CROSS-ATTENTION-GUIDED WAVENET FOR MEL SPECTROGRAM RECONSTRUCTION IN THE ICASSP 2024 AUDITORY EEG CHALLENGE

Yuan Fang, Hao Li, Xueliang Zhang, Fei Chen, Guanglai Gao

Inner Mongolia University, China
Southern University of Science and Technology, China

Reporter: Yuan Fang
32209021@mail.imu.edu.cn

2024.ieeeicassp.org
Outline

- BACKGROUND
- PROPOSED MODEL
- EXPERIMENT
- CONCLUSIONS
Task 2: regression

Fig.1 Task 2 of the Auditory EEG Challenge: EEG-to-MEL Spectrogram Reconstruction.

① The ICASSP 2024 Auditory EEG Challenge Task 2 is a regression task.
② Predicting the mel spectrogram based on the input EEG signal.
③ The model is evaluated using Pearson correlation.
Background - Drawbacks

1) Inter-individual differences.

2) Low signal-to-noise ratio.

3) EEG to speech is a challenging problem due to its nonlinear nature
Outline

- BACKGROUND
- PROPOSED MODEL
- EXPERIMENTS
- CONCLUSIONS
① Cross-Attention-Guided WaveNet for Mel spectrogram reconstruction.
② The coarse-to-fine granularity strategy.
③ Cross-attention mechanism is used to fuse two different modalities.
④ A combined loss function is used to optimize multiple outputs.
⑤ The Mixup augmentation technique to mitigate overfitting and improve generalization performance.
① In the field of deep learning, multi-objective sequential learning has become a common strategy.

② The coarse-to-fine granularity approach is used to estimate multiple objectives.

③ The effectiveness of this strategy was validated through experimental ablation studies.
① WaveNet effectively learns features from sequential data by utilizing dilated convolutions.

② WaveNet showed significant performance in the ICASSP 2023 Auditory EEG Challenge.

Fig. 4 WaveNet Architecture
① Cross-Attention mechanism is a multi-head attention mechanism commonly used in deep learning-based methods as a modality fusion module.

② Cross-Attention mechanism captures dependencies between different scales of features and modalities, facilitating effective information exchange and fusion.
PROPOSED MODEL - Loss

① multiple loss functions jointly to ensure stable training of the model.

② L1 norm

③ Negative Pearson correlation coefficient (NP)

④ Kullback-Leibler Divergence (KL divergence)

\[ \text{Loss} = \alpha \cdot L_1 + NP + KL \]

\[ L_1 = L_1(\text{Env}) + L_1(\text{Mel10}) + L_1(\text{Mel80}) + L_1(\text{Mag}) \]

\[ NP = NP(\text{Env}) + NP(\text{Mel10}) + NP(\text{Mel80}) + NP(\text{Mag}) \]

\[ KL = KL(\text{Mel10}) \]
Considering the constraints of a limited dataset, the Mixup data augmentation technique was adopted to alleviate overfitting and improve performance:

\[ x = \lambda x_i + (1 - \lambda) x_j \]
\[ y = \lambda y_i + (1 - \lambda) y_j \]

In the Mixup data augmentation technique, \( x_i \) and \( x_j \) represent two segments of EEG from different participants, while \( y_i \) and \( y_j \) represent the corresponding audio signals. The parameter \( \lambda \) is randomly sampled from the range \([0,1]\).
Outline

- BACKGROUND
- PROPOSED MODEL
- EXPERIMENTS
- CONCLUSIONS
• **Auditory EEG corpus:**
  • Auditory EEG challenge
  • **Train set:**
    • Sub-01 to Sub-26
    • Sub-43 to Sub-85
  • **Val set:**
    • Sub-27 to Sub-42
  • **Test set:**
    • Sub-86 to Sub-104
① The proposed model achieved a PCC score of 0.0651, outperforming other baseline models.
② The proposed model ranked second out of 48 teams in the Auditory EEG Challenge 2024 Task 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>PCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLAII</td>
<td>0.0470</td>
</tr>
<tr>
<td>DPRNN</td>
<td>0.0554</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.0651</td>
</tr>
</tbody>
</table>

Table 1 Comparative Analysis of Models on validation set
This ablation method solely utilizes the WaveNet module to reconstruct the Mel spectrogram.
This ablation method involves removing the last two decoding blocks. The purpose is to examine the influence of the coarse-to-fine granularity strategy.
This ablation method omits the mixed data augmentation technique. The purpose is to evaluate the impact of data augmentation operations on the model's performance.
① Each module of the model has made a significant contribution to the overall performance.
② The coarse-to-fine granularity strategy improved the performance by 0.002.
③ The decoding block and coarse-to-fine granularity strategy led to a 0.0071 improvement.
④ Mixup contributed an improvement effect of 0.0039.

<table>
<thead>
<tr>
<th>Model</th>
<th>PCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ablation-1</td>
<td>0.0580</td>
</tr>
<tr>
<td>Ablation-2</td>
<td>0.0631</td>
</tr>
<tr>
<td>Ablation-3</td>
<td>0.0612</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.0651</td>
</tr>
</tbody>
</table>

Table 2 Ablation experiments results
Outline

- BACKGROUND
- PROPOSED MODEL
- EXPERIMENTS
- CONCLUSIONS
Conclusions

- The proposed CAT-guided WaveNet model leverages CAT to bridge the gap between different modalities and utilizes WaveNet with a coarse-to-fine granularity to construct the Mel spectrogram.

- Compared to baseline, the proposed method demonstrates stronger performance and improved generalization ability on unseen data.

- The code has been uploaded to GitHub.

THANK YOU

Speech Signal Processing Group, Inner Mongolia University