × H × T × C, in which W, H, T, and C represent the acoustic representation, modulation representation, temporal, and channel, respectively. Subsequently, the multi-level attention module (MAM) is used to capture significant emotional segment information. The MAM extracts emotional information from three dimensions, namely channel (C), space (W × H), and time (T), accurately locating areas with significant emotions. The channel-level attention module is used to capture the dependence of the channel structure from the 3D convolution feature map, the spatial-level attention module is used to capture the dependence of the spectral–temporal spatial structure of the feature representation, and the temporal-level attention module is used to capture the significant emotional regions from the concatenated feature sequence of the channel and spatial attention map. Among them, the channel-level attention and spatial-level attention are responsible for capturing the dependencies between the channel and spatial dimension of the feature map in a parallel mode, respectively. Finally, attention-based feature representations obtained from the temporal-level attention module are fed into a softmax layer, which generates the distribution of the emotional state based on the input speech features.

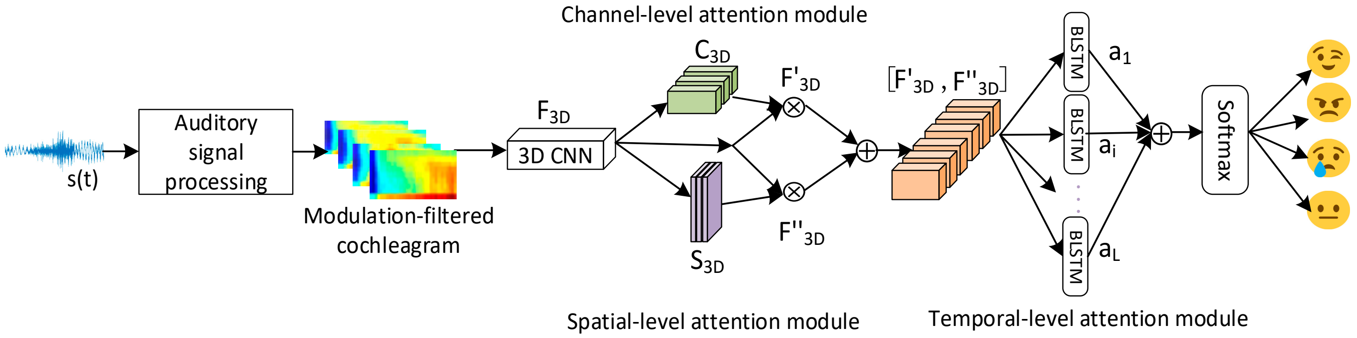

Channel-level attention module
s(t)
Auditory signal processing
Modulation-filtered cochleagram
F3D
3D CNN
C3D
S3D
F'3D
F''3D
[F'3D , F''3D]
BLSTM
BLSTM
BLSTM
a1
ai
aL
Softmax
Spatial-level attention module
Temporal-level attention module

Figure 2. Overview of multi-level attention-based emotion recognition model.

3.2. Channel-Level Attention

The channel-level attention module is used to calculate the channel-wise attention map from the 3D convolution feature map. This attention map helps to recalibrate the weights of each channel, allowing the model to focus on informative parts of the input. The design of the channel-level attention module is similar to that of the CBAM, with a slight difference. In this module, two additional 3D convolutional layers are inserted to capture spatial and temporal information from the feature maps specific to each channel. To compute the channel-level attention efficiently, the spatial and temporal dimensions of the input feature map are squeezed. This operation reduces the feature map's spatial and temporal dimensions while preserving the channel information. The resulting squeezed feature map is then passed through two separate 3D convolutional layers to extract spatial and temporal information for each channel. The outputs of these convolutional layers are used to compute the channel-level attention map. Figure 3 illustrates the structure of the channel-level attention module. The channel-level attention map is first obtained through adaptive learning, and then element-level multiplication with the input feature map F_{3D} is used to obtain a refined feature map F'_{3D}. The calculation formula is provided as follows:

$$F'_{3D} = C_{3D}(F_{3D}) \bigotimes F_{3D}, \tag{8}$$

where C_{3D} represents the channel-level attention map, with $\otimes$ representing the element-level multiplication.

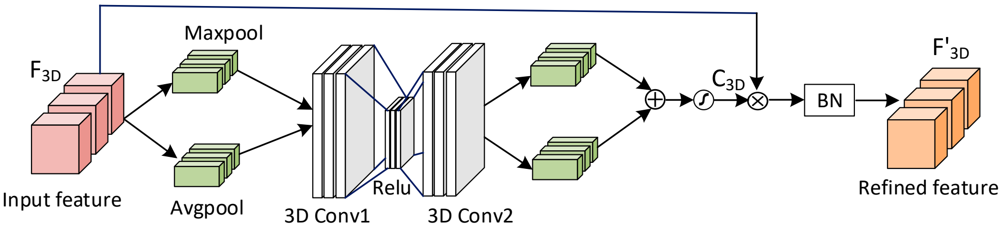


Figure 3. The channel-level attention module.

We first aggregate spatial information of a feature map F_{3D} by using both average-pooling and max-pooling operations, generating two different spatial context descriptors: $Avgpool(F_{3D})$ and $Maxpool(F_{3D})$, which denote adaptive average-pooling features and max-pooling features, respectively.

Both descriptors are then fed into two 3D convolutional layers with a rectified linear unit (ReLU) function. Subsequently, the features are fused using element-wise summation, and the sigmoid activation function is applied to obtain the channel attention map $C_{3D} \in R^{1\times1\times1\times C}$. The channel-level attention map indicates how important each channel is for the emotion recognition results. The calculation formula is as follows:

$$C_{3D}(F_{3D}) = \sigma(Conv_2(Relu(Conv_1(Maxpool(F_{3D})))) + Conv_2(Relu(Conv_1(Avgpool(F_{3D}))))), \tag{9}$$

where $Conv_1$ and $Conv_2$ represent the first and second 3D convolution operations, respectively, and σ denotes a sigmoid operation. Both convolutions are $1 \times 1 \times 1$ convolution kernels, the number of output channels is $\frac{C}{r}$ and C, and r is the dimensionality reduction coefficient in the channel-level attention, with a value of 16. The batch normalization after the channel feature map C_{3D} is used to obtain the same network input distribution and improve the effectiveness of different channels on the feature maps.

3.3. Spatial-Level Attention

The spatial-level attention module is used to calculate the spatial-wise attention map from the 3D convolution feature map. Unlike the channel-level attention module, which focuses on informative channels, the spatial attention module focuses on identifying informative spatial regions in the feature maps. The spatial attention is complementary to the channel-level attention and helps the model determine where the informative parts are located. Figure 4 illustrates the structure of the spatial-level attention module. The spatial-level attention map generated through the spatial-level attention is used for element-level multiplication with the F_{3D} to obtain a refined feature map F''_{3D}. The calculation formula is provided as follows:

$$F''_{3D} = S_{3D}(F_{3D}) \bigotimes F_{3D}, \tag{10}$$

where S_{3D} represents a spatial-level attention map, with $\otimes$ representing element-level multiplication. The feature map F_{3D} integrates the feature map through maximum pooling and average pooling, respectively, to obtain global information. Further, 3D convolution with a kernel size $3 \times 3 \times 1$ is used to obtain spatial regions of emotionally significant spectral–temporal space, thus obtaining a spatial-level attention map $S_{3D} \in R^{W\times H\times1\times1}$. The spatial-level attention map represents the importance of each region in the feature map F_{3D}. The calculation formula is provided as follows:

$$S_{3D}(F_{3D}) = \sigma\Big(f^{3\times3\times1}([Maxpool(F_{3D}), Avgpool(F_{3D})]), \tag{11}$$

where $f^{3\times3\times1}$ is a convolution kernel of size $3 \times 3 \times 1$.

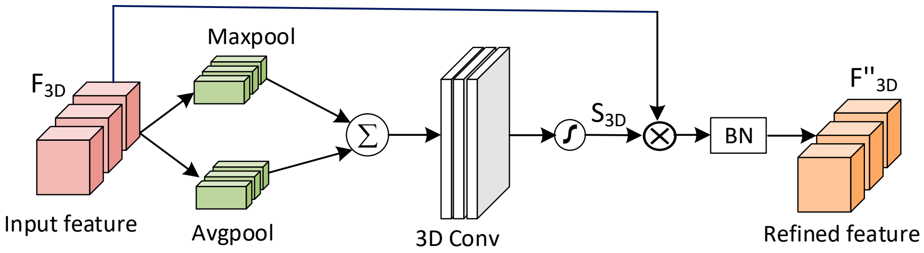


Figure 4. The spatial-level attention module.

3.4. Temporal-Level Attention

The temporal-level attention module is designed to focus on the significant emotional regions by considering the concatenation of the refined feature maps obtained from the channel attention map F'_{3D} and the spatial attention map F''_{3D}. Figure 5 illustrates the structure of the temporal-level attention module. In this module, a bidirectional LSTM (BLSTM) network is utilized. The speech frame sequence is fed into the BLSTM network in both the forward and backward directions. This allows the network to capture temporal dependencies and extract contextual information from both past and future frames. The output of the BLSTM network is a sequence of hidden states h_i. The last states of the forward and backward LSTM cells are concatenated to produce the final sequence of hidden states. This concatenation enables the model to capture long-range dependencies and capture the contextual information from the entire sequence. Subsequently, the ReLU activation function is used to produce non-linear transformations $\mathcal{R}(h_k)$.

$$\mathcal{R}(h_i) = U_i ReLU(W_i h_i + b_i), \tag{12}$$

where W_i and U_i are the trainable parameter matrices, and b_i is the bias vector. We use the non-linear function of the ReLU due to its good convergence performance. For each h_i, the α_i can be computed as follows:

$$\alpha_i = \frac{exp(\mathcal{R}(h_i))}{\sum_{i=1}^{L} exp(\mathcal{R}(h_i))}. \tag{13}$$

We then obtain the attention weights α_i of each sequence from the attention model. The output of the attention layer, att_sum, is the weighted sum of h.

$$att\_sum = \sum_{i=1}^{L} \alpha_i h_i \tag{14}$$

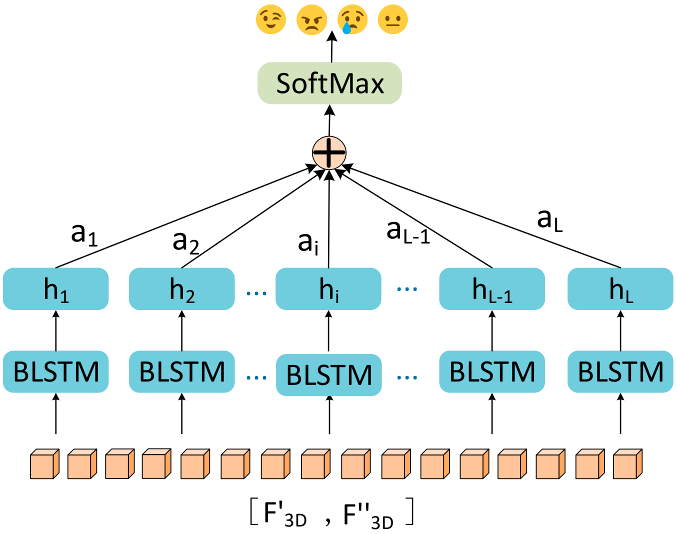


Figure 5. The temporal-level attention module.

Figure 6 illustrates a flowchart of the multi-level attention module for categorical emotion recognition using MCG features. The MCG features are extracted from the speech signal with a shape of 32 × 27 × 200 × 1. These features include modulation representation, which incorporates deltas and delta-deltas of nine modulation filters. To obtain a high-level feature representation, a 3D CNN is used with a shape of 8 × 7 × 50 × 64. The representation is then squeezed to a three-dimensional shape. Next, channel-level attention and spatial-level attention are used to generate emotional saliency maps for both the channel and spatial feature representations. Additionally, temporal-level attention is used to capture significant emotional regions from the concatenated feature sequence of the emotional saliency maps. Finally, the obtained feature representations are fed into a fully connected layer, followed by a softmax function, to derive the emotion state.

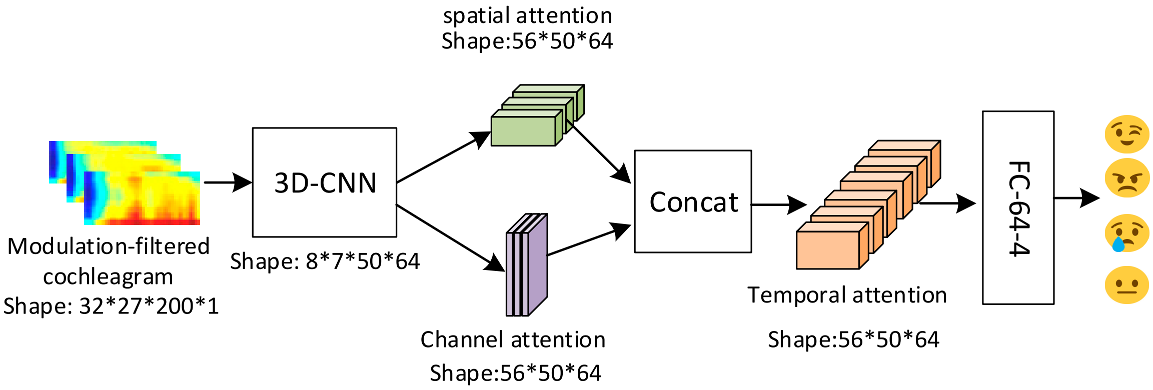


Figure 6. Flowchart of multi-level attention module. The shape of feature maps is denoted by [W, H, T, C].

4. Experimental Results and Analysis

In this section, we introduce the categorical emotion dataset and experimental result analysis in this study.

4.1. Dataset Description and Evaluation Metrics

In this study, the IEMOCAP database is used in the experiment for categorical emotion recognition. Only four emotional categories are used in this database: happy, sad, angry, and neutral. Since the speech from scripted data may contain an undesired relationship between linguistic information and the emotion labels, we only use the improvised data. We calculate MCG features from the speech signal within the IEMOCAP database and split those MCG features into 2 s segments. Each segment split from one sentence retains the same emotion label of the original sentence. The 2 s segments are performed during the training stage, while the entire sentences are used for evaluation during the testing stage.

Figure 7 illustrates the data distribution, revealing the following counts for each category: neutral (1099), happy (947), angry (289), and sad (608). Because the class distribution of the IEMOCAP database is not balanced, the number of utterances belonging to happy/neutral is more than 3-times that of angry. In this paper, unweighted accuracy (UA) is used as the performance metric of the proposed model to avoid bias towards the larger class. The calculation formula of UA is as follows:

$$acc_i = \frac{TP_i}{TP_i + FP_i} \times 100\%, \tag{15}$$

$$UA = \frac{1}{E}\sum_{i=1}^{E} acc_i \times 100\%, \tag{16}$$

where TP_i and FP_i represent the true and false positive of each category, acc_i represents the accuracy of each category, and E represents the number of emotional categories.

Additionally, the $F1$ score is also calculated to evaluate the performance of the model, as it takes into account the harmonic mean of precision and recall, providing a comprehensive measure of accuracy. The calculation formula of $F1$ score is as follows:

$$F1 = \frac{1}{E}\sum_{i=1}^{E} \frac{2TP_i}{2TP_i + FP_i + FN_i} \times 100\%, \tag{17}$$

where FN_i represents the false negative of each category.

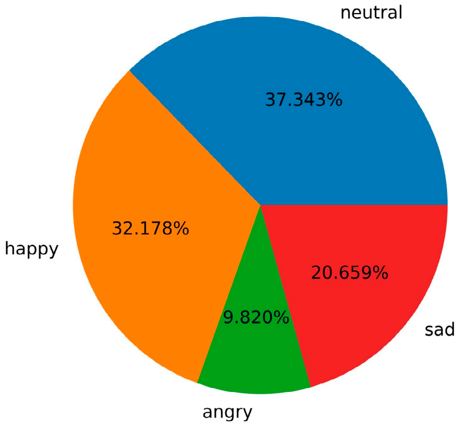


Figure 7. Distribution of the experimental data.

4.2. Experimental Setup

The deep learning model is trained using leave-one-session-out cross-validation with a batch size of 50. The TensorFlow deep learning framework is utilized to implement our methods. In all experiments, the models are trained with the Adam optimizer, employing a learning rate of 1×10^{-4} to minimize the likelihood of a cross-entropy objective. Additionally, we incorporate the ReLU activation function to introduce non-linearity into the networks. To mitigate overfitting during network training, a dropout rate of 0.5 is applied after the recurrent layer.

4.3. Experimental Results Analysis

To compare the performance of speech emotion recognition using MCG features and multi-level attention, two types of experimental comparisons are conducted. Firstly, we evaluate the emotional recognition performance of traditional acoustic features (MFCC, emobase2010, IS09 [37]), spectrograms, MSF, and MCG under the same deep model. Acoustic features are obtained by calculating the HSF using the openSMILE toolkit [38]. The spectrogram is generated by dividing the speech signal into frames and applying windowing, zero padding, and Fast Fourier Transform (FFT) to each frame. A cochleagram, which simulates the frequency selective characteristics of the human cochlea, is generated using a gammatone filterbank with 64 channels, ranging from 50 to 8000 Hz. MSF is obtained by calculating statistical features, such as spectral centroid, flatness, skewness, and kurtosis, from temporal modulation representation. All features are first normalized via specific z-normalization. For each feature set, we train convolutional recurrent neural networks (CRNNs) to recognize the speech emotion. The CRNN model consists of two convolutional blocks, one bidirectional LSTM block, and a fully connected layer. Each convolutional block consists of a convolutional layer with a convolutional kernel of 3*3, followed by a batch normalization (BN) layer, ReLU activation function layer, and a max-pooling layer.

Table 1 shows the performance comparison of the seven features on the IEMOCAP database. MFCC features yielded the lowest results at 58.5%, potentially due to their small number of 39-dimensional features compared to IS09, emobase2010, and MSF. The spectrogram achieved a slightly higher accuracy compared to MFCC. This representation captures the spectral content of the speech signal and provides better performance than MFCC. The cochleagram achieved a slightly higher accuracy than the spectrogram. This representation takes into account the frequency selectivity of the human auditory system, which contributes to its improved performance compared to the spectrogram. Among all the tested features, MCG achieved the highest accuracy of 63.8%. This outcome indicates that MCG features effectively capture emotional information within the same model.

Table 1. Performance comparison between different features on the IEMOCAP database (%).

FEATURE	UA
MFCC	58.5
emobase2010	60.9
IS09	59.4
MSF	59.7
Spectrogram	61.6
Cochleagram	62.1
MCG	63.8

Furthermore, in order to demonstrate the effectiveness of utilizing multi-level attention for extracting advanced emotional features from MCG, we conducted a comparative analysis between the proposed method and various baseline approaches. We compare our approach with several baselines.

(1) Three-dimensional CRNN max-pooling—similar to the CRNN model in hierarchical structure, but each convolutional block uses 3D convolution operations instead of 2D operations to extract high-level feature representations from MCG features. The max-pooling operation is used on the output of the LSTM network and then is fed into the fully connected layer for classifying.

(2) Three-dimensional CRNN attention—different from our proposed 3D CRNN max-pooling, the max-pooling operation is replaced with a temporal attention layer.

(3) Triple attention—the channel and spatial and temporal attention modules obtain their respective weights of the feature map in parallel, and then the concatenated attention maps are fed into the LSTM network.

Table 2 presents a performance comparison between different architectures on the IEMOCAP database, measured in terms of UA and F1 score.

Table 2. Performance comparison between different architectures on the IEMOCAP database (%).

METHOD	UA	F1 Score
3D CRNN-max-pooling	67.5	64.2
3D CRNN-attention	67.8	65.4
Triple-attention	69.4	68.1
Proposed method	71.0	69.2

The results indicate that the recognition accuracy of the 3D CNN based on max-pooling is 3.7% higher compared to the 2D CNN when using MCG features. Additionally, when incorporating the attention mechanism, the recognition accuracy of the 3D CNN increases by 4.0% compared to the 2D CNN. These findings suggest that the utilization of the 3D CNN architecture, along with max-pooling and attention mechanisms, leads to improved performance in recognizing emotions from spectral–temporal modulation representations of MCG. Moreover, the utilization of the attention method outperforms the deep model with the max-pooling operation, indicating that attention can capture discriminative emotional information from high-dimensional spatial information.

Furthermore, the results also show that the multi-level attention network outperforms other models, achieving the highest UA of 71.0% and F1 score of 69.2%. In comparison to the 3D CRNN attention model, the proposed multi-level attention network exhibits a considerable absolute improvement of +3.2% and +3.8% in UA and F1-score metrics, respectively. Similarly, when compared to the triple attention model, the multi-level attention network shows an absolute improvement of +1.6% and +1.1% in UA and F1-score metrics, respectively. These results suggest that the proposed method with the multi-level attention network is capable of extracting more informative features from speech data compared to the 3D CRNN attention and triple attention models. This indicates the superiority of the multi-level attention network in recognizing emotions, which can use a channel and spatial attention to obtain complementary attention maps and use temporal attention to obtain significant emotional regions. It can capture salient emotional information from the multi-dimensional space of the MCG feature and greatly reduce the parameters of conventional neural networks, which is important for high-dimensional representation space.

Overall, these findings support the efficacy of the proposed multi-level attention network in emotion recognition tasks, showcasing its potential for improving the accuracy and interpretability of emotion recognition systems.

In summary, Table 2 provides strong evidence for the improved performance of the proposed architecture in emotion recognition, outperforming other models in terms of both UA and F1 score. These statistical findings emphasize the effectiveness of the multi-level attention network, along with the advantages of incorporating 3D CNN, max-pooling, and attention mechanisms in recognizing emotions from MCG features.

Figure 8 presents a confusion matrix obtained from the experiments. The experimental results reveal that the proposed method obtains the highest recognition rate for sad and the lowest recognition rate for neutral emotion. There is a tendency for confusion between sad and neutral emotions, with instances of misclassification in both directions. Additionally, anger is more easily confused with happy than happy is confused with anger. In general, the ability of the multi-level attention model based on MCG features to recognize emotions is the same as that of the human auditory system.

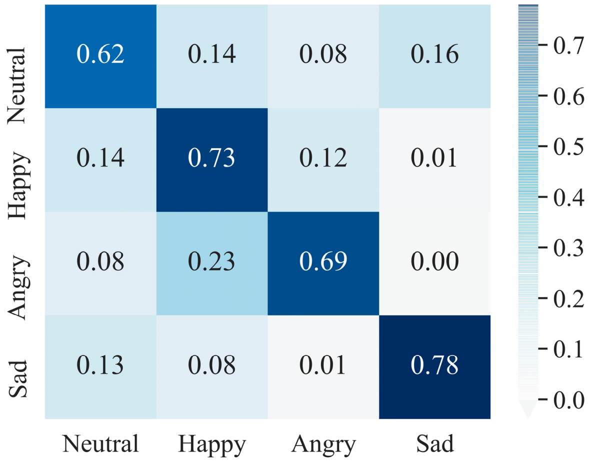


Figure 8. Confusion matrix of the multi-level attention-based emotion recognition model on the IEMOCAP dataset, where each row presents the confusion of the ground-truth emotion.

To show the benefit of the proposed model, we compare our results with the studies presented in Table 3. In [34], the authors proposed an end-to-end speech emotion recognition system using multi-level acoustic information, including MFCC, spectrogram, and wav2vec2, along with a newly designed co-attention module. In [39], the authors used log-Mel filterbank features as the input to an autoencoder and used an attentive CNN for representation learning. In [40], the authors used a 3D attention-based CRNN for learning discriminative features for SER, utilizing a Mel spectrogram with deltas and delta-deltas as inputs. In [41], the authors proposed a parallel network based on a connection attention mechanism (AMSNet) for multi-scale SER. In comparison to these studies, our approach achieves a comparable result of 71% accuracy on the IEMOCAP, employing a multi-level attention module with MCG features. This finding indicates that the MCG features provide effective spectral–temporal representations, and the multi-level attention module successfully extracts emotional information for accurate emotion recognition.

Table 3. The results of various approaches on the IEMOCAP database (%).

Literature	Features	Models	UA
Ramet et al. [39]	LLDs	ARNN	63.7
Mirsamadi et al. [22]	MFCC and spectrum	ARNN	58.8
Chen et al. [40]	Spectrogram	ACRNN	64.74 ± 5.44
Peng et al. [23]	Modulation spectrum	ASRNN	62.6
Zou et al. [34] *	wav2vec2	Co-attention	68.65
Jiang et al. [32]	Mel-spectrum	CRNN-MA	60.6
Chen et al. [41]	Spectrogram and LLDs	AMSNet	70.51
Our work	MCG	MAM	71.0

* Only list the results using the wav2vec2 in this study.

4.4. Ablation Experiment

To evaluate the effectiveness of the multi-level attention-based emotion recognition framework, this study carried out four ablation experiments, each focusing on different attention modules. The following ablation experiments were performed:

MAM: This experiment employed the multi-level attention method, incorporating channel-level, spatial-level, and temporal-level modules.

STM: This experiment utilized an attention method with spatial-level and temporal-level modules.

CTM: This experiment utilized an attention method with channel-level and temporal-level modules.

SCM: This experiment utilized an attention method with spatial-level and channel-level modules.

By conducting these ablation experiments, the study aimed to evaluate the impact of each attention module and determine the effectiveness of the multi-level attention-based emotion recognition framework.

The results of the ablation experiments are shown in Figure 9. It can be observed that both channel-level attention and spatial-level attention have similar effects on emotion recognition. On the other hand, temporal-level attention demonstrates a greater influence on emotion recognition compared to the former two attention models. However, channel-level attention and spatial-level attention have the effect of complementary information to some extent, thus strengthening the expression ability of auditory features and improving the model performance. The comparative analysis through ablation experiments shows that the multi-level attention model outperforms the individual attention modules in emotion recognition. This model exhibits better performance and acquires a more comprehensive representation of auditory emotion features. The bar chart trends in Figure 9 clearly show that the proposed emotion recognition model with the multi-level attention strategy offers a better approach in improving detection performance and enhancing accurate measurements, indicating the effectiveness of all the structures of the multi-level attention networks.

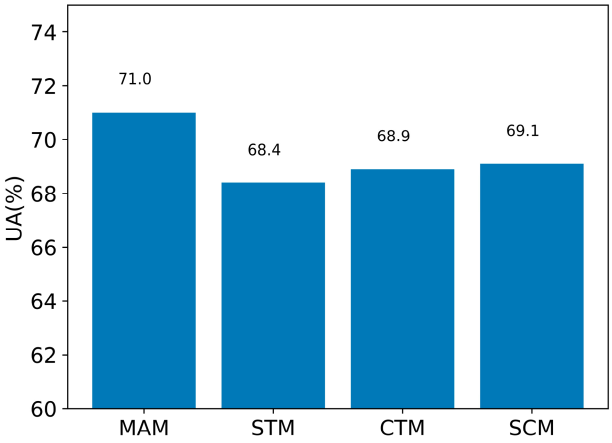


Figure 9. Results of ablation experiments.

5. Conclusions

Speech emotion recognition plays a crucial in enabling natural human–computer interaction. In this paper, we propose a multi-level attention-based framework that utilizes modulation-filtered cochleagram (MCG) features for categorical emotion recognition. Our approach takes into account channel, spatial, and temporal relationships in spectral–temporal modulation representations of MCG features. Channel-level and spatial-level attention modules are used to capture emotional saliency maps of channel and spatial feature representations from the 3D convolution feature maps. Additionally, the temporal-level attention module captures significant emotion regions. The experimental results demonstrate that our approach significantly outperforms the baseline model in terms of unweighted accuracy, highlighting the effectiveness of multi-level attention in SER. Furthermore, our proposed framework addresses the variability in emotional characteristics across time, which is an improvement on existing models. Auditory-inspired modulation-filtered cochleagram features present notable advantages in enhancing speech perception and comprehension. However, they suffer from two key limitations: high computational complexity and inefficient feature extraction. To overcome these challenges, future research endeavors should aim to investigate alternative approaches for rapid and efficient feature-processing methods. Overall, our proposed multi-level attention-based framework provides a promising approach for speech emotion recognition, and it opens avenues for further advancements in this field.

Author Contributions: Conceptualization, investigation, writing, Z.P.; supervision, J.D.; methodology, Y.L. and Y.D.; resources, W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Hunan Provincial Natural Science Foundation of China (Grant No.2021JJ30379) and was partially supported by the Youth Fund of the National Natural Science Foundation of China (Grant No. 62201571).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset used in this paper is available through the following link: (https://sail.usc.edu/iemocap/).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ACNN	Attentive convolutional neural network
ARNN	Attention-based recurrent neural network
ASRNN	Attention-based sliding recurrent neural network
BN	Batch normalization
CBAM	Convolutional block attention module
CNN	Convolutional neural network
CRNN	Convolutional and recurrent neural network
DFT	Discrete Fourier transform
ERB	Equivalent rectangular bandwidth
FFT	Fast Fourier transform
HSF	High-level statistics function
HRI	Human–robot interaction
IEMOCAP	Interactive Emotional Dyadic Motion Capture Database
IOE	Internet of Everythings
LLD	Low-level descriptors
LSTM	Long short-term memory
BLSTM	Bidirectional LSTM

MAM	multi-level attention module
MFCC	Mel frequency cepstral coefficient
MCG	Modulation-filtered cochleagram
MRCG	Multi-resolution cochleagram
MSF	Modulation spectral feature
ReLU	Rectified linear unit
RNN	Recurrent neural network
SER	Speech emotion recognition
UA	Unweighted accuracy

References

1. Zorzos, I.; Kakkos, I.; Ventouras, E.M.; Matsopoulos, G.K. Advances in Electrical Source Imaging: A Review of the Current Approaches, Applications and Challenges. *Signals* **2021**, *2*, 378–391. [CrossRef]
2. Du, Y.; Lim, Y.; Tan, Y. A Novel Human Activity Recognition and Prediction in Smart Home Based on Interaction. *Sensors* **2019**, *19*, 4474. [CrossRef]
3. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proceedings of the 2014 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
4. Neumann, M.; Vu, N.T. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. In Proceedings of the 2017 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 1263–1267. [CrossRef]
5. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213. [CrossRef]
6. Lim, W.; Jang, D.; Lee, T. Speech emotion recognition using convolutional and Recurrent Neural Networks. In Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Republic of Korea, 13–16 December 2016; pp. 1–4. [CrossRef]
7. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
8. Chen, J.; Wang, Y.; Wang, D. A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1993–2002. [CrossRef]
9. Santoro, R.; Moerel, M.; De Martino, F.; Goebel, R.; Ugurbil, K.; Yacoub, E.; Formisano, E. Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLoS Comput. Biol.* **2014**, *10*, e1003412. [CrossRef]
10. Zhu, Z.; Nishino, Y.; Miyauchi, R.; Unoki, M. Study on linguistic information and speaker individuality contained in temporal envelope of speech. *Acoust. Sci. Technol.* **2016**, *37*, 258–261. [CrossRef]
11. McDermott, J.H.; Simoncelli, E.P. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron* **2011**, *71*, 926–940. [CrossRef]
12. Wu, S.; Falk, T.H.; Chan, W.Y. Automatic speech emotion recognition using modulation spectral features. *Speech Commun.* **2011**, *53*, 768–785. [CrossRef]
13. Avila, A.R.; Momin, Z.A.; Santos, J.F.; OShaughnessy, D.; Falk, T.H. Feature Pooling of Modulation Spectrum Features for Improved Speech Emotion Recognition in the wild. *IEEE Trans. Affect. Comput.* **2018**, *3045*, 177–188. [CrossRef]
14. Tamulevičius, G.; Korvel, G.; Yayak, A.B.; Treigys, P.; Bernatavičienė, J.; Kostek, B. A Study of Cross-Linguistic Speech Emotion Recognition Based on 2D Feature Spaces. *Electronics* **2020**, *9*, 1725. [CrossRef]
15. Borwankar, S.; Shah, D.; Verma, J.P.; Tanwar, S. Automatic Speech Emotion Recognition Using Cochleagram Features. In *Recent Innovations in Computing*; Singh, P.K., Singh, Y., Chhabra, J.K., Illés, Z., Verma, C., Eds.; Springer: Singapore, 2022; pp. 453–466.
16. Peng, Z.; Dang, J.; Unoki, M.; Akagi, M. Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech. *Neural Netw.* **2021**, *140*, 261–273. [CrossRef] [PubMed]
17. Keren, G.; Schuller, B. Convolutional RNN: An enhanced model for extracting features from sequential data. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 3412–3419. [CrossRef]
18. Satt, A.; Rozenberg, S.; Hoory, R. Efficient emotion recognition from speech using deep learning on spectrograms. In Proceedings of the 2017 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 1089–1093.
19. Alluhaidan, A.S.; Saidani, O.; Jahangir, R.; Nauman, M.A.; Neffati, O.S. Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. *Appl. Sci.* **2023**, *13*, 4750. [CrossRef]
20. Ghosh, R.; Phadikar, S.; Deb, N.; Sinha, N.; Das, P.; Ghaderpour, E. Automatic Eyeblink and Muscular Artifact Detection and Removal From EEG Signals Using k-Nearest Neighbor Classifier and Long Short-Term Memory Networks. *IEEE Sens. J.* **2023**, *23*, 5422–5436. [CrossRef]
21. Kim, Y.; Choi, A. EEG-Based Emotion Classification Using Long Short-Term Memory Network with Attention Mechanism. *Sensors* **2020**, *20*, 6727. [CrossRef]

22. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.
23. Peng, Z.; Li, X.; Zhu, Z.; Unoki, M.; Dang, J.; Akagi, M. Speech Emotion Recognition Using 3D Convolutions and Attention-Based Sliding Recurrent Networks With Auditory Front-Ends. *IEEE Access* **2020**, *8*, 16560–16572. [CrossRef]
24. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
25. Wu, W.; Zhang, Y.; Wang, D.; Lei, Y. SK-Net: Deep learning on point cloud via end-to-end discovery of spatial keypoints. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 6422–6429.
26. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 13–19.
27. Xu, K.; Wang, Z.; Shi, J.; Li, H.; Zhang, Q.C. A2-net: Molecular structure estimation from cryo-em density volumes. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 1230–1237.
28. Xue, H.; Liu, C.; Wan, F.; Jiao, J.; Ji, X.; Ye, Q. Danet: Divergent activation for weakly supervised object localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6589–6598.
29. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
30. Ma, W.; Cui, Y.; Shao, N.; He, S.; Zhang, W.-N.; Liu, T.; Wang, S.; Hu, G. TripleNet: Triple attention network for multi-turn response selection in retrieval-based chatbots. *arXiv* **2019**, arXiv:1909.10666.
31. Liu, Z.; Zhao, X.; Huang, T.; Hu, R.; Zhou, Y.; Bai, X. Tanet: Robust 3d object detection from point clouds with triple attention. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11677–11684.
32. Jiang, P.; Xu, X.; Tao, H.; Zhao, L.; Zou, C. Convolutional-Recurrent Neural Networks with Multiple Attention Mechanisms for Speech Emotion Recognition. *IEEE Trans. Cogn. Dev. Syst.* **2021**, *14*, 1564–1573. [CrossRef]
33. Li, X.; Zhao, B.; Lu, X. MAM-RNN: Multi-level attention model based RNN for video captioning. *IJCAI Int. Jt. Conf. Artif. Intell.* **2017**, *2017*, 2208–2214.
34. Zou, H.; Si, Y.; Chen, C.; Rajan, D.; Chng, E.S. Speech emotion recognition with co-attention based multi-level acoustic information. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 7367–7371.
35. Glasberg, B.R.; Moore, B.C.J. Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* **1990**, *47*, 103–138. [CrossRef] [PubMed]
36. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
37. Schuller, B.; Steidl, S.; Batliner, A. The interspeech 2009 emotion challenge. In Proceedings of the 2009 10th Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009.
38. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, New York, NY, USA, 25–29 October 2010; pp. 1459–1462.
39. Ramet, G.; Garner, P.N.; Baeriswyl, M.; Lazaridis, A. Context-aware attention mechanism for speech emotion recognition. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 126–131.
40. Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1440–1444. [CrossRef]
41. Chen, Z.; Li, J.; Liu, H.; Wang, X.; Wang, H.; Zheng, Q. Learning multi-scale features for speech emotion recognition with connection attention mechanism. *Expert Syst. Appl.* **2023**, *214*, 118943. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.