electronics

MDPI

Article

Enhancing Dimensional Emotion Recognition from Speech through Modulation-Filtered Cochleagram and Parallel Attention Recurrent Network

Zhichao Peng [1,*], Hua Zeng [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; zenghua@huhst.edu.cn
[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 300072, China
[5] Pengcheng Laboratory, Shenzhen 518066, China
* Correspondence: zcpeng@tju.edu.cn (Z.P.); jdang@jaist.ac.jp (J.D.)

Abstract: Dimensional emotion can better describe rich and fine-grained emotional states than categorical emotion. In the realm of human–robot interaction, the ability to continuously recognize dimensional emotions from speech empowers robots to capture the temporal dynamics of a speaker's emotional state and adjust their interaction strategies in real-time. In this study, we present an approach to enhance dimensional emotion recognition through modulation-filtered cochleagram and parallel attention recurrent neural network (PA-net). Firstly, the multi-resolution modulation-filtered cochleagram is derived from speech signals through auditory signal processing. Subsequently, the PA-net is employed to establish multi-temporal dependencies from diverse scales of features, enabling the tracking of the dynamic variations in dimensional emotion within auditory modulation sequences. The results obtained from experiments conducted on the RECOLA dataset demonstrate that, at the feature level, the modulation-filtered cochleagram surpasses other assessed features in its efficacy to forecast valence and arousal. Particularly noteworthy is its pronounced superiority in scenarios characterized by a high signal-to-noise ratio. At the model level, the PA-net attains the highest predictive performance for both valence and arousal, clearly outperforming alternative regression models. Furthermore, the experiments carried out on the SEWA dataset demonstrate the substantial enhancements brought about by the proposed method in valence and arousal prediction. These results collectively highlight the potency and effectiveness of our approach in advancing the field of dimensional speech emotion recognition.

Keywords: modulation-filtered cochleagram; parallel attention recurrent neural network; dimensional emotion recognition; auditory signal processing; noise-robust

Citation: Peng, Z.; Zeng, H.; Li, Y.; Du, Y.; Dang, J. Enhancing Dimensional Emotion Recognition from Speech through Modulation-Filtered Cochleagram and Parallel Attention Recurrent Network. *Electronics* **2023**, *12*, 4620. https://doi.org/10.3390/electronics12224620

Academic Editor: Wojciech M. Zabołotny

Received: 24 September 2023
Revised: 31 October 2023
Accepted: 9 November 2023
Published: 12 November 2023

1. Introduction

The utilization of vocal emotion cues proves highly advantageous in helping robots or virtual agents to understand speakers' true intentions. Hence, the exploration of emotion recognition in human speech emerges as an area of significant research interest within the domain of natural human–robot interaction (HRI). Categorical emotions and dimensional emotions are the two main ways of describing emotional states. Dimensional emotions describe emotional states as points in a multidimensional emotional space, with each dimension corresponding to a different psychological attribute of the emotion [1]. In HRI, continuous dimensional emotion can help a robot capture the temporal dynamics of a speaker's emotional state and adjust both the manner of the interaction and its content in real time according to the changing state [2]. Therefore, dimensional emotion can better

meet the needs of HRI than categorical emotion. Researchers have accordingly shown an increasing interest in the representation and recognition of dimensional emotions [3]. Valence and arousal are the two most basic primitive forms in the dimensional emotional space. Valence represents the subjective evaluation or experience of positive or negative emotions. Arousal represents the high or low intensity level of emotional arousal. Speech is the most direct and effective way to achieve natural human–machine interaction. Therefore, dimensional emotion recognition from speech has received extensive attention from researchers in recent years [3].

In the pursuit of continuous dimensional emotion recognition from speech, the initial stage involves the extraction of sequential acoustic features that can represent the discriminative characteristics within each short-term segment. These features may be derived directly from sequential low-level descriptors (LLDs) or from the statistical features of sequential LLDs calculated on a block of continuous frames. Temporal dynamic information plays a crucial role in dimensional emotion recognition, primarily due to the continuous nature of the target dimensional values and the short time gap between two adjacent predictions [4]. However, as it is difficult to use LLD-based and functional-based acoustic features for capturing the temporal dynamics in this task, especially for the suprasegmental information of emotional speech. As a result, valence prediction performances tend to be comparatively lower. Previous studies have shown that temporal modulation, derived from an auditory perceptual model, is capable of effectively capturing temporal dynamics for speech perception and understanding [5–7]. Several studies have explored the extraction of modulation spectral features (MSF) from temporal modulation cues by computing spectral skewness, kurtosis, and other statistical characteristics. These investigations have demonstrated the noteworthy contribution of MSF to the perception of vocal emotion. [8,9]. Cognitive neuroscience studies indicate that the auditory cortex encodes sound into spectral temporal representations of different resolutions [10]. Chen et al. [11] proposed the multi-resolution cochleagram (MRCG) feature for speech separation, which extracts cochleagrams of different resolutions to obtain spectral–temporal information at varying scales. This approach achieved the best separation performance among all evaluated features. Inspired by the MRCG feature, Peng et al. [2] proposed the multi-resolution modulation-filtered cochleagram (MMCG) feature for dimensional emotion recognition, which shows significant effects in predicting valence and arousal.

In the realm of speech emotion recognition tasks, several computational models have been widely employed, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), Transformers, and attention-based models. Among these, CNNs are often utilized to extract higher-level feature representations from speech due to their ability to maintain spectral–temporal invariance [12]. RNNs are favored for their capacity to capture long-term temporal dependencies within speech sequences and are frequently combined with CNNs to model sequence dependencies [13,14]. Long short-term memory (LSTM), a specific type of RNN, has demonstrated considerable success in time series modeling due to its memory cells that capture long-term temporal dependencies within sequential data. This has led to its widespread adoption in modeling emotional representations from speech sequences [15]. Recently, some studies have proposed parallel attention or multi-headed attention through multi-scale, multi-modal, multi-channel, and other methods to obtain salient features related to target tasks [16–18]. Zhang et al. [19] proposed a multi-parallel attention network (MPAN) model for Session-based Recommendation. This innovative model incorporates a time-aware attention mechanism to capture users' short-term interests and a refined multi-head attention mechanism to extract diverse long-term interests from distinct latent subspaces. Zhu et al. [20] designed an attention-based multi-channel LSTM architecture to predict influenza outbreaks. Xu et al. [21] integrated multi-scale region attention into CNNs to emphasize different granularities of emotional features. In the emotion recognition process, irrelevant emotional information can act as noise, affecting system performance. Zhang et al. [22] proposed an adaptive interactive attention network (AIA-Net), a model that leverages text as the primary modality and audio as the auxil-

iary modality. This model employs interactive attention weights to effectively model the dynamic interaction between both modalities.

In reference [2], the LSTM recurrent unit outputs of different MCG features were directly fused without considering their distinctiveness. Within the MMCG features, the modulation-filtered cochleagram (MCG) features across various scales bring forth different aspects of emotional expression, with each MCG feature exerting different degrees of influence on emotional states. To tackle this variability, we propose a parallel attention recurrent network (PA-net) based on modulation-filtered cochleagram to predict both valence and arousal dimensions of emotions. Initially, MMCG features are extracted from speech signals using auditory signal processing. Subsequently, the PA-net employs parallel recurrent networks that simultaneously utilize multiple recurrent units to capture the temporal and contextual dependencies of MCG features. Finally, the attention mechanism is employed to facilitate the fusion of MCG features from different scales.

The main contributions of this study are as follows:

(1) We propose a parallel attention recurrent network for dimensional emotion recognition to model multiple temporal dependencies from modulation-filtered cochleagrams at different resolutions.
(2) The results of comprehensive experiments show that the modulation-filtered cochleagram performs better than traditional acoustic-based features and other auditory-based features for valence and arousal prediction.
(3) The proposed method consistently achieves the highest value of concordance correlation coefficient for valence and arousal prediction across different signal-to-noise ratio levels, suggesting that this method is more robust to noise overall.

The remainder of this study is organized as follows. In Section 2, we briefly review the related work. In Section 3, we describe the proposed dimensional emotional recognition method through modulation-filtered cochleagram and parallel attention recurrent network. Experimental evaluations and result analysis are presented in Section 4. We conclude the study in Section 5 with future perspectives.

2. Related Work

In recent decades, there has been significant exploration of categorical models for the classification of emotions into discrete classes. While these categories effectively encompass the most prevalent emotional states, real-life emotional responses often exhibit greater complexity, comprising compound and occasionally ambiguous elements. As an alternative approach, emotions can be modeled within a dimensional framework, wherein human affect is represented as a low-dimensional vector, encompassing dimensions such as arousal, valence, liking, and dominance. This dimensional representation allows for the modeling of affective states as continuous signals over time, which in turn facilitates the development of more realistic applications. The typical approach to dimensional emotion recognition comprises two primary stages: feature extraction and regression modeling. In this section, we provide a brief overview of the techniques employed in these two stages.

2.1. Speech Feature Extraction

Acoustic-based feature. Currently, acoustic-based features employed for speech emotion recognition can be categorized into three main types: prosody features (including duration, F0, energy, zero-crossing rate, and speaking rate), sound quality features, and spectrum-based features (such as LPC, MFCC, and LPCC features). Commonly used acoustic-based features can be extracted using two strategies: one based on low-level descriptors (LLDs), which involves capturing features such as 20 ms to 40 ms frame-based acoustic, spectral, and prosodic characteristics, and another based on High-level Statistics Functions (HSFs), which computes statistical values over LLD frame sequences to yield segment-level or utterance-level statistics. LLD features exhibit poor robustness in "in-the-wild" environments, leading to a sharp decline in recognition performance. On the other hand, HSF features lack temporal information from speech and are unsuitable for

constructing regression models for dimensional emotions. Researchers predominantly focus on extracting salient features from conventional acoustic features to address diverse emotion recognition tasks. It is worth noting that while El Ayadi et al. [23] have contended that this approach using HSFs can potentially lead to the loss of temporal information and may suffer from the diminutive size of the features, Atmaja et al. [24] have demonstrated that HSFs can yield superior results compared to LLDs in the same dataset and model. However, utilizing HSF-based acoustic features to capture the temporal dynamics within this task, especially with regard to suprasegmental information in emotional speech, often results in lower valence prediction performance.

Auditory-based feature. Based on the physiological and psychological characteristics of the human auditory system, researchers designed computational auditory models to simulate the various stages of the auditory processing. These models encompass cochlear mechanics, inner hair cells (IHC), and auditory nerve and brainstem signal processing. Dau et al. [25], for instance, proposed an auditory perception model to emulate signal processing in the human auditory system. In this model, temporal modulation cues are obtained using auditory filtering of the speech signal and modulation filtering of the temporal amplitude envelope in a cascade manner. The auditory filter mimics the time-frequency signal decomposition occurring in the cochlea, the temporal amplitude envelope simulates the transduction of IHC, and the modulation filter simulates the signal modulation of the inferior colliculus (IC). As a result, this process yields temporal modulation cues with high-frequency domain resolution, encapsulating rich spectral–temporal information that enables the perception of variations in loudness, timbre, and pitch in speech. These cues contain rich spectral–temporal information to perceive variations of the loudness, timbre, and pitch of speech [6] and have been widely used in sound texture perception [26], speaker individuality perception [27], speech recognition [28,29], acoustic event recognition [30], and emotion recognition. Psychological acoustic research reveals that after the time-frequency decomposition of speech signals within the cochlea, spectral–temporal modulation occurs during transmission, resulting in the formation of a spectral–temporal modulation representation [31,32]. This type of modulation plays a crucial role in speech perception and understanding. Wu et al. [33] employed statistical functions such as spectral kurtosis and spectral skewness on the spectral–temporal modulation representation to derive MSF for speech emotion recognition. However, such statistical features lack temporal dynamics and fail to capture genuine emotional states in speech. Kshirsagar et al. [34] proposed a robust emotion recognition method that combines bag-of-audio-words and modulation spectral features to form a modulation frequency spectrum feature bag. Previous study proposed the MMCG feature to extract high-level auditory representation from temporal modulation cues for dimensional emotion recognition and designed a multi-channel parallel LSTM network architecture to track the temporal dynamics of auditory representation sequence.

2.2. Emotion Recognition Model

Convolutional and recurrent neural networks have demonstrated remarkable success in the realm of dimensional emotion prediction [35]. Trigeorgis et al. [36] proposed an innovative approach, employing one-dimensional CNNs to directly acquire high-level emotion feature representations from speech signals. Subsequently, they harnessed LSTM networks to capture the temporal dependencies within these representations, enabling the prediction of dimensional emotions. Similarly, Wöllmer et al. [37] introduced a method grounded in LSTM for the automatic recognition of audio and video cues. Interestingly, research has shown that leveraging audio information tends to yield superior results in dimensional emotion prediction compared to relying solely on video information. Furthermore, the application of attention mechanisms has become prevalent and proven to be highly effective across various tasks, including machine translation and image captioning. Yang et al. [38] proposed a CNN-BLSTM network model designed to monitor continuous changes in emotions within the arousal-valence two-dimensional space. This model achieves this by integrating inputs from both raw

waveform signals and spectrograms. To harness the temporal dynamics inherent in emotions, many studies have employed temporal attention models to capture important emotional information within speech utterances. These methods are all designed to extract various channels and spatial attention maps from LLDs, spectrograms, or waveforms, and subsequently fuse these attention maps to recognize emotions [38]. The research on the temporal attention model is mainly concentrated in categorical emotional recognition. Neumann et al. [39] introduced the attentive convolutional neural network (ACNN), which employs attention models to recognize emotions from log-Mel filterbank features. Mirsamadi et al. [40] proposed the attentive recurrent neural network (ARNN), which takes frame-level LLD inputs to the RNN and then identifies emotions using local attention as a weighted pooling method. Peng et al. [41] proposed an attention-based sliding recurrent neural network (ASRNN) to simulate the sustained attention and selective attention behavior of humans during emotion perception and recognition. Makhmudov et al. [42] developed a novel emotion recognition model that leverages attention-oriented parallel CNN encoders to concurrently capture essential features for use in emotion classification. Karnati et al. [43] proposed a texture-based feature-level ensemble parallel network (FLEPNet) to address the challenges mentioned previously and enhance the performance of a facial emotion recognition system.

However, the temporal attention model has relatively few studies in dimension emotional recognition tasks [44]. Avila et al. [45] introduced a feature pooling technique that combines MSFs and 3D spectral–temporal representations to enhance the robustness of emotion recognition. Peng et al. [2] proposed the multi-resolution modulation-filtered cochleagram (MMCG) feature for dimensional emotion recognition, which shows significant effects in predicting valence and arousal. These methods do not consider using temporal attention to capture significant emotional regions within the advanced feature sequences of speech signals. The role of different resolution features of MMCG may be different. Therefore, attention mechanisms are employed to capture salient emotional information from multi-resolution MCG features in this study.

3. Research Method

3.1. Overall Structure

The proposed dimensional emotion recognition framework, based on a modulation-filtered cochleagram and parallel attention recurrent network, is illustrated in Figure 1. The speech signal s(t) is filtered through the cochlear auditory filterbank, Hilbert transform, and modulation filterbank to generate the modulation spectrogram representation [32]. From this representation, modulation units are constructed, yielding multi-resolution modulation-filtered cochleagram features. Subsequently, the parallel attention recurrent network (utilizing LSTM as recurrent units) extracts high-level auditory modulation features from different resolution MCG inputs. The parallel recurrent network establishes multi-scale dependencies from various-resolution MCG features, and the attention mechanism facilitates feature fusion from the output feature representations of the parallel recurrent network. Finally, employing a multi-task learning approach, the emotion model is jointly trained to predict valence and arousal dimensions.

3.2. Multi-Resolution Modulation-Filtered Cochleagram

The MCG simulates the auditory processing of the human ear and encodes the 3D spectral–temporal modulation representation, yielding multi-resolution spectral–temporal features [2]. The process involves the use of Gammatone cochlear filters to mimic the cochlear basilar membrane's decomposition of the speech signal into multiple acoustic frequency channel signals. The Hilbert transform is then applied to emulate the inner hair cell's extraction of the temporal envelope for each channel. Following this, modulation filters are used to simulate the thalamus' modulation filtering of the temporal envelope, generating modulation frequency channel signals. From these modulation channels, modulation units are created. To extract multi-resolution temporal modulation cues from the

modulation units and obtain multi-scale information, each modulation unit is convolved with itself in a discrete convolution operation. Additionally, a non-linear logarithmic operation is performed on each time-frequency modulation unit to enhance the energy information of lower frequencies. In the MMCG features, the first and second modulation cochleagram (MCG1, MCG2) respectively yield cochleagram features with high and low temporal resolutions from the modulation units. By performing 2D convolution operations with rectangular windows centered on different frequency channels and time frames composed of MCG1, and subsequently applying mean pooling, the third and fourth modulation cochleagram (MCG3, MCG4) are obtained. If the window extends beyond the cochleagram's range, zero-padding is applied. The MMCG feature employs 1D or 2D convolution operations (including convolution kernels with various receptive field sizes) to create multi-resolution features. These features inherently possess strong expressive capabilities for feature representation.

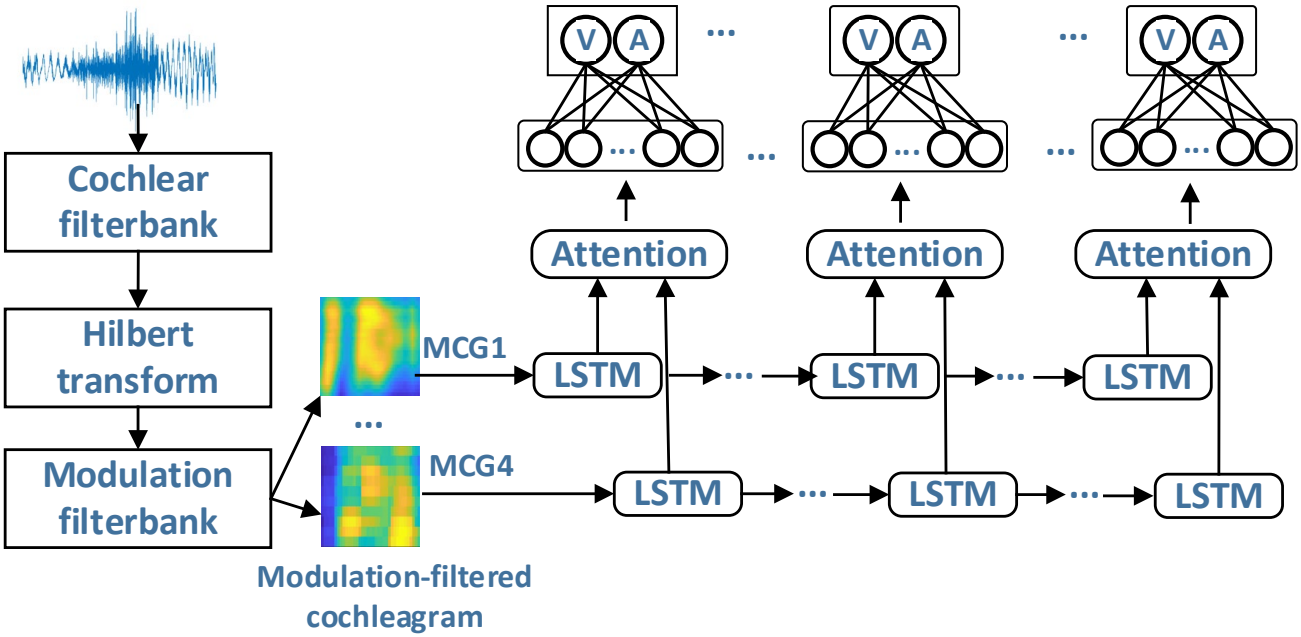

Figure 1. Dimensional emotion recognition framework based on modulation-filtered cochleagram and parallel attention recurrent network.

Figure 2 illustrates the multi-resolution modulation-filtered cochleagrams of clean speech and noisy speech. The left panel displays the modulation-filtered cochleagram features of clean speech, while the right panel shows the modulation-filtered cochleagram under a noise environment with a signal-to-noise ratio (SNR) of 5 dB. In this figure, the x-axis represents the number of modulation units, and the y-axis represents the auditory filtering channels. On the left panel, the modulation-filtered cochleagram of the first modulation channel is shown for the clean speech scenario. The MMCG is constructed by combining four modulation cochleagrams (MCG1-MCG4) with different spectral–temporal resolutions. Each modulation channel in this feature contains multi-resolution temporal information and contextual spectral–temporal information. On the right panel, the same speech is depicted in a noisy environment with an SNR of 5 dB. Despite significant distortion in the signal due to the low SNR, the salient features in the modulation-filtered cochleagram remain discernible even in the presence of noise.

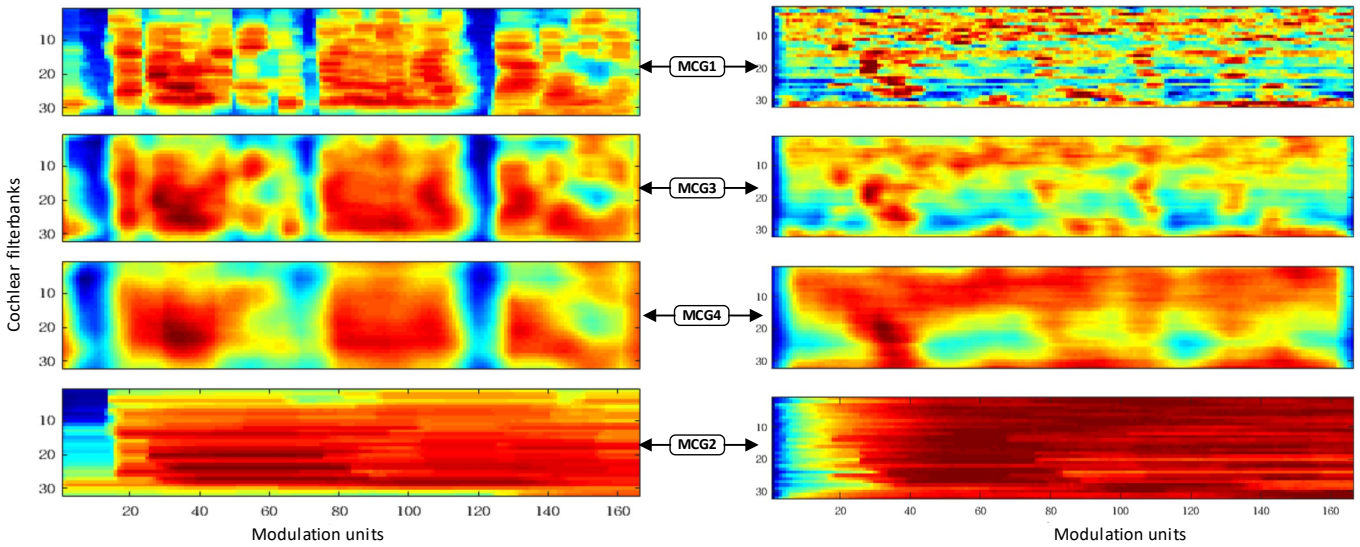

Figure 2. Multi-resolution modulation-filtered cochleagram of clean and noise speech. The left panel shows the modulation-filtered cochleagram of the first modulation channel, The right panel shows the modulation-filtered cochleagram at SNR 5 dB.

3.3. Parallel Attention Recurrent Network

The MCG1-MCG4 within MMCG encompasses temporal and contextual information at various scales. The challenge lies in amalgamating these diverse-scale MCGs cohesively. Since a single-channel recurrent network cannot simultaneously extract the interdependencies of cochleagram features at different scales, this study introduces a parallel attention recurrent network (PA-net), as depicted in Figure 3. In this approach, parallel recurrent networks are utilized, employing multiple recurrent units concurrently to capture the temporal and contextual dependencies within cochleagram features. This is facilitated by incorporating an attention mechanism, enabling the fusion of MCG features across different scales. $MCG_k(n, m, i)$ refers to the n th acoustic frequency channel of the i th modulation unit and the m th modulation frequency channel in the k th modulation-filtered cochleagram. The k th modulation-filtered cochleagram is $MCG_k(n, m, i)$, indicated as follows:

$$MCG_k(n, m, i) \in R^{N \times M \times I}, \tag{1}$$

where N, M, and I represent the number of cochlear filter channels, the number of modulation channels, and the temporal modulation units, respectively. Subsequently, different scales $MCG_k(n, m, i)$ are sent to the loop network to generate S_k, and then ReLU is used to generate the nonlinear transformation $\mathcal{R}(S_k)$.

$$\mathcal{R}(S_k) = U_k ReLU(W_k S_k + b_k), \tag{2}$$

where, W_k, U_k are the trainable parameter matrix and b_k are biased. Using the ReLU nonlinear function, which has good convergence performance. For each S_k, the α_k is calculated as follows:

$$\alpha_k = \frac{exp(\mathcal{R}(S_k))}{\sum_{k=1}^{4} exp(\mathcal{R}(S_k))}. \tag{3}$$

The weight of the recurrent unit output S_k is obtained through the attention module, and the weighted fusion features are obtained by multiplying with the S_k, which are expressed as follows:

$$att\_sum = \sum_{k=1}^{4} \alpha_k S_k. \tag{4}$$

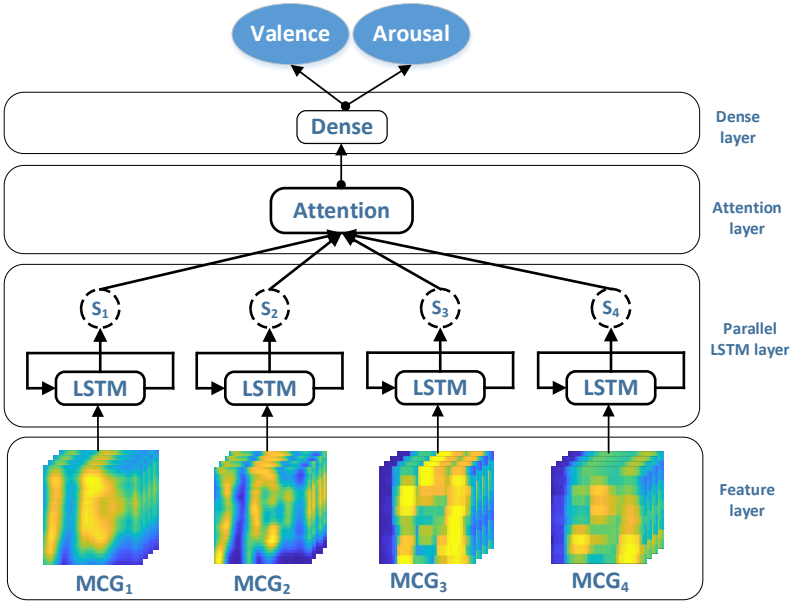

Figure 3. Parallel attention recurrent neural network.

Finally, predictive models of valence and arousal are constructed by fully connected Dense layers.

4. Experimental Results and Analysis

4.1. The Emotional Speech Data

In this study, experiments for dimensional emotion recognition were conducted using subsets of the RECOLA (remote collaborative and affective interactions) [46] and SEWA (sentiment analysis in the wild) [47] datasets. Both datasets consist of spontaneous emotional dialogue data and their subsets were used for the 2016 and 2017 AVEC Emotion Challenge [48,49]. The RECOLA dataset represents a multi-modal corpus, capturing remote collaborative and affective interactions. This comprehensive dataset comprises 27 French-speaking individuals and is thoughtfully partitioned into three subsets, each containing nine participants: a training set, a development set, and a testing set. These partitions are designed to ensure a balanced representation of various demographic characteristics, including gender, age, and primary language spoken by the participants. The SEWA dataset is a collection of mixed audiovisual content, featuring interactions between 64 target speakers and their conversational partners. This dataset is systematically divided into three distinct subsets: 34 in the training set, 14 in the development set, and 16 in the testing set. The emotion dimensions, including arousal, valence, and liking, were continuously annotated for these recorded segments. The primary distinction between RECOLA and SEWA lies in the annotation frequency, where in RECOLA, each valence and arousal value is annotated every 40 milliseconds frame, and in SEWA, annotations are performed every 100 milliseconds frame. In this study, predictions for valence and arousal were made on these two data subsets. The proposed dimensional speech emotion recognition model was trained and validated on the same training and development sets as in references [2,45,50].

4.2. Multitask Learning and Evaluation Metrics

The experiment used the evaluation index CCC (concordance correlation coefficient) officially recommended by the AVEC Challenge. ρ_c is the concordance correlation coefficient between the prediction values of emotion dimensions and the gold-standard measurement, and the calculation formula is as follows:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{5}$$

where ρ is the Pearson correlation coefficient between the time series prediction and the gold standard, x is the predictive value of a certain emotional dimension, y is the gold standard corresponding to x, σ_x^2 and σ_y^2 are the variance of two sequences, and μ_x and μ_y are the mean of two sequences. In the valence–arousal emotion space, due to the strong correlation between valence and arousal [2], a multi-task learning method is used to predict both valence and arousal simultaneously in this study, and use CCC-based loss function (L_c) as the objective function of the depth model. L_c be defined as:

$$L_c = \frac{2 - \rho_c^a - \rho_c^v}{2}, \tag{6}$$

where ρ_c^a and ρ_c^v are the CCC for valence and arousal, respectively.

4.3. Experimental Results

4.3.1. Benchmark Experiments

For the RECOLA dataset, a comparative experiment was conducted involving the extraction of MFCC, the extended version of Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [51], Modulation Spectral Feature (MSF), Modulation-filtered Cochleagram (MRCG), and Multi-resolution Modulation-filtered Cochleagram (MMCG) features, as well as LLD and HSF strategies. Firstly, the speech signals underwent pre-emphasis and normalization operations. Subsequently, the processed data were segmented into multiple sub-sequences, which were then used as inputs for the model. In the context of dimensional emotion recognition, where valence and arousal values are annotated continuously over a sequence of frames, the LLD-based strategy employed frame stacking to extract frame-level feature sequences. Specifically, a four-frame stacking approach was used to obtain frame-level features in the RECOLA dataset. On the other hand, the HSF-based strategy involved applying statistical functions to compute 4 s long LLD features, which were then offset by 40 milliseconds to generate frame-level feature sequences. To model these feature sequences, Support Vector Regression (SVR) and a single-channel Long Short-Term Memory (LSTM) were employed as baseline models. These baseline models were used for comparison purposes in the experimentation.

The baseline LSTM network consists of an input layer, two hidden layers with 128 and 64 nodes, respectively, followed by a fully connected layer and a regression layer. The hidden layers are connected using a fully connected layer with ReLU activation for non-linearity. During model training, a dropout rate of 0.75 is applied before the regression layer to prevent overfitting. Finally, the regression layer is used to predict the valence and arousal values of emotions. Table 1 presents the prediction results of the two regression models on different features using the RECOLA dataset. It is evident from the table that MMCG features achieved the highest arousal prediction result (CCC of 0.742) using the LSTM-based regression method, while they also yielded the highest valence prediction result (CCC of 0.371) using the SVR-based regression method. Within the same regression model, auditory perception-based features (MSF, MRCG, and MMCG) outperformed the acoustic features based on LLD and HSF in dimensional emotion recognition. This observation highlights that auditory features extracted from the perspective of speech perception exhibit stronger feature expression and better predictive power for valence and arousal emotion dimensions compared to acoustic features extracted from the perspective of speech generation.

Table 1. The CCC using different feature sets (RECOLA).

Feature	SVR		LSTM	
	Arousal	**Valence**	**Arousal**	**Valence**
MFCC_LLD	0.595	0.269	0.679	0.320
MFCC_HSF	0.606	0.305	0.651	0.331

Table 1. *Cont.*

Feature	SVR		LSTM	
	Arousal	Valence	Arousal	Valence
eGeMAPS_LLD	0.610	0.293	0.662	0.312
eGeMAPS_HSF	0.602	0.314	0.701	0.329
MSF	0.641	0.304	0.709	0.368
MRCG	0.685	0.353	0.734	0.351
MMCG	0.694	0.371	0.742	0.362

4.3.2. Noise Environment Valence and Arousal Prediction

To further analyze the impact of noise environments on the dimensional emotion recognition of different features, this study employed the same LSTM network to investigate the performance of valence and arousal prediction with the addition of Gaussian white noise at various SNR levels in the RECOLA dataset. Table 2 displays the valence and arousal prediction for different features with long-time and delta feature conditions at various SNR levels. The results indicate that the predictive ability of acoustic features in noisy environments is significantly lower compared to that of auditory modulation-based features. For instance, in a 20 dB SNR environment, the arousal CCC based on MFCC features is only 0.426, whereas it increases to 0.772 when using MMCG features. Similarly, valence CCC improves from 0.193 to 0.418. This indicates that the valence and arousal predictive abilities of acoustic features are more susceptible to noise interference compared to auditory features. Comparing the prediction of valence and arousal in noisy environments to those in clean speech environments, there is a noticeable decrease in prediction performance. Moreover, auditory perception-based features demonstrate a significant advantage in noise robustness compared to acoustic features. MMCG consistently achieves the highest valence and arousal CCC values across different SNR levels, suggesting that MMCG features are more robust to noise overall. This advantage might stem from auditory modulation filtering, which further decomposes the noisy signal, allowing extraction of low-frequency information that remains relatively unaffected by noise interference.

Table 2. The CCC using different feature sets under different SNR (RECOLA).

Feature	0 dB		5 dB		10 dB		20 dB	
	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence
MFCC_HSF	0.361	0.173	0.367	0.182	0.403	0.195	0.426	0.193
eGeMAPS_HSF	0.403	0.202	0.421	0.196	0.446	0.201	0.451	0.203
MSF	0.594	0.252	0.548	0.266	0.478	0.318	0.574	0.226
MRCG	0.658	0.304	0.674	0.318	0.696	0.316	0.718	0.364
MMCG	0.700	0.344	0.744	0.400	0.750	0.446	0.772	0.418

4.3.3. Valence and Arousal Prediction Based on PA-Net

The parallel attention recurrent network, PA-net, captures the significant emotional modulation features in the speech spectral–temporal modulation space from different resolution MCG features and models their feature dependencies. Table 3 presents the valence and arousal prediction results for single-channel LSTM, multi-channel LSTM, and PA-net on the RECOLA and SEWA datasets. In the RECOLA experiments, training sequences with a length of approximately 30 s were used, and testing did not require segmentation. The highest CCC was achieved from PA-net on RECOLA, whose arousal and valence were 0.859 and 0.529, respectively. Moreover, compared with the single-channel LSTM, the arousal prediction was relatively improved by 15.7% (from 0.742 to 0.859), and the valence prediction by 46.1% (from 0.362 to 0.529). In the SEWA experiments, due to variable sequence lengths in the dataset, zero-padding was applied to align all sequences before training the deep regression model. Sequence lengths were around 90 s, and testing

did not involve segmentation or padding operations [49]. The highest CCC was achieved from PA-net on SEWA, whose arousal and valence were 0.557 and 0.531, respectively, which is consistent with the results obtained on RECOLA. The experimental results indicate that PA-net outperforms single-channel and multi-channel LSTMs in both datasets for valence and arousal prediction. This suggests that the attention-based parallel recurrent network is better at modeling the dependency relationships of different scale MCG features, leading to improved prediction performance.

Table 3. The CCC of different recurrent networks under RECOLA and SEWA datasets.

Dataset	Single-Channel LSTM		Multi-Channel LSTM		PA-Net	
	Arousal	Valence	Arousal	Valence	Arousal	Valence
RECOLA	0.742	0.362	0.824	0.474	0.859	0.529
SEWA	0.472	0.342	0.523	0.519	0.557	0.531

In order to further analyze the dimension emotion recognition performance of PA-net in noisy environments, this study compared the valence and arousal prediction results of PA-net and LSTM networks at different signal-to-noise ratios (SNR) on the RECOLA dataset. Table 4 presents the valence and arousal prediction CCC scores for PA-net and LSTM networks under various SNR conditions. It can be seen that the prediction of valence and arousal emotion is severely affected by the presence of noise. However, the experimental findings indicate that PA-net outperforms the single-channel LSTM network in valence and arousal prediction with higher CCC under varying SNR. This suggests that PA-net exhibits superior noise robustness in predicting valence and arousal compared to the single-channel LSTM network.

Table 4. The CCC using different deep models under different SNR.

Model	0 dB		5 dB		10 dB		20 dB	
	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence
LSTM	0.700	0.344	0.744	0.400	0.750	0.446	0.772	0.418
PA-net	0.743	0.372	0.779	0.427	0.795	0.463	0.817	0.476

Figure 4 illustrates the valence and arousal prediction examples of the single-channel LSTM and PA-net models based on MMCG features. The green curves represent the prediction sequences of arousal (Figure 4a) and valence (Figure 4b) from the single-channel LSTM network in continuous speech signals. The orange curves depict the prediction sequences of valence and arousal from the PA-net model in continuous speech signals. The deep blue curves represent the corresponding ground truth values. From the figure, it can be seen that for the prediction of arousal and valence, the PA-net obtains CCCs of 0.93 and 0.63, respectively, while the LSTM network obtains CCCs of 0.88 and 0.59, respectively. This suggests that there is more significant variability in the valence and arousal prediction values when modeling MCG features at different resolutions with LSTM. In contrast, PA-net demonstrates a better capability to closely match the ground truth. This indicates that the PA-net fits the ground truth curves better than the LSTM network.

Finally, this study compares the CCC scores obtained by different methods on the RECOLA dataset, as shown in Table 5. The proposed emotion recognition approach based on MCG features and PA-net achieves the best performance in both valence and arousal predictions. Specifically, PA-net outperforms the multi-channel approach by an improvement of 5.8% in arousal prediction and 10% in valence prediction. This suggests that incorporating an attention mechanism for advanced feature fusion yields better results than a simple concatenated approach for feature fusion.

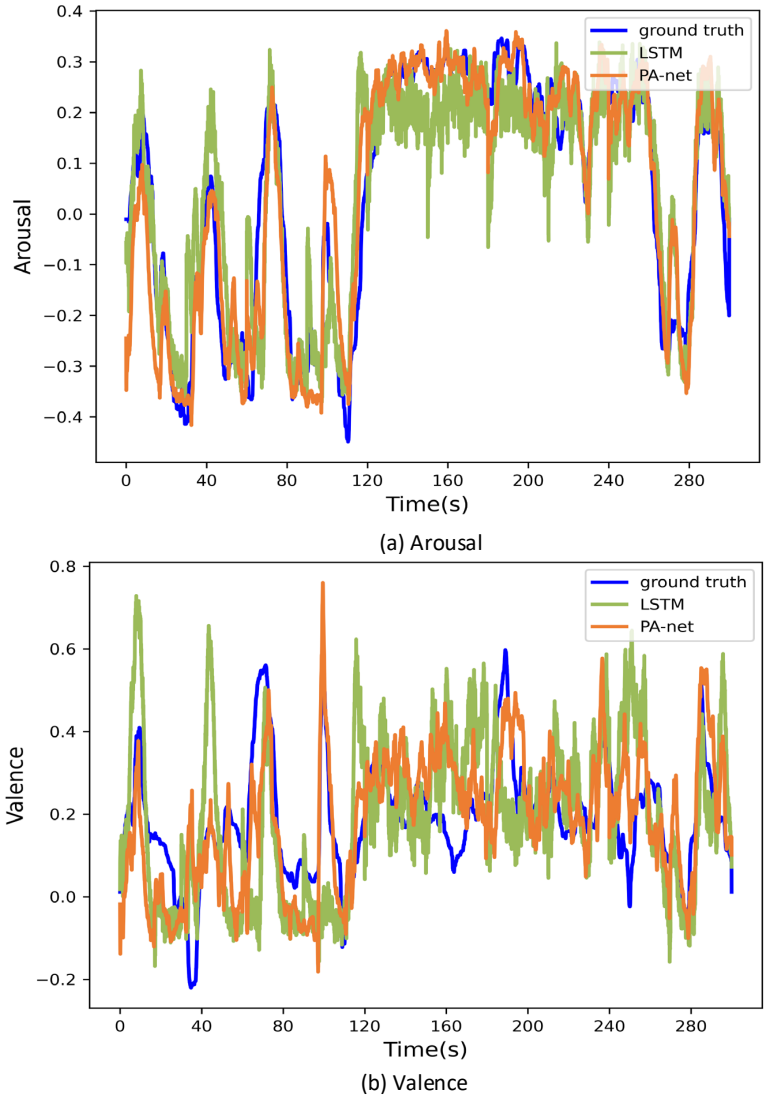

Figure 4. An example of (**a**) arousal and (**b**) valence prediction of the MMCG features on LSTM and PA-net obtained for subject P26 in RECOLA.

Table 5. The CCC comparison under different features and models on the RECOLA dataset.

Method	Feature	Model	Arousal	Valence
Zhang et al. [52]	eGeMAPS	RNN	0.783	0.495
Avila et al. [45]	MSFs	Single-channel LSTM	0.795	0.265
Peng et al. [2]	MMCG	Multi-channel LSTM	0.812	0.481
Proposed method	MMCG	PA-net	0.859	0.529

5. Conclusions

Speech emotion recognition plays a crucial role in enabling natural human–robot interaction. In this study, we propose a dimension emotion recognition method based on multi-resolution modulation cochleargram (MMCG) and parallel attention recurrent network (PA-net). The PA-net is utilized to capture temporal and contextual information at different scales from MMCG features and establish multiple temporal dependencies to track the dynamic changes of dimensional emotions in auditory representation sequences. Our experimental findings consistently demonstrate the superiority of our proposed method, as it consistently achieves the highest concordance correlation coefficient values for valence and arousal prediction across a range of signal-to-noise ratio levels. At the feature level, MMCG surpasses other assessed features in its ability to predict valence and arousal, with

remarkable efficacy particularly in high signal-to-noise ratio scenarios. Furthermore, at the model level, the PA-net exhibits the highest predictive performance for both valence and arousal, significantly outperforming alternative regression models.

In summary, our results collectively underscore the potency and effectiveness of our approach in advancing the field of dimensional emotion recognition. In the future, we plan to conduct further research on modulation cochleargram features based on human auditory characteristics, and then plan to use some pre-trained models to obtain salient emotional information from MMCG features.

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

Funding: 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).

Data Availability Statement: The Remote collaborative and affective interactions dataset (RECOLA) used in this paper is available through the following link: (https://diuf.unifr.ch/main/diva/recola/, accessed on 1 January 2019). The Sentiment analysis in the wild (SEWA) used in this paper is available through the following link: (https://db.sewaproject.eu/, accessed on 1 March 2019).

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
BLSTM	Bidirectional LSTM
CCC	Concordance correlation coefficient
CNN	Convolutional neural network
CRNN	Convolutional and recurrent neural network
eGeMAPS	Extended version of Geneva Minimalistic Acoustic Parameter Set
HSF	High-level statistics function
HRI	Human–robot interaction
IC	Inferior colliculus
IHC	Inner hair cells
LLD	Low-level descriptors
LSTM	Long short-term memory
MFCC	Mel frequency cepstral coefficient
MCG	Modulation-filtered cochleagram
MMCG	Multi-resolution modulation-filtered cochleagram
MRCG	Multi-resolution cochleagram
MSF	Modulation spectral feature
RNN	Recurrent neural network
PA-net	Parallel attention recurrent network
RECOLA	Remote collaborative and affective interactions
SEWA	Sentiment analysis in the wild
SNR	Signal-to-noise ratio

References

1. Li, H.F.; Chen, J.; Ma, L.; Bo, H.J.; Xu, C.; Li, H.W. Dimensional speech emotion recognition review. *Ruan Jian Xue Bao J. Softw.* **2020**, *31*, 2465–2491. [CrossRef]
2. 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]
3. Mencattini, A.; Martinelli, E.; Ringeval, F.; Schuller, B.; Di Natale, C. Continuous Estimation of Emotions in Speech by Dynamic Cooperative Speaker Models. *IEEE Trans. Affect. Comput.* **2017**, *8*, 314–327. [CrossRef]

4. Chen, S. Multi-modal Dimensional Emotion Recognition using Recurrent Neural Networks. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, Brisbane, Australia, 26–30 October 2015; pp. 49–56.
5. Drullman, R. Temporal envelope and fine structure cues for speech intelligibility. *J. Acoust. Soc. Am.* **1995**, *97*, 585–592. [CrossRef]
6. Atlas, L.; Shamma, S.A. Joint Acoustic and Modulation Frequency. *EURASIP J. Appl. Signal Process.* **2003**, 668–675. [CrossRef]
7. Unoki, M.; Zhu, Z. Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech. *Acoust. Sci. Technol.* **2020**, *41*, 233–244. [CrossRef]
8. Zhu, Z.; Miyauchi, R.; Araki, Y.; Unoki, M. Contribution of modulation spectral features on the perception of vocal-emotion using noise-vocoded speech. *Acoust. Sci. Technol.* **2018**, *39*, 379–386. [CrossRef]
9. Peng, Z.; Zhu, Z.; Unoki, M.; Dang, J.; Akagi, M. Dimensional Emotion Recognition from Speech Using Modulation Spectral Features and Recurrent Neural Networks. In Proceedings of the 11th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 524–528.
10. Chi, T.; Ru, P.; Shamma, S.A. Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* **2005**, *118*, 887–906. [CrossRef]
11. 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]
12. 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]
13. 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]
14. Satt, A.; Rozenberg, S.; Hoory, R. Efficient emotion recognition from speech using deep learning on spectrograms. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20-24 August 2017; pp. 1089–1093.
15. Tzirakis, P.; Zhang, J.; Schuller, B.W. End-to-end speech emotion recognition using deep neural networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5089–5093.
16. Fan, J.; Zhang, K.; Huang, Y.; Zhu, Y.; Chen, B. Parallel spatio-temporal attention-based TCN for multivariate time series prediction. *Neural Comput. Appl.* **2023**, *35*, 13109–13118. [CrossRef]
17. Chen, S.; Jin, Q.; Zhao, J.; Wang, S. Multimodal multi-task learning for dimensional and continuous emotion recognition. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, CA, USA, 23–27 October 2017; pp. 19–26.
18. Huang, J.; Li, Y.; Tao, J.; Lian, Z.; Wen, Z.; Yang, M.; Yi, J. Continuous multimodal emotion prediction based on long short term memory recurrent neural network. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, CA, USA, 23–27 October 2017; pp. 11–18.
19. Zang, T.; Zhu, Y.; Zhu, J.; Xu, Y.; Liu, H. MPAN: Multi-parallel attention network for session-based recommendation. *Neurocomputing* **2022**, *471*, 230–241. [CrossRef]
20. Fu, B.; Yang, Y.; Ma, Y.; Hao, J.; Chen, S.; Liu, S.; Li, T.; Liao, Z.; Zhu, X. Attention-Based Recurrent Multi-Channel Neural Network for Influenza Epidemic Prediction. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 3–6 December 2018; pp. 1245–1248. [CrossRef]
21. Xu, M.; Zhang, F.; Cui, X.; Zhang, W. Speech Emotion Recognition with Multiscale Area Attention and Data Augmentation. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6319–6323. [CrossRef]
22. Zhang, T.; Li, S.; Chen, B.; Yuan, H.; Chen, C.L.P. AIA-Net: Adaptive Interactive Attention Network for Text–Audio Emotion Recognition. *IEEE Trans. Cybern.* **2022**, 1–13. [CrossRef] [PubMed]
23. 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]
24. Atmaja, B.T.; Akagi, M. Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM. *Speech Commun.* **2020**, *126*, 9–21. [CrossRef]
25. Dau, T.; Kollmeier, B.; Kohlrausch, A. Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *J. Acoust. Soc. Am.* **1997**, *102*, 2906–2919. [CrossRef]
26. 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]
27. Zhu, Z.; Miyauchi, R.; Araki, Y.; Unoki, M. Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech. *Acoust. Sci. Technol.* **2018**, *39*, 234–242. [CrossRef]
28. Moritz, N.; Anemuller, J.; Kollmeier, B. An Auditory Inspired Amplitude Modulation Filter Bank for Robust Feature Extraction in Automatic Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 1926–1937. [CrossRef]
29. Yin, H.; Hohmann, V.; Nadeu, C. Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency. *Speech Commun.* **2011**, *53*, 707–715. [CrossRef]
30. Sharan, R.V.; Moir, T.J. Acoustic event recognition using cochleagram image and convolutional neural networks. *Appl. Acoust.* **2019**, *148*, 62–66. [CrossRef]

31. 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] [PubMed]
32. 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]
33. Wu, S.; Falk, T.H.; Chan, W.-Y. Automatic speech emotion recognition using modulation spectral features. *Speech Commun.* **2011**, *53*, 768–785. [CrossRef]
34. Kshirsagar, S.R.; Falk, T.H. Quality-Aware Bag of Modulation Spectrum Features for Robust Speech Emotion Recognition. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1892–1905. [CrossRef]
35. Zhang, Z.; Ringeval, F.; Han, J.; Deng, J.; Marchi, E.; Schuller, B. Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks. In Proceedings of the Annual Conference International Speech Communication Association, Interspeech, San Francisco, CA, USA, 8–12-September 2016; pp. 3593–3597. [CrossRef]
36. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20–25 March 2016; pp. 5200–5204. [CrossRef]
37. Wöllmer, M.; Kaiser, M.; Eyben, F.; Schuller, B.; Rigoll, G. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image Vis. Comput.* **2013**, *31*, 153–163. [CrossRef]
38. Yang, Z.; Hirschberg, J. Predicting Arousal and Valence from Waveforms and Spectrograms using Deep Neural Networks. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 3092–3096.
39. 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 Interspeech 2017 18th Conference International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 1263–1267. [CrossRef]
40. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.
41. 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]
42. Makhmudov, F.; Kutlimuratov, A.; Akhmedov, F.; Abdallah, M.S.; Cho, Y.-I. Modeling Speech Emotion Recognition via Attention-Oriented Parallel CNN Encoders. *Electronics* **2022**, *11*, 4047. [CrossRef]
43. Karnati, M.; Seal, A.; Yazidi, A.; Krejcar, O. FLEPNet: Feature Level Ensemble Parallel Network for Facial Expression Recognition. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2058–2070. [CrossRef]
44. Wagner, J.; Triantafyllopoulos, A.; Wierstorf, H.; Schmitt, M.; Burkhardt, F.; Eyben, F.; Schuller, B.W. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10745–10759. [CrossRef]
45. Avila, A.R.; Akhtar, Z.; 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**, *12*, 177–188. [CrossRef]
46. Ringeval, F.; Sonderegger, A.; Sauer, J.; Lalanne, D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013. [CrossRef]
47. Kossaifi, J.; Walecki, R.; Panagakis, Y.; Shen, J.; Schmitt, M.; Ringeval, F.; Han, J.; Pandit, V.; Toisoul, A.; Schuller, B.W.; et al. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *arXiv* **2019**, arXiv:1901.02839. [CrossRef] [PubMed]
48. Valstar, M.; Schuller, B.; Smith, K.; Almaev, T.; Eyben, F.; Krajewski, J.; Cowie, R.; Pantic, M. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In Proceedings of the AVEC 2016—Depression, Mood, and Emotion Recognition Workshop and Challenge, Amsterdam, Netherlands, 15–19 October 2016; pp. 3–10. [CrossRef]
49. Ringeval, F.; Schuller, B.; Valstar, M.; Gratch, J.; Cowie, R.; Scherer, S.; Mozgai, S.; Cummins, N.; Schmitt, M.; Pantic, P. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, CA, USA, 23–27 October 2017; pp. 3–9.
50. Ouyang, A.; Dang, T.; Sethu, V.; Ambikairajah, E. Speech based emotion prediction: Can a linear model work? In Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, Graz, Austria, 15–19 September 2019; pp. 2813–2817. [CrossRef]
51. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; Andre, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2015**, *7*, 190–202. [CrossRef]
52. Zhang, Z.; Han, J.; Coutinho, E.; Schuller, B.W. Dynamic Difficulty Awareness Training for Continuous Emotion Prediction. *IEEE Trans. Multimed.* **2019**, *21*, 1289–1301. [CrossRef]