DBPNet: Dual-Branch Parallel Network with Temporal-Frequency Fusion for Auditory Attention Detection

Qinke Ni, **Hongyu Zhang**, **Cunhang Fan**, **Shengbing Pei**, **Chang Zhou** and **Zhao Lv**

Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China

{e02114031, e22201103}@stu.ahu.edu.cn, {cunhang.fan, kjlz}@ahu.edu.cn

Abstract

Auditory attention decoding (AAD) aims to recognize the attended speaker based on electroencephalography (EEG) signals in multi-talker environments. Most AAD methods only focus on the temporal or frequency domain, but neglect the relationships between these two domains, which results in the inability to simultaneously consider both time-varying and spectral-spatial information. To address this issue, this paper proposes a dual-branch parallel network with temporal-frequency fusion for AAD, named DBPNet, which consists of the temporal attentive branch and the frequency residual branch. Specifically, the temporal attentive branch aims to capture the time-varying features in the EEG time-series signal. The frequency residual branch aims to extract spectral-spatial features of multi-band EEG signals by the residual convolution. Finally, these dual branches are fused to consider both EEG signals time-varying and spectral-spatial features and get classification results. Experimental results show that compared with the best baseline, DBPNet achieves a relative improvement of 20.4% with a 0.1-second decision window for the MM-AAD dataset, but the number of trainable parameters is reduced by about 91 times.

1 Introduction

In the context of a cocktail party, people possess the ability to isolate and concentrate on a single sound source within a noisy, multi-talker environment, a skill commonly referred to as the cocktail party effect [Haykin and Chen, 2005]. However, individuals with hearing impairments may reduce or completely lose the capacity to focus on the specific speaker amidst background noise. Most existing hearing aids fall short in addressing the fundamental challenge of pinpointing the target speaker [Puffay *et al.*, 2023]. According to previous neuroscience studies, there is a connection between brain activity and auditory attention [Mesgarani and Chang, 2012]. Decoding auditory attention from brain neural activities is the auditory attention decoding (AAD) task. Therefore, handling the cocktail party problem is one of the applications of AAD practical research [Ciccarelli *et al.*, 2019].

Electroencephalography (EEG) [O'sullivan *et al.*, 2015], as the primary technology for recording brain activity, several studies have confirmed the feasibility of EEG used in AAD [Choi *et al.*, 2013].EEG records the brain's electrical activity via scalp-placed sensors, producing a sequence of non-linear time series signals. Through Fast Fourier Transform (FFT) [Stone, 1966], the signals are converted from the temporal domain to the frequency domain. Based on the frequency ranges, these signals are typically divided into several main types: delta (1–3 Hz), theta (4–7 Hz), alpha (8–13 Hz), beta (14–30 Hz) and gamma (31–50 Hz) [Aftanas *et al.*, 2004; Liu *et al.*, 2017]. Different frequency bands have different spatial features, which depict different states of the human body. Numerous methods have been proposed for extracting EEG features, like differential entropy (DE) [Shi *et al.*, 2013] and power spectral density (PSD) [Frantzidis *et al.*, 2010].

AAD research primarily focuses on two paradigms: speaker identification and tracking spatial attention [Puffay *et al.*, 2023]. Recent advancements in neuroscience have uncovered connections between neural activities and auditory detection [Ding and Simon, 2012]. According to conclusions, some researches employ a stimulus-reconstruction or speech envelope reconstruction method, which necessitates the use of clean auditory stimuli as input [O'sullivan *et al.*, 2015; Cai *et al.*, 2021a]. However, most real-world scenarios are multi-sound environments. Listeners receive a mixture of multiple sounds, making it a challenge to obtain clean auditory stimuli. Therefore, our studies focus on using only EEG signals as input to achieve tracking spatial attention.

Traditional AAD tasks rely on linear methods to process EEG signals [Geirnaert *et al.*, 2021]. However, these linear methods often struggle with non-linear mapping relationships, leading to the need for longer decision windows [Miran *et al.*, 2018]. Previous studies have proven each frequency band contains different spatial features, which depict different states of the human body [Zheng and Lu, 2015]. Therefore, some studies focus on the frequency of EEG signals. They employ the convolutional neural network (CNN) to decode from frequency bands and have good performance, extracting DE values from frequency bands and projecting them onto 2D topological maps [Jiang *et al.*, 2022]. However, it overlooks the dynamic time-varying features of the EEG signals. Given that EEG signals are essentially time-series data [Bassett and Sporns, 2017], other studies focus on the temporal as-

pects of EEG signals [Su *et al.*, 2022], introducing the attention mechanism, which obtains successful results. Although EEG time-series signals have abundant time-varying information, their limited spectral-spatial features in the temporal domain indicate a weak correlation with human spatial attention [Wöstmann *et al.*, 2016]. Therefore, it's necessary to fuse the temporal and frequency domain to get a more comprehensive feature and direct concatenation with two features is an effective method. However, there is currently no work exploring the fusion of temporal and frequency domain features in AAD. Without any one of the temporal and frequency domain information, EEG features cannot be comprehensively represented, which means its time-varying or spectral-spatial information can not be considered simultaneously.

To address this issue, this paper proposes a novel dual-branch parallel network with temporal and frequency fusion (DBPNet) for AAD, which comprehensively considers both temporal and frequency domains and employs feature fusion to exploit time-varying and spectral-spatial features simultaneously. Specifically, our network is achieved by two branches and one module: (1) *Temporal Attentive Branch*. The temporal attentive branch focuses on the temporal correlations within EEG signals different from concentration on the channels correlations in previous studies. It can capture the dynamics of time-varying information in EEG time-series signals as temporal features. (2) *Frequency Residual Branch*. The frequency residual branch employs residual convolution rather than repeated stacking of convolutional layers to extract spectral-spatial features as frequency features from multi-band EEG signals (3) *Feature Fusion & Classifier*. These dual branches are fused to simultaneously consider the temporal and frequency domain features and input them into a classifier, which predicts the direction class label by mapping the fused features into label space. Furthermore, we assess the decoding performance of the DBPNet in three datasets: KUL, DTU and MM-AAD. Especially, in MM-AAD, the proposed DBPNet with smaller trainable parameters achieves 20.4% relative improvements over the best baseline in terms of 0.1-second decision windows. These experimental results indicate that DBPNet can effectively extract and fuse the temporal and frequency domain features of EEG signals. The major contributions of our paper are outlined as follows:

- We propose a novel dual-branch network for AAD, which consists of a temporal attentive branch and a frequency residual branch. Therefore, the proposed method can make full use of the time-varying and spectral-spatial information simultaneously.
- The DBPNet, compared to the best baseline, demonstrates a relative improvement of approximately 20.4% for 0.1-second decision window in the MM-AAD dataset. Additionally, it achieves a significant reduction in trainable parameters, approximately 91 times fewer than the best baseline.

2 Our Proposed DBPNet Method

Previous researches on EEG-based AAD only concentrate on either the temporal domain or the frequency domain, neglecting the advantages of the fusion of two domain features. To address this issue, we propose DBPNet, as shown in Figure 1, a novel EEG-based model with temporal-frequency domain fusion. It includes a temporal attentive branch (TABNet) that learns time-varying features from EEG signals [Vaswani *et al.*, 2017], a frequency residual branch (FRBNet) that extracts spectral-spatial features from the frequency domain, along with a feature fusion and classifier module that synthesizes these two features and get final classification results.

Employing a conventional method to split EEG data into moving windows, we can get a series of decision windows. Each decision window contains a small duration of EEG signals, collectively represented by $R = [r_1, \ldots, r_i, \ldots, r_T] \in \mathbb{R}^{N \times T}$, forming a $N \times T$ matrix. Here, N denotes the number of EEG channels and T represents the length of the decision window, so $r_i \in \mathbb{R}^{N \times 1}$ means EEG data with N channels at the i-th decision window.

2.1 Temporal Attentive Branch

Previous studies have shown different individuals have variations in response to the same stimuli [Deng *et al.*, 2020; Bednar *et al.*, 2017; Golumbic *et al.*, 2013], which may be caused by their distinct physiology and psychology [Viswanathan *et al.*, 2019; Choi *et al.*, 2014]. During listening to the auditory stimuli, an individual's attention varies in responding to the content of auditory attention [Jones and Boltz, 1989; Jones *et al.*, 2002]. The transformer encoder can dynamically allocate attention weights to capture the time-varying information from the EEG signals. On this basis, we implement a three-step process to extract temporal features.

Previous studies have shown that the common spatial pattern (CSP) algorithm [Geirnaert *et al.*, 2020] performs excellently in addressing brain-computer interface tasks. So firstly, before processing EEG signals in DBPNet, we apply the CSP algorithm to enhance the signal-to-noise ratio of raw signals [Ramoser *et al.*, 2000; Lotte and Guan, 2010; Cheng *et al.*, 2020],

$$E = CSP(R) \tag{1}$$

where $CSP(\cdot)$ is representation of the CSP algorithm, $E \in \mathbb{R}^{N \times T}$ representing processed EEG signal.

Secondly, we employ the transformer encoder, as shown in Figure 1. Considering the change of temporal signal in EEG data, the transformer encoder [Vaswani *et al.*, 2017] can dynamically allocate weights and encode the raw EEG signal E. It can be formulated as follows:

$$S = TransformerEncoder(E) \tag{2}$$

where $TransformerEncoder(\cdot)$ denotes transformer encoder algorithm, S is encoded EEG data through transformer encoder and $S \in \mathbb{R}^{N \times T}$ has the same shape as E.

Finally, we aggregate EEG signals S through a 1D convolution layer along with an adaptive average pooling layer to get average feature values,

$$P = AdaptiveAvgPool(relu(Conv(S))) \tag{3}$$

where $Conv(\cdot)$ represents the convolutional layer along with the rectified linear unit function $relu(\cdot)$.

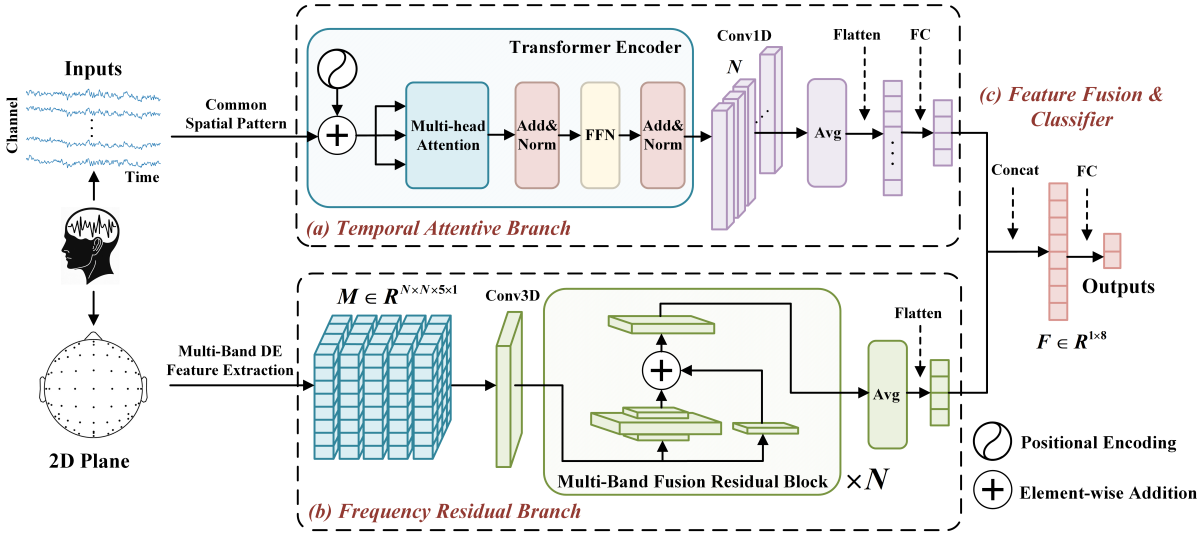


Figure 1: An overview of the DBPNet, containing a dual-pathway temporal-frequency fusion approach and a classifier to get classification results. (a) is the temporal attentive branch, consisting of a transformer encoder layer, a 1D convolutional layer, an adaptive average pooling layer and fc layers, dedicated to the extraction of temporal domain features. (b) is the frequency residual branch, consisting of a 3D convolutional layer, a max pooling layer, multi-band fusion residual blocks and an adaptive average pooling layer, focusing on the extraction of frequency domain features. (c) is the feature fusion and classifier, serving to get final detection results through direct concatenating.

$AdaptiveAvgPool(\cdot)$ denotes an adaptive average pooling layer. $P \in \mathbb{R}^{1\times N}$ is the collection of average feature values. Then P is employed two fully-connected (fc) layers to extract temporal features, as follows:

$$F_1 = relu(w_2(\sigma(w_1 P + b_1)) + b_2) \tag{4}$$

where w_1 and w_2 are the parameters of two fc layers' weights, b_1 and b_2 are the biases of fc layers. $\sigma(\cdot)$ is the sigmoid activation function. F_1, temporal features, is a N size vector, extracted from EEG signals in the temporal domain.

2.2 Frequency Residual Branch

Prior research has demonstrated that multi-band DE features can significantly improve the accuracy of AAD [Jiang *et al.*, 2022]. Our method involves extracting DE values from each band, projecting them onto 2D topological maps to utilize their topological patterns [Snyder, 1987] and extracting spectral-spatial features through multi-band fusion residual blocks [He *et al.*, 2016]. Specifically, we implement a three-step process to extract frequency features.

Firstly, the EEG data is decomposed into five frequency bands [Liu *et al.*, 2017]. Regarding each frequency band containing different spatial characteristics, we extract DE values from five frequency bands and then project them onto five 2D planes. Therefore, there is a $M = [m_1, \ldots, m_i, \ldots, m_5] \in \mathbb{R}^{N\times N\times 5\times 1}$ concatenating five 2D planes, where N is the number of grids in a 2D plane $m_i \in \mathbb{R}^{N\times N\times 1}$. m_i represents a 2D topological plane.

Secondly, we employ a convolutional layer followed by a max pooling layer to process the whole 3D topological collection M as a one-channel feature. This step aims to extract the significant information from the topological collection M.

$$U = Max(tanh(Conv(M))) \tag{5}$$

where $Conv(\cdot)$ denotes the 3D convolutional layer, $Max(\cdot)$ denotes a max pooling layer and $tanh(\cdot)$ denotes the hyperbolic tangent (tanh) function. The processed EEG signals are represented by $U \in \mathbb{R}^{N\times N\times K\times C}$, where K denotes the number of frequency bands and C denotes the channel number.

Finally, as illustrated in Figure 1, we implement multi-band fusion residual blocks to fuse DE features across different frequency bands. These blocks consist of 3D convolutional layers and 3D batch normalization layers. This step fuses all spectral-spatial information contained within topological maps collection U. Additionally, we stack N such multi-band fusion residual blocks to augment the model's capability in extracting frequency features, as follows:

$$\begin{aligned} O^{(0)} &= ConvLayer(ResBlock(U)) \\ O^{(i)} &= ConvLayer(ResBlock(O^{(i-1)})) \end{aligned} \tag{6}$$

where $ConvLayer(\cdot)$ denotes 3D convolutional layer along with 3D batch normalization layer, $ResBlock(\cdot)$ represents the multi-band fusion residual block. And $O^{(i)} \in \mathbb{R}^{N\times N\times K\times C}$ means the i-th hidden output. What is different from the previous multi-band fusion residual blocks is that $O^{(N-1)}$ is employed by a 3D convolution layer to reduce its dimensionality following batch normalization. An adaptive average pooling layer extracts average features, as follows:

$$O^{(N)} = relu(ConvLayer(O^{(N-1)})) \tag{7}$$

$$F_2 = relu(AdaptiveAvgPool(O^{(N)})) \tag{8}$$

where $ConvLayer(\cdot)$, consistent with its previous usage, represents a 3D convolution layer with a 3D batch normalization layer, $relu(\cdot)$ means the rectified linear unit function and $O^{(N)}$ denotes the N-th hidden output of the model. In Eq.

(8), $AdaptiveAvgPool(\cdot)$ refers to an adaptive average pooling layer. F_2, frequency features, presents a feature vector of size N, extracted from EEG signals in frequency domain.

2.3 Feature Fusion & Classifier

To fuse the temporal-frequency domain features, we concatenate them into a new vector. Then through a linear transformation, the model derives the final classification result.

Firstly, we concatenate two vectors to get a new dual-domain feature vector, as follows:

$$F = [F_1, F_2] \tag{9}$$

where F is the new vector concatenating F_1 temporal features and F_2 frequency features. Then, we employ a fc layer to get the final result, as follows:

$$predict = wF + b \tag{10}$$

where w and b are two parameters of the fc layer. $predict$ is the predicted direction label. In the training stage, we apply the binary cross-entropy function to update the parameter.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log Q_i + (1 - y_i) \cdot \log(1 - Q_i)] \tag{11}$$

where y_i means the ground-truth label of i-th decision window, N means the number of samples, Q_i is the corresponding possibility of $predict$ with softmax function processing.

3 Experiments

3.1 AAD Dataset

We extensively evaluate the proposed method in three datasets, namely KUL [Das *et al.*, 2019; Das *et al.*, 2016], DTU [Fuglsang *et al.*, 2018; Fuglsang *et al.*, 2017] and MM-AAD. KUL and DTU are widely used publicly available datasets that contain EEG data for only auditory stimulus scenes. MM-AAD is our self-built dataset for simulating real scenes, which contains EEG data of the audio-only scene and the audio-visual scene.

1) **KUL:** In this study, 16 subjects with normal hearing listen to four Dutch short stories in a soundproof room. Narrated by three male Flemish speakers, the audio is delivered through in-ear headphones, filtered at 4 kHz and set at 60 dB. The study features 'dichotic' (dry) and 'head-related transfer function' (HRTF) listening scenarios, focusing on one of two overlapping male voices from either 90° left or right. Each subject completes 8 trials, each lasting 6 minutes, in varying order. The 64-channel EEG data is recorded using the BioSemi ActiveTwo system at an 8196 Hz sampling rate.

2) **DTU:** In this study, 18 subjects with normal hearing listen to Danish audiobooks through ER-2 earphones at 65 dB and 48 kHz. The audiobooks, with varied reverberation levels and narrated by a male and female speaker, simulate different environments. Subjects focus on one voice of two competing voices played concurrently at a 60° orientation. They complete 60 trials, each 50 seconds long, with changing sound streams and narrator genders. EEG data is captured from 64 channels using the BioSemi ActiveTwo system at a 512 Hz sampling rate, in line with the 10/20 international system.

Dataset	Scene	Subjects	Duration (minutes)	Stimulus Language
KUL	audio-only	16	48	Dutch
DTU	audio-only	18	50	Danish
MM-AAD	audio-only	50	55	Chinese
	audio-visual	50	55	Chinese

Table 1: Details of three datasets used in experiments

3) **MM-AAD:** This dataset is a multi-modal AAD dataset constructed by ourselves, which well simulates the multi-modal input of visual and auditory in real scenes. This dataset contains EEG recordings from 50 subjects (34 males and 16 females) with normal hearing, tested with audio-only and audio-visual stimuli. There are 40 stories chosen from a classical Chinese short story collection, narrated by male and female voices. Subjects should pay attention to stories playing from their left or right spatial direction. Each subject completes 20 trials of approximately 165 seconds, with varying audio stream locations and narrator genders. EEG data is captured from 32 channels at a 4000 Hz sampling rate, following the 10/20 international system.

3.2 Data Preprocessing

To maintain fairness in experiments with different EEG datasets, specific preprocessing is applied to each. For the KUL dataset, EEG data are re-referenced to mastoid electrodes' average response, bandpass filtered between 0.1Hz and 50 Hz and downsampled to 128 Hz. The DTU dataset is filtered to remove 50Hz line noise, downsampled to 128 Hz and high-pass filtered at 0.1 Hz. Eye artifacts are removed using joint decorrelation and data are re-referenced to the average EEG channel response [Fuglsang *et al.*, 2017]. For the MM-AAD, data are bandpass filtered from 0.1 Hz to 50 Hz, with 50 Hz noise removed by a notch filter and downsampled to 128 Hz. Eye artifacts are eliminated and independent component analysis (ICA) is applied for further noise removal.

The performance of the proposed model is evaluated under the three different sliding decision window lengths.

3.3 Baseline Methods

We compare our proposed DBPNet with the two following baselines:

- **SSF-CNN [Cai *et al.*, 2021b]:** A CNN based on alpha power neural signals to extract spectral-spatial features for AAD.
- **MBSSFCC [Jiang *et al.*, 2022]:** A CNN network and convolutional-long-short-term-memory (ConvLSTM) to extract spectro-spatial-temporal features from multiple frequency bands for AAD.

3.4 Network Configuration

Our primary task in AAD involves determining the direction of sound (left or right), essentially a binary classification problem. The resulting accuracy is defined as the percentage

Branch	Layers	Value
TABNet	Transformer encoder layer	1
	Convolution kernel	64×7
	fc layer	2
FRBNet	Convolution kernel	32×1×7×7
	Max pool kernel	1×3×3
	Residual blocks	3
	Convolution kernel	4×1×1
	Adaptive average pool output	1×1×1

Table 2: The detailed parameters of layers in TABNet and FRBNet.

of the correct classification result divided by the total number of classifications. We refine the model parameters by reducing the loss value and evaluate the effectiveness using the accuracy percentages derived from the test set.

Taking the 1-second decision window in KUL as an example, after data preprocessing, there is a total of 5,752 decision windows per subject, including 4,658 decision windows for training and 576 decision windows for testing.

For TABNet, 1-second EEG data with 128 samples and 64 channels are denoted as $R \in \mathbb{R}^{128\times64}$. Firstly, the data is employed by the CSP algorithm. And, through the transformer encoder with self-attention mechanism, we can get the encoded data $S \in \mathbb{R}^{128\times64}$. Then there is a convolutional layer and two fc layers (input: 64, hidden: 16, output: 4) to extract temporal features $F_1 \in \mathbb{R}^{1\times4}$ from S. For FRBNet, we extract DE values from five frequency bands. Then through 3D to 2D projection method, we can get five 2D planes, each of which can be denoted as $m_i \in \mathbb{R}^{32\times32\times1}$ and concatenate them to a 3D map collection $M = [m_1, m_2, m_3, m_4, m_5] \in \mathbb{R}^{32\times32\times5\times1}$. Furthermore, after multi-band fusion residual blocks and adaptive average pooling processing, we can get frequency features $F_2 \in \mathbb{R}^{1\times4}$. Finally, in the feature fusion and classifier stage, we concatenate temporal feature F_i and frequency feature F_2 to a new vector $F \in \mathbb{R}^{1\times8}$. Then, through a fc layer (input: 8, output: 2), we can get the final classification result.

The detailed values of layers in TABNet and FRBNet can be found in Table 2. All studies are performed using PyTorch. We implement dropout layers and the early stopping scheme.

4 Results

4.1 Auditory Attention Analysis

To evaluate the proposed model's decoding performance in AAD, we compare DBPNet with other AAD models for three decision windows, detailed in Table 3. We reproduce results for available models and cite the results from the corresponding papers for not open ones.

DBPNet showcases strong performance in KUL, DTU and MM-AAD datasets. In KUL, the accuracy is from 87.1% (SD: 6.55%) for 0.1-second to 96.5% (SD: 3.50%) for 2-second and in DTU from 75.1% (SD: 4.87%) for 0.1-second to 86.5% (SD: 5.34%) for 2-second. In MM-AAD, the detection accuracy is from 91.4% (SD: 4.63%) for 0.1-second to 92.5% (SD: 4.59) for 2-second in the audio-only scene and from 92.1% (SD: 4.47%) for 0.1-second to 93.4% (SD: 4.86%) for 2-second in the audio-visual scene. Their accuracy increases consistently with longer decision windows, aligning with previous studies [Jiang *et al.*, 2022; Cai *et al.*, 2023b; Cai *et al.*, 2023a].

Expanding on initial observations, it's clear that the accuracy of classification from the DTU is approximately 11% lower than those from the KUL, a trend echoed in previous studies [Jiang *et al.*, 2022; Cai *et al.*, 2023b; Cai *et al.*, 2023a]. This discrepancy is influenced by various factors. (1) The orientation of auditory stimuli. In KUL, the speech stimuli originate from 90° left and 90° right directions, but in DTU, 60° to the left and right (2) The absence of ambient reverberation. The DTU includes varying levels of room reverberation, whereas the KUL dataset does not. (3) The gender of speakers. The KUL exclusively includes male speech stimuli, but the DTU includes both male and female speakers.

Besides, under the evaluation of all models in MM-AAD, the accuracy in the audio-visual scene is commonly higher than the accuracy in the audio-only scene, which shows that vision can aid in accurately identifying the source of sound.

4.2 Ablation Study

In the ablation study, we compare two single branches (TABNet and FRBNet) with DBPNet. Both experiments are evaluated under identical conditions to those of preceding studies. The results of the ablation study are displayed in Table 3.

In KUL, TABNet's accuracy notably increases from 82.5% (SD: 6.64%) for 0.1-second to 93.5% (SD: 5.41%) for 2-second, for FRBNet, from 80.4% (SD: 7.08%) for 0.1-second to 89.5% (SD: 8.16) for 2-second. Both accuracy of the two branches is lower than DBPNet's results, from 87.1% (SD: 6.55%) for 0.1-second to 96.5% (SD: 3.50%) for 2-second. As shown in Table 3 and Figure 2, the trends are similar in DTU. Besides, in MM-AAD, the detection accuracy of TABNet is improved from 90.9% (SD: 4.38%) for 0.1-second to 91.9% (SD: 4.93%) for 2-second in the audio-only scene and from 91.4% (SD: 4.88) for 0.1-second to 92.4% (SD: 5.24%) for 2-second in the audio-visual scene. FRBNet has the same trend as TABNet in both scenes. Similar to previous results of DBPNet, the decoding accuracy of single-branch networks in audio-visual scenes is higher than networks in audio-only scenes. A comparative analysis of DBPNet against these single-branch networks reveals its consistent superiority across three datasets, displaying a notable average 5.73% relative improvement in KUL, 7.42% in DTU, 8.11% in the audio-only scene of MM-AAD and 7.27% in audio-visual scene of MM-AAD, which further underscores our DBPNet effectively fuses temporal and frequency features to improve the decoding performance of AAD.

5 Discussion

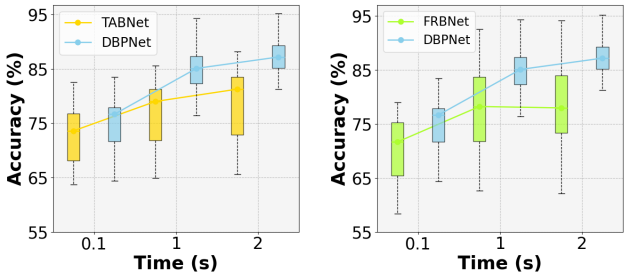
5.1 Comparative Study

We compare our proposed DBPNet with the other six models for AAD classification, as shown in Table 3. The results show our DBPNet achieves state-of-the-art performance and proves the effectiveness of temporal-frequency fusion.

In DTU, for 0.1-second, DBPNet achieves respectively relative improvements of 32.5%, 14.3%, 3.59%, 20.2% and

Dataset	Scene	Model	Decision Window		
			0.1-second	1-second	2-second
KUL	audio-only	CNN [Vandecappelle *et al.*, 2021]	74.3	84.1	85.7
		STAnet [Su *et al.*, 2022]	80.8	90.1	91.4
		BSAnet [Cai *et al.*, 2023a]	-	93.7 ± 4.02	95.2 ± 3.08
		SSF-CNN* [Cai *et al.*, 2021b]	76.3 ± 8.47	84.4 ± 8.67	87.8 ± 7.87
		MBSSFCC* [Jiang *et al.*, 2022]	79.0 ± 7.34	86.5 ± 7.16	89.5 ± 6.74
		DBPNet (ours)	**87.1 ± 6.55**	**95.0 ± 4.16**	**96.5 ± 3.50**
DTU	audio-only	CNN [Vandecappelle *et al.*, 2021]	56.7	63.3	65.2
		STAnet [Su *et al.*, 2022]	65.7	71.9	73.7
		EEG-Graph Net [Cai *et al.*, 2023b]	72.5 ± 7.41	78.7 ± 6.47	79.4 ± 7.16
		BSAnet [Cai *et al.*, 2023a]	-	83.1 ± 6.75	85.6 ± 6.47
		SSF-CNN* [Cai *et al.*, 2021b]	62.5 ± 3.40	69.8 ± 5.12	73.3 ± 6.21
		MBSSFCC* [Jiang *et al.*, 2022]	66.9 ± 5.00	75.6 ± 6.55	78.7 ± 6.75
		DBPNet (ours)	**75.1 ± 4.87**	**83.9 ± 5.95**	**86.5 ± 5.34**
MM-AAD	audio-only	SSF-CNN* [Cai *et al.*, 2021b]	56.5 ± 5.71	57.0 ± 6.55	57.9 ± 7.47
		MBSSFCC* [Jiang *et al.*, 2022]	75.3 ± 9.27	76.5 ± 9.90	77.0 ± 9.92
		DBPNet (ours)	**91.4 ± 4.63**	**92.0 ± 5.42**	**92.5 ± 4.59**
	audio-visual	SSF-CNN* [Cai *et al.*, 2021b]	56.6 ± 3.82	57.2 ± 5.59	58.2 ± 6.39
		MBSSFCC* [Jiang *et al.*, 2022]	77.2 ± 9.01	78.1 ± 10.1	78.4 ± 9.57
		DBPNet (ours)	**92.1 ± 4.47**	**92.8 ± 5.94**	**93.4 ± 4.86**

Table 3: AAD accuracy(%) achieved by the proposed DBPNet in KUL, DTU and MM-AAD, compared with other models for three decision windows (0.1-second, 1-second, 2-second). ”-” means there are no corresponding experiments conducted or no results in the corresponding paper. The results of ”*” marked baseline models have been reproduced.



(a) TABNet-DBPNet (b) FRBNet-DBPNet

Figure 2: AAD average accuracy(%) of the ablation study across all subjects in DTU. (a) represents the box map of TABNet and DBPNet average accuracy. (b) represents the box map of FRBNet and DBPNet average accuracy.

12.3% compared with CNN, STAnet, EEG-Graph Net, SSF-CNN and MBSSFCC. For 1-second, the relative improvements are 32.5%, 16.7%, 6.61%, 0.96%, 20.2% and 11.0%, compared with CNN, STAnet, EEG-Graph Net, BSAnet, SSF-CNN and MBSSFCC. In addition, the relative improvements, compared with those six models, are 32.7%, 17.4%, 8.94%, 1.05%, 18.0% and 9.91%. Similarly in KUL and MM-AAD, the relative improvements of DBPNet are higher than other AAD models. Especially in MM-AAD, DBPNet achieves 20.4% relative improvement for 0.1-second, compared with the best baseline. These results highlight DBPNet's effectiveness and feasibility of temporal and frequency domain extraction and fusion under different datasets.

5.2 Ablation Analysis

DBPNet's success validates the effectiveness of the temporal-frequency fusion strategy. The temporal attentive branch extracts temporal features and the frequency residual branch extracts spectral-spatial features. Both branches can obtain corresponding features and are fused effectively.

Effects of TABNet

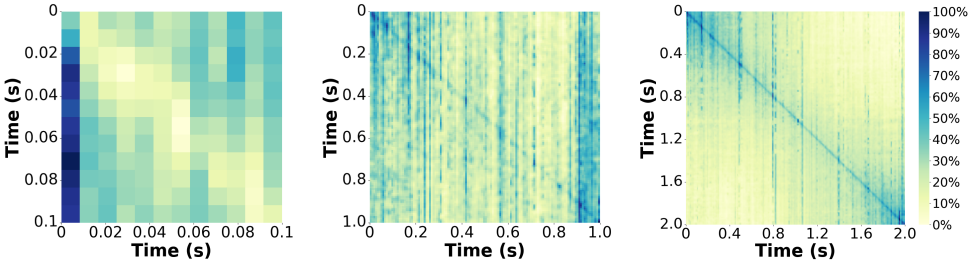
Previous studies have shown there is a rhythmic relationship between neural activities and external stimuli [Fiebelkorn and Kastner, 2019]. Therefore, there is a certain connection between EEG signals, with some signals being more valuable for AAD. The transformer encoder can learn weights from EEG signals and encode them automatically. In Figure 3(a)-(c), EEG signals at different times have different weighted relationships. Especially in Figure 3 (b),(c), obvious vertical strips confirm partial signals hold higher value for tracking spatial attention. However, for Figure 3(a), strips look insignificant suggesting limitations in capturing effective EEG signals, which possibly be caused by the latency of the human brain [Power *et al.*, 2012]. Overall, the attention mechanism of the temporal attentive branch can automatically discover the time-varying information from EEG time-series signals.

Effects of FRBNet

As shown in Table 3 and Table 4, the classification accuracy of FRBNet in the KUL, DTU and MM-AAD datasets matches that frequency-focused model: SSF-CNN [Cai *et al.*, 2021b] and MBSSFCC [Jiang *et al.*, 2022]. The performance suggests that FRBNet with multi-band fusion residual blocks can effectively extract spectral-spatial features through multiple frequency bands of EEG signals [Wöstmann *et al.*, 2016].

Dataset	Scene	Model	Decision Window		
			0.1-second	1-second	2-second
KUL	audio-only	TABNet	82.5 ± 6.64	91.2 ± 5.60	93.5 ± 5.41
		FRBNet	80.4 ± 7.08	87.5 ± 7.68	89.5 ± 8.16
		DBPNet (ours)	**87.1 ± 6.55**	**95.0 ± 4.16**	**96.5 ± 3.50**
DTU	audio-only	TABNet	72.9 ± 5.34	76.8 ± 6.25	79.0 ± 6.68
		FRBNet	70.4 ± 5.86	77.8 ± 8.11	79.0 ± 8.27
		DBPNet (ours)	**75.1 ± 4.87**	**83.9 ± 5.95**	**86.5 ± 5.34**
MM-AAD	audio-only	TABNet	90.9 ± 4.38	91.2 ± 5.34	91.9 ± 4.93
		FRBNet	77.4 ± 9.49	80.2 ± 10.1	81.3 ± 10.1
		DBPNet (ours)	**91.4 ± 4.63**	**92.0 ± 5.42**	**92.5 ± 4.59**
	audio-visual	TABNet	91.4 ± 4.88	92.0 ± 5.04	92.4 ± 5.24
		FRBNet	79.2 ± 9.37	82.5 ± 9.01	83.3 ± 9.88
		DBPNet (ours)	**92.1 ± 4.47**	**92.8 ± 5.94**	**93.4 ± 4.86**

Table 4: AAD accuracy(%) achieved by the ablation study in KUL, DTU and MM-AAD using DBPNet. TABNet is the network only containing a temporal attentive branch. FRBNet is the network only containing a frequency residual branch.



(a) (b) (c)

Figure 3: Heat maps of average attention weights using DBPNet across all subjects for three decision windows in DTU. Each row represents the attention weights of a specific EEG signal in the sequence with all signals. (a) The average attention weights for 0.1-second. (b) The average attention weight for 1-second. (c) The average attention weight for 2-second.

Model	Trainable Parameters
SSF-CNN [Cai *et al.*, 2021b]	4.21M
MBSSFCC [Jiang *et al.*, 2022]	83.91M
DBPNet (ours)	**0.91M**

Table 5: The number of trainable parameters of DBPNet and two open-source models. "M" denotes million, equal to 10^6.

5.3 Computational Cost

Table 5 compares the parameter counts of DBPNet with SSF-CNN [Cai *et al.*, 2021b] and MBSSFCC [Jiang *et al.*, 2022], revealing that DBPNet requires 0.91M trainable parameters in total, approximately 3.6 times fewer than SSF-CNN and 91 times fewer than MBSSFCC. It demonstrates that FRBNet reduces the number of trainable parameters of convolution networks in AAD. This stark reduction in computational cost makes DBPNet highly suitable for low-power devices like hearing aids.

6 Conclusion

In this paper, to address the lack of work on fusion between the temporal and frequency domains in AAD, we propose the DBPNet, a novel dual-branch parallel network with temporal-frequency fusion. The model utilizes a dual-branch approach to parallelly extract features from the temporal and frequency domains. For the temporal attentive branch, by incorporating the transformer encoder, extract time-varying information from EEG time-series signals as temporal features. For the frequency residual branch, we employ multi-band fusion residual blocks to extract spectral-spatial features from multi-band EEG signals as frequency features. Finally, through the fusion of temporal-frequency features, we can obtain comprehensive EEG features that contain both time-varying and spectral-spatial information and then get the classification results. Our evaluation is conducted in three datasets: KUL, DTU and MM-AAD. Especially in MM-AAD, our DBPNet achieves 20.4% relative improvement for 0.1-second decision windows, but trainable parameter counts are reduced by about 91 times, compared with the best baseline. Experimental results show that DBPNet achieves the effective extraction and fusion of temporal-frequency domain features. For further study, we will explore a more effective strategy of temporal-frequency feature fusion and use GNN or other neural networks to further improve the performance of AAD.

Acknowledgements

This work is supported by the STI 2030—Major Projects (No. 2021ZD0201500), the National Natural Science Foundation of China (NSFC) (No.62201002), Distinguished Youth Foundation of Anhui Scientific Committee (No. 2208085J05), Special Fund for Key Program of Science and Technology of Anhui Province (No. 202203a07020008), Open Fund of Key Laboratory of Flight Techniques and Flight Safety, CACC (No, FZ2022KF15).

Contribution Statement

Qinke Ni, Hongyu Zhang and Cunhang Fan are equal contributors. Cunhang Fan and Zhao Lv are corresponding authors.

References

[Aftanas *et al.*, 2004] Lyubomir I Aftanas, Natalya V Reva, Anton A Varlamov, Sergey V Pavlov, and Victor P Makhnev. Analysis of evoked eeg synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics. *Neuroscience and behavioral physiology*, 34:859–867, 2004.

[Bassett and Sporns, 2017] Danielle S Bassett and Olaf Sporns. Network neuroscience. *Nature neuroscience*, 20(3):353–364, 2017.

[Bednar *et al.*, 2017] Adam Bednar, Francis M Boland, and Edmund C Lalor. Different spatio-temporal electroencephalography features drive the successful decoding of binaural and monaural cues for sound localization. *European Journal of Neuroscience*, 45(5):679–689, 2017.

[Cai *et al.*, 2021a] Siqi Cai, Peiwen Li, Enze Su, and Longhan Xie. Auditory attention detection via cross-modal attention. *Frontiers in neuroscience*, 15:652058, 2021.

[Cai *et al.*, 2021b] Siqi Cai, Pengcheng Sun, Tanja Schultz, and Haizhou Li. Low-latency auditory spatial attention detection based on spectro-spatial features from eeg. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5812–5815. IEEE, 2021.

[Cai *et al.*, 2023a] Siqi Cai, Peiwen Li, and Haizhou Li. A bio-inspired spiking attentional neural network for attentional selection in the listening brain. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Cai *et al.*, 2023b] Siqi Cai, Tanja Schultz, and Haizhou Li. Brain topology modeling with eeg-graphs for auditory spatial attention detection. *IEEE Transactions on Biomedical Engineering*, 2023.

[Cheng *et al.*, 2020] Nicholas Cheng, Kok Soon Phua, Hwa Sen Lai, Pui Kit Tam, Ka Yin Tang, Kai Kei Cheng, Raye Chen-Hua Yeow, Kai Keng Ang, Cuntai Guan, and Jeong Hoon Lim. Brain-computer interface-based soft robotic glove rehabilitation for stroke. *IEEE Transactions on Biomedical Engineering*, 67(12):3339–3351, 2020.

[Choi *et al.*, 2013] Inyong Choi, Siddharth Rajaram, Lenny A Varghese, and Barbara G Shinn-Cunningham. Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography. *Frontiers in human neuroscience*, 7:115, 2013.

[Choi *et al.*, 2014] Inyong Choi, Le Wang, Hari Bharadwaj, and Barbara Shinn-Cunningham. Individual differences in attentional modulation of cortical responses correlate with selective attention performance. *Hearing research*, 314:10–19, 2014.

[Ciccarelli *et al.*, 2019] Gregory Ciccarelli, Michael Nolan, Joseph Perricone, Paul T Calamia, Stephanie Haro, James O'sullivan, Nima Mesgarani, Thomas F Quatieri, and Christopher J Smalt. Comparison of two-talker attention decoding from eeg with nonlinear neural networks and linear methods. *Scientific reports*, 9(1):11538, 2019.

[Das *et al.*, 2016] Neetha Das, Wouter Biesmans, Alexander Bertrand, and Tom Francart. The effect of head-related filtering and ear-specific decoding bias on auditory attention detection. *Journal of neural engineering*, 13(5):056014, 2016.

[Das *et al.*, 2019] Neetha Das, Tom Francart, and Alexander Bertrand. Auditory attention detection dataset kuleuven. *Zenodo*, 2019.

[Deng *et al.*, 2020] Yuqi Deng, Inyong Choi, and Barbara Shinn-Cunningham. Topographic specificity of alpha power during auditory spatial attention. *Neuroimage*, 207:116360, 2020.

[Ding and Simon, 2012] Nai Ding and Jonathan Z Simon. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29):11854–11859, 2012.

[Fiebelkorn and Kastner, 2019] Ian C Fiebelkorn and Sabine Kastner. A rhythmic theory of attention. *Trends in cognitive sciences*, 23(2):87–101, 2019.

[Frantzidis *et al.*, 2010] Christos A Frantzidis, Charalampos Bratsas, Christos L Papadelis, Evdokimos Konstantinidis, Costas Pappas, and Panagiotis D Bamidis. Toward emotion aware computing: an integrated approach using multichannel neurophysiological recordings and affective visual stimuli. *IEEE transactions on Information Technology in Biomedicine*, 14(3):589–597, 2010.

[Fuglsang *et al.*, 2017] Søren Asp Fuglsang, Torsten Dau, and Jens Hjortkjær. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *Neuroimage*, 156:435–444, 2017.

[Fuglsang *et al.*, 2018] Søren A Fuglsang, Daniel DE Wong, and Jens Hjortkjær. Eeg and audio dataset for auditory attention decoding. *Zenodo*, 2018.

[Geirnaert *et al.*, 2020] Simon Geirnaert, Tom Francart, and Alexander Bertrand. Fast eeg-based decoding of the directional focus of auditory attention using common spatial patterns. *IEEE Transactions on Biomedical Engineering*, 68(5):1557–1568, 2020.

[Geirnaert *et al.*, 2021] Simon Geirnaert, Servaas Vandecappelle, Emina Alickovic, Alain de Cheveigne, Edmund Lalor, Bernd T Meyer, Sina Miran, Tom Francart, and

Alexander Bertrand. Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices. *IEEE Signal Processing Magazine*, 38(4):89–102, 2021.

[Golumbic *et al.*, 2013] Elana M Zion Golumbic, Nai Ding, Stephan Bickel, Peter Lakatos, Catherine A Schevon, Guy M McKhann, Robert R Goodman, Ronald Emerson, Ashesh D Mehta, Jonathan Z Simon, et al. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, 77(5):980–991, 2013.

[Haykin and Chen, 2005] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Jiang *et al.*, 2022] Yifan Jiang, Ning Chen, and Jing Jin. Detecting the locus of auditory attention based on the spectro-spatial-temporal analysis of eeg. *Journal of Neural Engineering*, 19(5):056035, 2022.

[Jones and Boltz, 1989] Mari Riess Jones and Marilyn Boltz. Dynamic attending and responses to time. *Psychological review*, 96(3):459, 1989.

[Jones *et al.*, 2002] Mari Riess Jones, Heather Moynihan, Noah MacKenzie, and Jennifer Puente. Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological science*, 13(4):313–319, 2002.

[Liu *et al.*, 2017] Yong-Jin Liu, Minjing Yu, Guozhen Zhao, Jinjing Song, Yan Ge, and Yuanchun Shi. Real-time movie-induced discrete emotion recognition from eeg signals. *IEEE Transactions on Affective Computing*, 9(4):550–562, 2017.

[Lotte and Guan, 2010] Fabien Lotte and Cuntai Guan. Regularizing common spatial patterns to improve bci designs: unified theory and new algorithms. *IEEE Transactions on biomedical Engineering*, 58(2):355–362, 2010.

[Mesgarani and Chang, 2012] Nima Mesgarani and Edward F Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236, 2012.

[Miran *et al.*, 2018] Sina Miran, Sahar Akram, Alireza Sheikhattar, Jonathan Z Simon, Tao Zhang, and Behtash Babadi. Real-time tracking of selective auditory attention from m/eeg: A bayesian filtering approach. *Frontiers in neuroscience*, 12:262, 2018.

[O'sullivan *et al.*, 2015] James A O'sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral cortex*, 25(7):1697–1706, 2015.

[Power *et al.*, 2012] Alan J Power, John J Foxe, Emma-Jane Forde, Richard B Reilly, and Edmund C Lalor. At what time is the cocktail party? a late locus of selective attention to natural speech. *European Journal of Neuroscience*, 35(9):1497–1503, 2012.

[Puffay *et al.*, 2023] Corentin Puffay, Bernd Accou, Lies Bollens, Mohammad Jalilpour Monesi, Jonas Vanthornhout, Tom Francart, et al. Relating eeg to continuous speech using deep neural networks: a review. *arXiv preprint arXiv:2302.01736*, 2023.

[Ramoser *et al.*, 2000] Herbert Ramoser, Johannes Muller-Gerking, and Gert Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE transactions on rehabilitation engineering*, 8(4):441–446, 2000.

[Shi *et al.*, 2013] Li-Chen Shi, Ying-Ying Jiao, and Bao-Liang Lu. Differential entropy feature for eeg-based vigilance estimation. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6627–6630. IEEE, 2013.

[Snyder, 1987] John Parr Snyder. *Map projections–A working manual*, volume 1395. US Government Printing Office, 1987.

[Stone, 1966] Harold S Stone. R66-50 an algorithm for the machine calculation of complex fourier series. *IEEE Transactions on Electronic Computers*, (4):680–681, 1966.

[Su *et al.*, 2022] Enze Su, Siqi Cai, Longhan Xie, Haizhou Li, and Tanja Schultz. Stanet: A spatiotemporal attention network for decoding auditory spatial attention from eeg. *IEEE Transactions on Biomedical Engineering*, 69(7):2233–2242, 2022.

[Vandecappelle *et al.*, 2021] Servaas Vandecappelle, Lucas Deckers, Neetha Das, Amir Hossein Ansari, Alexander Bertrand, and Tom Francart. Eeg-based detection of the locus of auditory attention with convolutional neural networks. *Elife*, 10:e56481, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Viswanathan *et al.*, 2019] Vibha Viswanathan, Hari M Bharadwaj, and Barbara G Shinn-Cunningham. Electroencephalographic signatures of the neural representation of speech during selective attention. *Eneuro*, 6(5), 2019.

[Wöstmann *et al.*, 2016] Malte Wöstmann, Björn Herrmann, Burkhard Maess, and Jonas Obleser. Spatiotemporal dynamics of auditory attention synchronize with speech. *Proceedings of the National Academy of Sciences*, 113(14):3873–3878, 2016.

[Zheng and Lu, 2015] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3):162–175, 2015.