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

Zhichao Peng *, Wenhua He , Yongwei Li , Yegang Du , Jianwu Dang *

Posted Date: 1 May 2023

doi: 10.20944/preprints202305.0003.v1

Keywords: Categorical emotion recognition; Auditory signal processing; Modulation-filtered cochleagram; Multi-level attention

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, Hunan, China; zcpeng@tju.edu.cn
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China; yongwei.li@nlpr.ia.ac.cn
[3] Future Robotics Organization, Waseda University, Tokyo, Japan; yg.du@aoni.waseda.jp
[4] College of Intelligence and Computing, Tianjin University, Tianjin, China; jdang@jaist.ac.jp
[5] Pengcheng Laboratory, Shenzhen, China; jdang@jaist.ac.jp
* Correspondence: jdang@jaist.ac.jp, zcpeng@tju.edu.cn

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 recognize emotions from the modulation-filtered cochleagram. The channel-level attention and spatial-level attention modules are used to capture emotional saliency maps of channel and spatial feature representations from the 3D convolution feature maps, respectively. Furthermore, the temporal-level attention module captures significant emotional regions from the concatenated feature sequence of the emotional saliency maps. Our experiments on the 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 a better unweighted accuracy of 71% in categorical emotion recognition than several existing approaches in the experiments. 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

1. Introduction

The Internet of Everything (IoE) presents a plethora of opportunities for human-robot interaction. Speech is the most natural and convenient communication mode between humans and robots. Emotion information from speech can effectively help robots understand the speaker's intentions in natural human-robot interaction. Therefore, speech emotion recognition (SER) holds immense potential for diverse applications in human-robot interaction, including but not limited to intelligent driving, service robotics, online education, telemedicine, and criminal investigations[1].

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)[2], Mel-filter bank features[3], Spectrogam[4][5], etc. However, researchers have not identified the best speech features for SER, and still explore the effective features that can represent emotional states[6]. 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 can be an effective approach. However, the complexity of the human auditory system and its signal processing mechanism remain unclear. Researchers have simulated 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[7], etc, but there are still few studies on 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[8]. This modulation spectrum plays an essential role in speech perception and understanding[9][10]. Several studies have used statistical functions on the modulation spectrum to obtain modulation spectral features (MSF) for SER tasks [11]. Avila et al.[12] proposed a feature pooling scheme for dimensional emotion recognition using combined MSF and 3D spectrum representation. They extracted the amplitude envelope of the Gammatone auditory filterbank and applied discrete Fourier transform (DFT) to obtain the spectral-temporal modulation representation. However, this method uses DFT to convert the envelope signal into the frequency domain before temporal modulation, thus increasing the computational complexity. Peng et al. [13] proposed the 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 investigation.

In order to extract high-level feature representations from speech features, deep learning methods such as rr neural network (RNN), Transformer, etc., are mainly used for the SER task. CNN is often used to extract high-level speech feature representation because of its scale and rotation invariance[14]. RNN is often used to capture the sequence dependence[15][16] because of its long-term dependence in the speech sequence. Recently, attention mechanisms have been incorporated into deep learning methods to automatically capture salient emotion features in speech sequences. Neumann et al.[3] proposed attentive CNN (ACNN) based on the attention model to identify emotions from the log-Mel filterbank features. Mirsamadi et al.[17] introduced attentive RNN (ARNN) model recognize emotions from frame-level LLDs with local attention as a weighted pooling method. Peng et al.[18] proposed an attention-based sliding recurrent neural network (ASRNN) to effectively model auditory representation sequence by mimicking the auditory attention to capture salient emotion regions. In addition, 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. [7] introduced Key-Sparse Transformer to dynamically judge 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 is used to obtain the importance of different channels, such as SE-Net[19], SK-Net[20], and ECA-Net[21]. Spatial attention is transformed into another space through the spatial conversion module and retains key information, such as A2-Net[22], DANet[23], and convolutional block attention module(CBAM)[24]. In addition, some studies have constructed multi-level attention models from different dimensions. Ma et al.[25] proposed TripleNet that uses a hierarchical representation module to construct the representation of context, reply and query in multi-turn dialogue, in which the triple attention mechanism is applied to update the representation. Liu et al.[26] proposed TANet for object detection by jointly considering the triple attention of channel, point and voxel. for speech dialogue and object detection. Jiang et al.[27] 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 the self-attention layer to calculate the weights from Mel-spectrum features. Liu et al. [28] proposed a novel multi-level attention network, which contains a multiscale low-level feature extractor and a multi-unit attention module for SER. Zou et al.[29] proposed an end-to-end speech emotion recognition system using multi-level acoustic information with a newly designed co-attention module. These methods used multiple attention models to extract different channel and spatial attention maps from LLDs, spectrograms, and waveforms, and then fused these attention maps to recognize emotions, without considering capturing significant emotional regions of speech

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2023 doi:10.20944/preprints202305.0003.v1

sequences using temporal attention. To address this issue and investigate the effectiveness of MCG features in discrete emotion recognition, 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, 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 spectral-temporal feature representation. Finally, a temporal-level attention module is used to capture the significant emotional regions from the concatenated feature sequence of 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, 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, thereby improving emotion recognition performance.
- The proposed method is evaluated on Interactive Emotional Dyadic Motion Capture Database (IEMOCAP). It obtains an unweighted accuracy of 71%, showing the effectiveness of our approach.

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

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 application of the modulation-filtered cochleagram features in categorical emotion recognition. The emotional speech signal $s(t)$ is first filtered by 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 n-th channel envelope signal is used to obtain the spectral-temporal modulation signal $s_{mu}(n,m,i)$, it is defined as:

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

where $\mathrm{w}(t_w)$ is the window function, t_w is the time window size, and Len_s is the frame shift. $s_{mu}(n,m,i)$ refer 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_t/Len_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. $s_{mu}(n,m,i)$ represent the m modulation subchannel in the n acoustic subchannel. The calculation formula is as follows:

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2023 doi:10.20944/preprints202305.0003.v1

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

The $s_g(n,t)$ represents the speech signal s(t) of the n th channel of the speech signal processed by the auditory filter, using the following formula:

$$s_g(n,t) = g_t(n,t) * s(t), \ 1 \le n \le N, \quad (4)$$

where $g_t(n,t)$ represents the pulse response of the n th 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 cochlear basilar membrane, and its pulse response is the product of the Gamma distribution and the cosine signal:

$$g_t(n,t) = At^{n_f-1}\exp\left(-2\pi w_f ERB_N(f_n)t\right)\cos(2\pi f_n t + \varphi), \quad (5)$$

where A, n_f and w_f are the amplitude, order and bandwidth of the filter, $At^{n_f-1}\exp\left(-2\pi w_f ERB_N(f_n)t\right)$ is the amplitude term of the Gamma distribution representation, f_n is the central frequency of the n th channel of the filter, and $ERB_N(f_n)$ 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:

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

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 [30], with 9.26449 and 24.7, respectively.

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

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

2.2. MCG feature representation of different emotions

In the MCG feature, the weight of emotions expressed by different channels is different, mainly focusing on low-modulation frequency channels of about 4 Hz, in which neutral emotion and sadness are expressed at the lower modulation frequency, while anger and happiness are opposite [10]. Figure 1 shows examples of the MCG feature of the first modulation channel in different emotion speech on the IEMOCAP dataset [31]. The x axis represents the speech sequence, and the y axis is the number of acoustic channels n (n=16). Panels (a) to (d) in Figure 1 show the modulation-filtered cochleagram of sadness, anger, neutral emotion, and happiness, respectively. From these panels, we can find that the different emotion has a different acoustic channel, suggesting they could be discriminated from each other from MCG features. From the cochleagram, the energy of sadness is concentrated in the slow acoustic channel, and the energy of anger and happiness is concentrated in the higher acoustic channel. However, compared with 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2023 doi:10.20944/preprints202305.0003.v1

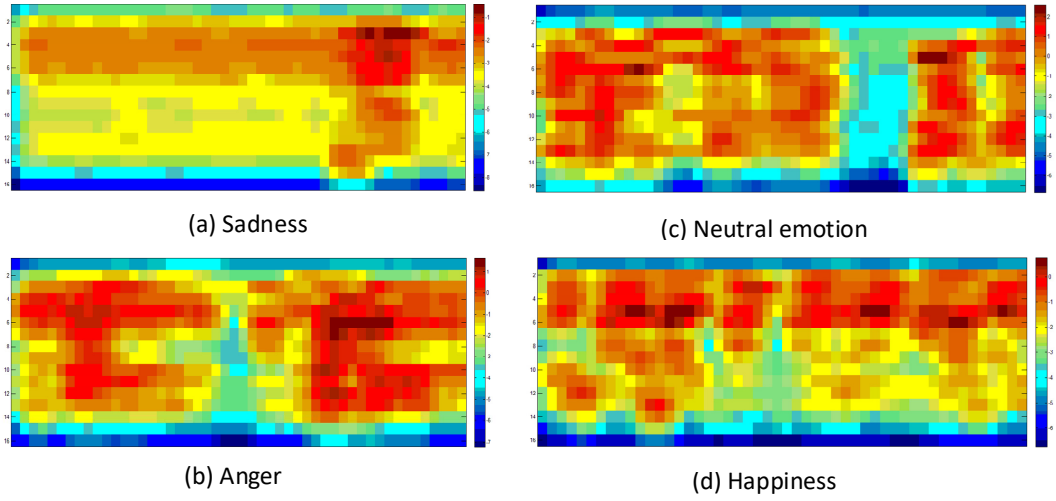

Figure 1. Modulation-filtered cochleagram feature representation of different emotions.

3. Emotional recognition model

In this section, we introduce 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 by 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 spectral-temporal feature representation, 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 to capture the dependencies between the channel and spatial dimension of the feature map in a parallel mode, respectively. Finally, attention-based feature representations are obtained through temporal-level attention and further passed to a softmax layer to generate the emotional state distribution.

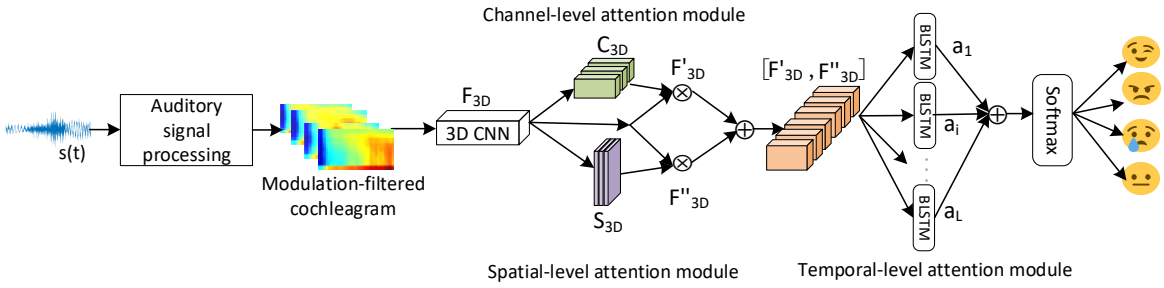

Figure 2. Overview of multi-level attention-based emotion recognition model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2023 doi:10.20944/preprints202305.0003.v1

3.2. Channel-level attention

A channel-level attention module is used to calculate the channel-wise attention map from the 3D convolution feature map. Channel-level attention adaptively recalibrates the weights of each channel to focuses on what is an informative part. The channel-level attention module is designed similarly to CBAM. The main difference between channel-level attention and CBAM is that we insert two 3D convolutional layers to characterize spatial and temporal information of feature maps for channel-wise features. To compute the channel-level attention efficiently, we squeeze the spatial and temporal dimension of the input feature map. The channel-level attention module is shown in Figure 3. Channel-level attention map is first obtained by adaptive learning, and then element-level multiplication with the input feature map F_{3D} to obtain a refined feature map F'_{3D}. The calculation formula is provided as follows:

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

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

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 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:

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

where $Conv_1$ and $Conv_2$ represent the first and second 3D convolution operations, respectively, and σ denotes a sigmoid operation. Both convolutions are 1x1x1 convolution kernels, the number of output channels is $\frac{C}{r}$ and C, 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.

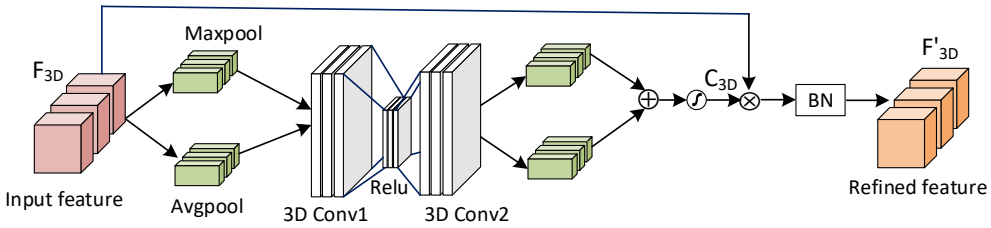

Figure 3. The channel-level attention module.

3.3. Spatial-level attention

A spatial-level attention module is used to calculate the spatial-wise attention map from the 3D convolution feature map. Different from channel-level attention, spatial attention focuses on where is an informative part of feature maps, which is complementary to channel-level attention. The spatial-level attention module is shown in Figure 4. The spatial-level attention map generated by the spatial-level attention is used element-level multiplication with the F_{3D} to obtain a refined feature map F''_{3D}. The calculation formula is provided as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2023 doi:10.20944/preprints202305.0003.v1

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

where S_{3D} represents a spatial-level attention map, $\otimes$ representing element-level multiplication. The feature map F_{3D} integrates the feature map through maximum pooling and average pooling, respectively, to obtain global information. 3D convolution with a kernel size 3x3x1 is used to obtain spatial regions of emotionally significant spectral-temporal space, thus obtaining spatial-level attention map $S_{3D} \in R^{W \times H \times 1 \times 1}$. The spatial-level attention the map represents the importance of each region in the feature map F_{3D}. The calculation formula is provided as follows:

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

where $f^{3\times3\times1}$ is a convolution kernel of size 3x3x1.

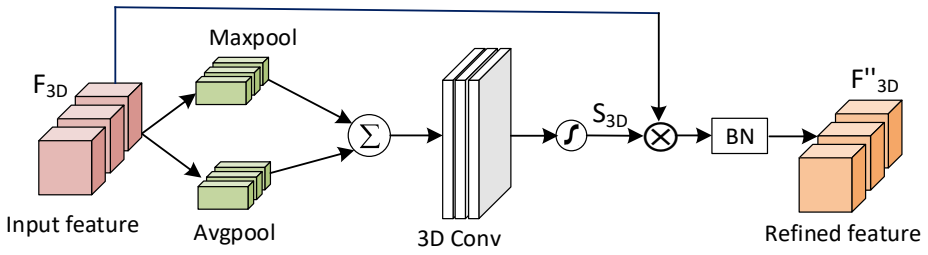

Figure 4. The spatial-level attention module.

3.4. Temporal-level attention

Because many speech frames are unrelated to the expressed emotion, such as silence, temporal-level attention is mainly used to focus on the significant emotional regions from the concatenation feature of the channel attention map F'_{3D} and the spatial attention map F''_{3D}.The temporal-level attention module is shown in Figure 5. Specifically, we use a bidirectional LSTM (BLSTM) network in this study, where the sequence of received signals is once fed in the forward direction into one LSTM cell, and once fed backward into another LSTM cell. We concatenate the last state of the forward and backward LSTM cells to produce a sequence of h_i. Subsequently, a ReLU 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, U_i are the trainable parameter matrices, 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.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2023 doi:10.20944/preprints202305.0003.v1

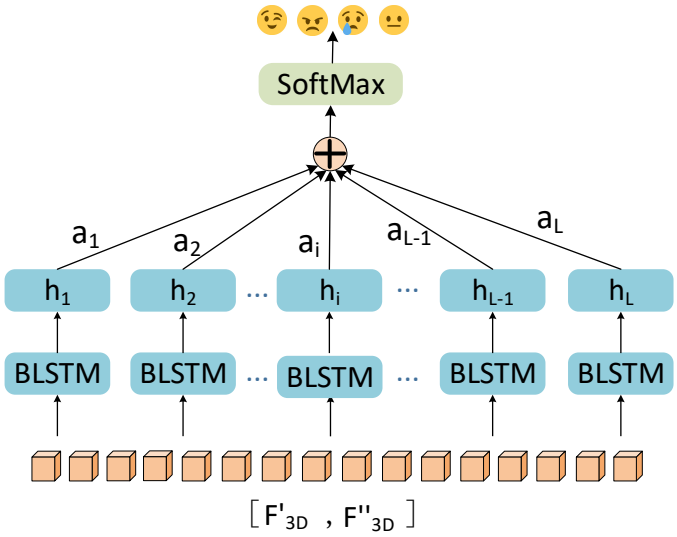

Figure 5. The temporal-level attention module.

4. Experimental results and analysis

In this section, we introduce the categorical emotion datasets and experimental results analysis in this study.

4.1. Dataset description

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 in IEMOCAP and split those MCG features into 2-second segments. Segments split from one sentence uses the same emotion label as the original sentence. The 2-second segment is performed during the training stage, while the entire sentence is used for evaluation during the testing stage. The data distribution is shown in Figure 6, where neutral, happy, angry and sad are 1099, 947, 289, 608, respectively. Because the class distribution of 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 the bias towards the larger class.

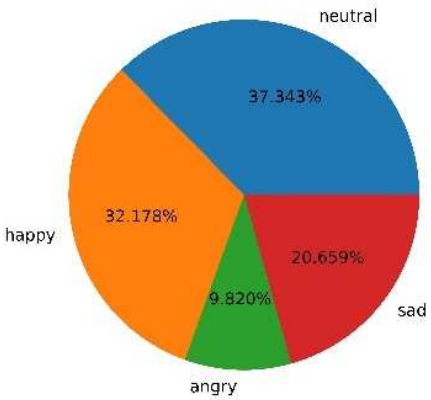

Figure 6. Distribution of the experimental data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2023 doi:10.20944/preprints202305.0003.v1

4.2. Experimental setup

The deep learning model is trained using leave-one-session-out cross-validation with a batch size of 50. We implement our methods with the TensorFlow deep learning framework. Our approach is implemented through the TensorFlow framework. The models were trained in all experiments using the Adam optimizer with a learning rate of 1e-4 to minimize the chance of having a cross-entropy objective. Moreover, we used a ReLU as the activation function, which brought the non-linearity into the networks. To avoid overfitting when training our networks, we used a dropout rate of 0.5 after the recurrent layer.

4.3. Experimental results analysis

To compare the performance of speech emotion recognition on MCG features and multi-level attention, two types of experimental comparisons are designed. Firstly, we investigate the emotional recognition performance of traditional acoustic features (MFCC, emobase2010, IS09[32]), spectrograms, MSF, and MCG under the same deep model. Acoustic features are obtained by calculating the HSF using the openSMILE toolkit[33]. The spectrogram is obtained by dividing the speech signal into frames and performing windowing, zero padding, and Fast Fourier Transform (FFT) on each frame. Cochleagram, simulating the frequency selective characteristics of the human cochlea, is generated using a gammatone filterbank having 64 channels from frequency 50 to 8000 Hz from speech signal. MSF is obtained by calculating the spectral centroid, flatness, skewness, kurtosis, and other statistical features from temporal modulation representation. All features are first normalized by specific z-normalization. For each feature set, we train convolutional recurrent neural networks (CRNN) to recognize the speech emotion. The CRNN model consists of two convolutional blocks, one bidirectional LSTM block, and a fully connected layer. Each of the convolutional block consists of a convolutional layer with a convolutional kernel of 3*3 followed by a Batch Normalization (BN) layer, a ReLU activation function layer, and a max-pooling layer. Table 1 shows the performance comparison of the seven features on the IEMOCAP database. The MFCC features yielded the worst results at 58.5% compared to IS09, emobase2010, and MSF, possibly due to the least number of 39-dimensional MFCC features. The best result in IEMOCAP was generated on MCG features with an accuracy of 63.8%, indicating that MCG features can effectively capture emotional information under the same model.

We compare our approach with several baselines. 1) 3D 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 LSTM network, and then is fed into the fully connected layer for classifying. 2) 3D CRNN-attention. Different from our proposed 3D CRNN-max-pooling, the max-pooling operation is replaced with a temporal attention layer. 2) Triple-attention. The channel, spatial, and temporal attention modules obtain their respective weights of feature map in parallel, and then the concatenated attention maps are fed into the LSTM network. The results obtained for each method are shown in Table 2. The results show that 3D convolution has a significant improvement in recognition performance compared to 2D convolution, indicating that 3D convolution obtains more spectral time spatial information. The use of attention method has a higher recognition rate than deep model with the max-pooling operation, indicating that attention can capture discriminative emotional information from high-dimensional spatial information. The results also show that the multi-level attention network achieves the best performance with 71.0% in UA measures. This indicates that multi-level attention methods can use a channel and spatial attention to obtain complementary attention maps and use temporal attention to obtain significant emotional regions.

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

FEATURE	UA
MFCC	58.5

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2023 doi:10.20944/preprints202305.0003.v1

emobase2010	60.9
IS09	58.4
MSF	59.7
Spectrogram	61.6
Cochleagram	62.1
MCG	63.8

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

METHOD	UA
3D CRNN-max-pooling	67.5
3D CRNN-attention	67.8
Triple-attention	69.4
Proposed method	71.0

The confusion matrix is shown in Figure 7. The experimental results show that the proposed method obtains the highest recognition rate on Sad and the lowest recognition rate on Neutral emotion. Sad is easily confused with Neutral emotion and vice versa. 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.

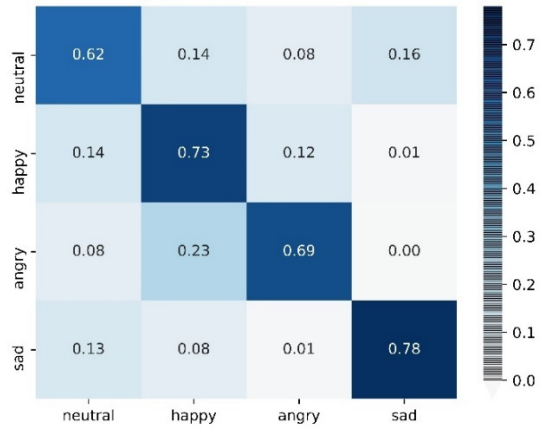

Figure 7. Confusion matrix of the multi-level attention-based emotion recognition model on the IEMOCAP dataset, where each row presents the confusion of the ground-truth emotion.

To show the benefit of the proposed model, we compare our results with the studies presented in Table 3. In [29], the authors proposed an end-to-end speech emotion recognition system using multi-level acoustic information including MFCC, spectrogram, and wav2vec2 with a newly designed co-attention module. In [34], the authors used Log-Mel filterbank features as the input to the autoencoder and used attentive CNN for representation learning. In [35], the authors used a 3D attention-based CRNN to learn discriminative features for SER, where the Mel-spectrogram with deltas and delta-deltas are used as input. In [36], the authors proposed a parallel network based on a connection attention mechanism (AMSNet) for multi-scale SER. Compared to these studies, we are achieving a better result of 71% on the IEMOCAP using a multi-level attention module from MCG features. This indicates that the MCG features can provide spectral-temporal representations, and the multi-level attention module can effectively extract emotional information for emotion recognition.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2023 doi:10.20944/preprints202305.0003.v1

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

Literature	Features	Models	UA
Ramet et al.[34]	LLDs	ARNN	63.7
Mirsamadi et al.[17]	MFCC and spectrum	ARNN	58.8
Chen et al.[35]	Spectrogram	ACRNN	64.74±5.44
Peng et al. [18]	Modulation spectrum	ASRNN	62.6
Zou et al.[29]	wav2vec2	Co-attention	68.65
Jiang et al.[27]	Mel-spectrum	CRNN-MA	60.6
Chen et al. [36]	Spectrogram and LLDs	AMSNet	70.51
Our work	MCG	MAM	71.0

4.4. Ablation experiment

To evaluate the effectiveness of the multi-level attention-based emotion recognition framework, this study carried out four ablation experiments on different attention modules.

MAM: the multi-level attention model with channel-level, spatial-level, and temporal-level modules.

STM: the attention model with spatial-level and temporal-level modules.

CTM: the attention model with channel-level and temporal-level modules.

SCM: the attention model with spatial-level and channel-level modules.

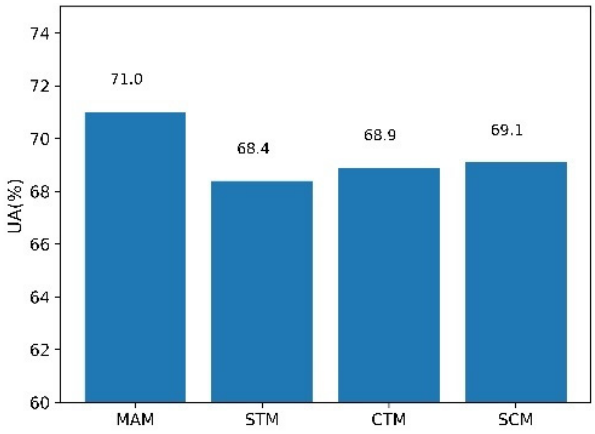

Figure 8. Results of ablation experiments.

The results are shown in Figure 8. Channel-level attention and spatial-level attention have similar effects on emotion recognition, while temporal-level attention has more influence on emotion recognition than 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. Comparative analysis through ablation experiments shows that the multi-level attention model has better emotion recognition performance and can acquire a more general representation of auditory emotion features. The bar chart trends in Figure 8 clearly show that the proposed emotion recognition model with the multi-level attention strategy offers a better approach in improving detection performance on all datasets and enhancing accurate measurements. Meanwhile, it is found in Figure 8 that the results of

the proposed model are improved significantly, indicating the effectiveness of all the structures of the multi-level attention networks.

5. Conclusions

Speech emotion recognition is critical in enabling natural human-computer interaction. In this paper, we propose a multi-level attention-based framework that utilizes modulation-filtered cochleagram features for categorical emotion recognition. Our approach takes into account channel, spatial, and temporal relationships in speech features, with channel-level and spatial-level attention used to capture emotional saliency maps of channel and spatial feature representations from the 3D convolution feature maps, and temporal-level attention capturing significant emotion regions. The experimental results demonstrate that our approach significantly outperforms the baseline model on 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 over existing models. Future work will explore the extension of our multi-level attention mechanism to capture emotions that exhibit varying characteristics over time.

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

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

References

1. Y. Du, Y. Lim, and Y. Tan, "A Novel Human Activity Recognition and Prediction in Smart Home Based on Interaction," *Sensors*, vol. 19, no. 20, 2019, doi: 10.3390/s19204474.
2. K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," 2014.
3. M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *Proc. INTERSPEECH 2017 18th Annu. Conf. Int. Speech Commun. Assoc.*, vol. 2017-Augus, no. 3, pp. 1263–1267, 2017, doi: 10.21437/Interspeech.2017-917.
4. Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimed.*, vol. 16, no. 8, pp. 2203–2213, 2014.
5. W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," *Proc. 8th Asia-Pacific Signal Inf. Process. Assoc. Annu. Conf.*, pp. 1–4, 2016, doi: 10.1109/APSIPA.2016.7820699.
6. M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011, doi: 10.1016/j.patcog.2010.09.020.
7. J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1993–2002, 2014.
8. R. Santoro *et al.*, "Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex," *PLoS Comput. Biol.*, vol. 10, no. 1, 2014, doi: 10.1371/journal.pcbi.1003412.
9. Z. Zhu, Y. Nishino, R. Miyauchi, and M. Unoki, "Study on linguistic information and speaker individuality contained in temporal envelope of speech," *Acoust. Sci. Technol.*, vol. 37, no. 5, pp. 258–261, 2016.
10. J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.
11. S. Wu, T. H. Falk, and W. Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, vol. 53, no. 5, pp. 768–785, 2011, doi: 10.1016/j.specom.2010.08.013.
12. A. R. Avila, Z. A. Momin, J. F. Santos, D. OShaughnessy, and T. H. Falk, "Feature Pooling of Modulation Spectrum Features for Improved Speech Emotion Recognition in the wild," *IEEE Trans. Affect. Comput.*, vol. 3045, no. c, pp. 1–1, 2018, doi: 10.1109/TAFFC.2018.2858255.

13. Z. Peng, J. Dang, M. Unoki, and M. Akagi, "Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech," *Neural Networks*, vol. 140, pp. 261–273, 2021, doi: 10.1016/j.neunet.2021.03.027.
14. A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network," *Appl. Sci.*, vol. 13, no. 8, 2023, doi: 10.3390/app13084750.
15. G. Keren and B. Schuller, "Convolutional RNN: An enhanced model for extracting features from sequential data," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2016-Octob, pp. 3412–3419, 2016, doi: 10.1109/IJCNN.2016.7727636.
16. A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," *Proc. Interspeech 2017*, pp. 1089–1093, 2017.
17. S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *in Proc. 40th IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 2227–2231.
18. Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech Emotion Recognition Using 3D Convolutions and Attention-Based Sliding Recurrent Networks With Auditory Front-Ends," *IEEE Access*, vol. 8, pp. 16560–16572, 2020.
19. J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," *CoRR*, vol. abs/1709.0, 2017, [Online]. Available: http://arxiv.org/abs/1709.01507
20. W. Wu, Y. Zhang, D. Wang, and Y. Lei, "SK-Net: Deep learning on point cloud via end-to-end discovery of spatial keypoints," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 04, pp. 6422–6429.
21. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Supplementary material for 'ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, WA, USA*, 2020, pp. 13–19.
22. K. Xu, Z. Wang, J. Shi, H. Li, and Q. C. Zhang, "A2-net: Molecular structure estimation from cryo-em density volumes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 01, pp. 1230–1237.
23. H. Xue, C. Liu, F. Wan, J. Jiao, X. Ji, and Q. Ye, "Danet: Divergent activation for weakly supervised object localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6589–6598.
24. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
25. W. Ma *et al.*, "TripleNet: Triple attention network for multi-turn response selection in retrieval-based chatbots," *arXiv Prepr. arXiv1909.10666*, 2019.
26. Z. Liu, X. Zhao, T. Huang, R. Hu, Y. Zhou, and X. Bai, "Tanet: Robust 3d object detection from point clouds with triple attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 07, pp. 11677–11684.
27. P. Jiang, X. Xu, H. Tao, L. Zhao, and C. Zou, "Convolutional-Recurrent Neural Networks with Multiple Attention Mechanisms for Speech Emotion Recognition," *IEEE Trans. Cogn. Dev. Syst.*, vol. 14, no. 4, pp. 1564–1573, 2021.
28. X. Li, B. Zhao, and X. Lu, "MAM-RNN: Multi-level attention model based RNN for video captioning," *IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 2208–2214, 2017.
29. H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7367–7371.
30. B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1–2, pp. 103–138, 1990.
31. C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.
32. B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," 2009.
33. F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM int. conf. Multimedia*, 2010, pp. 1459–1462.
34. G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 126–131.

35. M. Chen, X. He, J. Yang, and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, 2018, doi: 10.1109/LSP.2018.2860246.
36. Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, and Q. Zheng, "Learning multi-scale features for speech emotion recognition with connection attention mechanism," *Expert Syst. Appl.*, vol. 214, p. 118943, 2023.