CAT-Net: A Cross-Attention Tone Network for Cross-Subject EEG-EMG Fusion Tone Decoding

Yifan Zhuang[3], Calvin Huang[4], Zepeng Yu[5], Yongjie Zou[2*], Jiawei Ju[1,2*]

[1]Shanghai Center for Brain Science and Brain-Inspired Technology
[2] Lingang Laboratory
[3]Sony Interactive Entertainment, United States
[4]Amazon.com
[5]Independent Researcher
evelyn.zhuang@sony.com, chuang28@stevens.edu, jeffreyzepengyu@ucla.edu, zouyj@lglab.ac.cn, jiawei_ju@163.com

Abstract

Brain-computer interface (BCI) speech decoding has emerged as a promising tool for assisting individuals with speech impairments. In this context, the integration of electroencephalography (EEG) and electromyography (EMG) signals offers strong potential for enhancing decoding performance. Mandarin tone classification presents particular challenges, as tonal variations convey distinct meanings even when phonemes remain identical. In this study, we propose a novel cross-subject multimodal BCI decoding framework that fuses EEG and EMG signals to classify four Mandarin tones under both audible and silent speech conditions. Inspired by the cooperative mechanisms of neural and muscular systems in speech production, our neural decoding architecture combines spatial-temporal feature extraction branches with a cross-attention fusion mechanism, enabling informative interaction between modalities. We further incorporate domain-adversarial training to improve cross-subject generalization. We collected 4,800 EEG trials and 4,800 EMG trials from 10 participants using only twenty EEG and five EMG channels, demonstrating the feasibility of minimal-channel decoding. Despite employing lightweight modules, our model outperforms state-of-the-art baselines across all conditions, achieving average classification accuracies of 87.83% for audible speech and 88.08% for silent speech. In cross-subject evaluations, it still maintains strong performance with accuracies of 83.27% and 85.10% for audible and silent speech, respectively. We further conduct ablation studies to validate the effectiveness of each component. Our findings suggest that tone-level decoding with minimal EEG-EMG channels is feasible and potentially generalizable across subjects, contributing to the development of practical BCI applications.

Code — https://github.com/YifanZhuang/CAT-Net

Introduction

Brain-computer interface (BCI) speech decoding represents a transformative technology for individuals with speech impairments caused by stroke, ALS, and brainstem injuries (Brumberg et al. 2010). Among tonal languages, Mandarin Chinese presents unique challenges due to its four distinct tones, which fundamentally alter word meanings despite identical phonemes. The development of effective tone classification systems is crucial for creating practical BCI communication interfaces that can operate effectively in Chinese-speaking populations.

Electroencephalography (EEG) has emerged as a promising non-invasive technique for neural speech decoding, offering excellent temporal resolution and portability advantages (Sharon et al. 2020; Dash, Ferrari, and Wang 2021). EEG captures neural activities associated with speech planning, phonological processing, and auditory feedback mechanisms. However, extracting Mandarin tones from EEG signals presents significant challenges due to the subtle and spatially distributed nature of pitch-related neural signals. Traditional EEG systems suffer from limited spatial resolution, resulting in high misclassification rates, particularly for perceptually challenging tone pairs such as Tone 2 (rising) and Tone 4 (falling), which share similar pitch contour characteristics and are frequently confused in both neural decoding and perceptual studies (Li, Pun, and Chen 2021).

To address these limitations, researchers have increasingly explored multimodal approaches by integrating surface electromyography (sEMG) signals from facial and jaw muscles during both audible and silent speech conditions (Barrientos Rojas et al. 2022). EMG provides complementary articulatory information about muscle activation patterns that directly correlate with tone production mechanisms. The activation of specific muscle groups, including the zygomatic and masseter muscles, produces distinct patterns corresponding to different pitch contours in Mandarin tones. By combining EEG and EMG signals, researchers can simultaneously monitor neural speech intentions and peripheral articulatory execution, potentially improving decoding accuracy across both overt and covert speech conditions. In recent studies, Li et al. demonstrated that fusing EEG and EMG signals can lead to more robust silent speech recognition systems (Li et al. 2023).

Despite promising results, current EEG-EMG fusion systems face several critical limitations. First, existing approaches typically require high-density electrode configurations with numerous EEG channels and multiple EMG sen-

*Corresponding authors.

sors, making them impractical for daily use due to setup complexity and user discomfort (Zhang et al. 2024). Second, most current systems lack effective mechanisms for capturing the intricate interactions between neural and muscular signals, often relying on simple concatenation or basic fusion strategies. Third, and most importantly, these systems demonstrate poor generalization capabilities across different users, with performance degrading substantially when applied to unseen subjects due to anatomical differences, electrode impedance variations, and individual cortical activation patterns.

The cross-subject generalization challenge is particularly critical for practical BCI applications, as clinical systems must function reliably across diverse user populations without requiring extensive individual calibration (Chen et al. 2025; Lee and Lee 2022). Current approaches show significant performance drops when transitioning from within-subject to cross-subject evaluation scenarios, highlighting the need for more robust domain adaptation strategies.

To address these challenges, we introduce CAT-Net, [C]ross-[A]ttention [T]one Net, a novel multimodal fusion framework that integrates EEG and EMG signals for the classification of Mandarin tones. Our approach makes several key contributions: (1) We develop a cross-attention mechanism inspired by Transformer architectures (Vaswani et al. 2017) that enables sophisticated bidirectional interaction between EEG and EMG modalities, moving beyond simple concatenation to capture complex neural-muscular coordination patterns; (2) We demonstrate effective tone classification using a minimal channel configuration of only 20 EEG channels and 5 EMG channels, selected through attention-based channel importance analysis; (3) We incorporate domain adaptation techniques (Ganin et al. 2016) to enhance cross-subject generalization capabilities; (4) We provide comprehensive evaluation across both audible and silent speech conditions, demonstrating the system's versatility for different communication scenarios, assisting individuals with hearing or speech impairments.

Related Work

Neural Speech Decoding

EEGNet and DeepConvNet have been widely used as baseline architectures for EEG-based speech decoding tasks due to their lightweight design and strong feature extraction capabilities (Lawhern et al. 2018; Schirrmeister et al. 2017). These models have demonstrated promising results in brain-computer interface (BCI) applications, particularly for assisting individuals with speech impairments (Brumberg et al. 2010). For Mandarin tone classification, Li et al. (Li, Pun, and Chen 2021) achieved 42.9% accuracy using Riemannian manifold features, while Wang et al. (Wang et al. 2023a) improved the accuracy to 68% with end-to-end CNNs. However, EEG-only approaches suffer from limited spatial resolution and high artifact susceptibility, particularly when distinguishing similar tones like Tone 2 and Tone 4.

EMG signals provide complementary articulatory information that directly correlates with speech production (Bharali et al. 2024). Wu et al. achieved 90.76% accuracy on Mandarin phrase classification using parallel Inception CNNs with EMG (Wu et al. 2022). Additionally, Janke and Diener demonstrated that facial surface electromyographic (sEMG) signals can be directly transformed into audible speech, enabling silent speech communication while preserving critical paralinguistic cues (Janke and Diener 2017). These findings highlight the potential of EEG and EMG as viable modalities for decoding speech, especially in scenarios where traditional acoustic signals are unavailable or impaired.

Multimodal Fusion and Attention Mechanisms

Traditional multimodal fusion employs concatenation or late fusion, which fails to capture temporal dependencies and cross-modal interactions during speech production. Saha and Fels (Saha and Fels 2019) demonstrated that attention-based feature selection significantly improves EEG-based imagined speech decoding, achieving 23.45% improvement over baselines through hierarchical attention mechanisms on joint variability matrices. Li et al. (Li et al. 2023) trained sequence-to-sequence decoders that transform the combined EEG- sEMG signals collected during silent speech into audible speech, and resulting audio transcripts showed a mean character-error rate (CER) of 7.22% across eight speakers. However, these approaches either focus on single-modality self-attention mechanisms or employ simple concatenation-based fusion, without capturing the bidirectional interactions between multiple inputs.

Cross-attention mechanisms, inspired by Transformers (Vaswani et al. 2017), have shown success in capturing bidirectional interactions between modalities. Tang et al. (Tang et al. 2025) applied cross-attention to EMG and accelerometer data for tremor classification, outperforming single-modality baselines. Speech-related applications pose unique challenges that demand advanced attention mechanisms to capture neural-muscular interactions. Our cross-attention framework enables EEG and EMG to dynamically attend to each other's key features, going beyond simple concatenation to model the coordination essential for tone classification.

Cross-Subject Generalization

Subject variability remains a key challenge for BCI speech systems. EEG and EMG signal patterns differ highly across individuals due to anatomical, physiological, and electrode placement differences, often causing a model to fail on unseen subjects. General domain adaptation frameworks have shown that treating multiple sources as a unified domain can be suboptimal (Zhu, Zhuang, and Wang 2019). For EEG and EMG specific applications, Zhang et al. apply a conditional domain adversarial network to EMG-based silent speech recognition, while Zhao et al. introduce DRDA, a domain discriminator trained adversarially, to encourage the feature extractor to learn domain-invariant representations, thereby improving generalization to unseen subjects or sessions (Zhang et al. 2023; Zhao et al. 2021). These approaches underscore the importance of learning features that capture task-relevant information while minimizing subject-specific variance. Inspired by this, our CAT-Net integrates

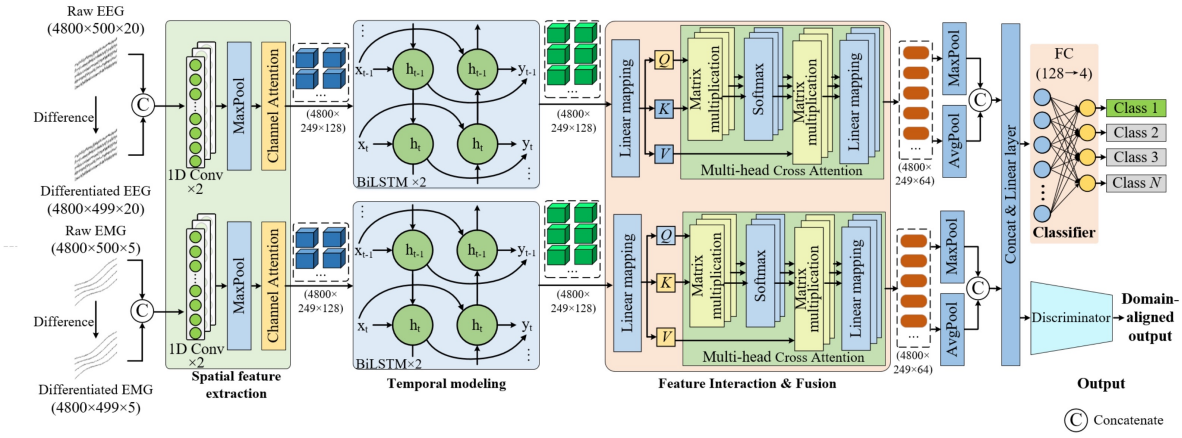

Figure 1: The architecture of our proposed CATNet.

domain-adversarial training into its architecture to explicitly disentangle class-discriminative features from subject-dependent patterns. This enables the model to maintain high performance when deployed to new subjects, particularly enhancing accuracy on abnormal unseen subjects, which is a critical requirement for practical BCI systems.

Methods

To effectively model the neural-muscular coordination in speech, we propose a model that we call CAT-Net, [C]ross-[A]ttention [T]one Net, a three-stage architecture designed for multimodal EEG-EMG fusion (Fig. 1). First, spatial-temporal features are extracted independently from EEG and EMG signals. Second, a cross-attention module, inspired by the Transformer architecture (Vaswani et al. 2017), enables dynamic information exchange between modalities, allowing each to attend to the most informative features of the other. Finally, the fused features serve dual purposes: they are directly fed to a tone classifier to predict the four Mandarin tones, and simultaneously passed through a gradient reversal layer to a domain discriminator that encourages subject-invariant feature learning. In addition, we leverage SHAP (SHapley Additive exPlanations) values to interpret the model's decision process and quantify the relative contributions of EEG and EMG features (see Appendix D).

Signal Inputs

Let $X_i = \{X_i^e, X_i^m\}_{i=1}^n$ denote the paired electrophysiological recordings for trial i, where the superscripts e and m stand for **EEG** and **EMG**, respectively. The labels are $Y_i = \{Y_i^t, Y_i^s\}_{i=1}^n$, with $Y_i^t \in \{1, 2, 3, 4\}$ encoding the Mandarin tone and Y_i^s identifying the speaker. After preprocessing, each trial is represented by the sequence $x = (x_1, \ldots, x_T) \in \mathbb{R}^{2C_e+2C_m}$ with $T = 500$, where C_e and C_m are the numbers of EEG and EMG channels. Every raw channel is augmented with its **first-order temporal difference** $\Delta x_t = x_t - x_{t-1}$, a finite-difference operator that accentuates rapid transients, improves stationarity, and has been shown to boost decoding accuracy in both EEG- and EMG-based BCIs (Andreou and Poli 2016; Phinyomark et al. 2013). Because Δx_t is undefined at $t = 1$, the effective sequence length becomes $T' = T - 1 = 499$; concatenating the original and differential streams explains the factor 2 in the feature dimension $2C_e + 2C_m$.

Spatial and Temporal Encoders

For each of the modalities of EEG and EMG, CAT-Net needs to first encode their spatial and temporal features. For spatial, we use two 1x1 pointwise Conv1D layers. This means we apply a kernel with size of $1 \times 2C_{EEG \text{ or } EMG} \times F$, where F is the feature dimension of size 64 and 128, to each individual timestep t.

$$\mathbf{H}_t = \text{ReLU}(X_t W_{conv}) \in \mathbb{R}^{1 \times F} \tag{1}$$

Intuitively, this means each column of the kernel weight matrix W_{conv} learns a spatial combination of our EEG or EMG channels into a feature in F. A 1D MaxPooling layer with a kernel size of 2 and stride of 2 halves the sequence dimension from 499 to 249, preserving key signal features like neural spikes while reducing computational complexity. To reduce the noise inherent in EEG & EMG data, we adopt the channel attention from CBAM (Woo et al. 2018) to re-weight feature importance. We apply a global average pooling to capture general channel scales and global max pooling to capture spikes across our feature representations. $s_a = \text{GlobalAveragePooling}(\mathbf{H}) \in \mathbb{R}^F$ and $s_m = \text{GlobalMaxPooling}(\mathbf{H}) \in \mathbb{R}^F$. We then pass s_a and s_m to dense layers with sigmoid activation, normalizing the outputs to 0 to 1, $s'_a = \text{sigmoid}(W_2\text{ReLU}(W_1 s_a))$. Similarly, $s'_b = \text{sigmoid}(W_2\text{ReLU}(W_1 s_m))$. To combine these two, we perform a simple sum, $s' = s'_a + s'_b \in \mathbb{R}^F$. The final weights are then broadcasted across the time dimension T.

$$\tilde{\mathbf{H}} = s' \odot \mathbf{H} \in \mathbb{R}^{T \times F} \tag{2}$$

We then use BiLSTM to model temporal dynamics with the same output of 64 features, as it captures long-range dependencies and bidirectional patterns essential for detecting

Algorithm 1: Single training iteration for CAT-Net

Input: EEG mini-batch $\tilde{X}^e$, EMG mini-batch $\tilde{X}^m$, tone labels $\tilde{Y}^t$, subject labels $\tilde{Y}^s$
Parameter: modality-specific encoders $\mathcal{E}_\theta^e$ and $\mathcal{E}_\theta^m$ (same arch, different weights); cross-attention module XAttn_ξ; tone classifier $\mathcal{C}_\phi$; domain discriminator $\mathcal{D}_\psi$; weights λ_{dom}; focal hyper-γ, $\boldsymbol{\alpha}$
Output: losses $\mathcal{L}_{\text{focal}}, \mathcal{L}_{\text{dom}}$

$Z^{(e)} \leftarrow \mathcal{E}_\theta^e(\tilde{X}^e)$
$Z^{(m)} \leftarrow \mathcal{E}_\theta^m(\tilde{X}^m)$
$C^{(e)} \leftarrow \text{XAttn}(Q{=}Z^{(e)},\ K{=}Z^{(m)},\ V{=}Z^{(m)})$
$C^{(m)} \leftarrow \text{XAttn}(Q{=}Z^{(m)},\ K{=}Z^{(e)},\ V{=}Z^{(e)})$
$Z \leftarrow \text{Fuse}\big(C^{(e)},\ C^{(m)}\big)$
$\hat{Y}^t \leftarrow \mathcal{C}_\phi(Z)$
$\hat{Y}^s \leftarrow \mathcal{D}_\psi\big(\text{GRL}_{\lambda_{\text{dom}}}(Z)\big)$
$\mathcal{L}_{\text{focal}} \leftarrow \text{FocalLoss}(\hat{Y}^t, \tilde{Y}^t)$
$\mathcal{L}_{\text{dom}} \leftarrow \text{CrossEntropy}(\hat{Y}^s, \tilde{Y}^s)$
$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda_{\text{dom}}\,\mathcal{L}_{\text{dom}}$
Back-propagate and update θ, ξ, ϕ, ψ with Adam; update each c_k with EMA
return $\mathcal{L}_{\text{focal}},\ \mathcal{L}_{\text{dom}}$

tonal variations, while being more stable and data-efficient than a full Transformer encoder. Due to having both a forward LSTM and backwards LSTM and concatenating their outputs, the final output is of shape $\mathbf{Z} \in \mathbb{R}^{T\times 2F}$.

Cross-Modal Attention Fusion

For the encodings above, we apply the same operations to both EEG and EMG, each has its unique set of weights. Thus, we have inputs $\mathbf{Z}^{EEG}, \mathbf{Z}^{EMG}$ to our cross attention layer. For each of these input we have three weight matrices $W_Q, W_K \in \mathbb{R}^{2F\times K}; W_V \in \mathbb{R}^{2F\times V}$ and corresponding outputs $Q^{(e,m)} = \mathbf{Z}^{(e,m)} W_Q^{(e,m)} \in \mathbf{R}^{T\times K}$. We let EEG query the key and value outputs of EMG and vice versa (Algorithm 1, lines 3-4).

$$\mathbf{C}^{(e,m)} = \text{MHA}(Q^{(e,m)}, W^{(m,e)}, K^{(m,e)}) \tag{3}$$

For each of EEG and EMG, we use a Multi-Head Cross Attention layer with 4 heads and dimension $K = V = 32$. The final output $\mathbf{C}^{(e,m)}$ is of shape $T \times 128$. We apply Layer Normalization to prevent exploding gradients. We then use a global avg pooling and global max pooling to capture activations and spikes across time, $P_{avg} = \text{GlobalAvgPool}(C) \in \mathbb{R}^{1\times 128}$ and concatenate them, $P = \text{Concat}(P_{avg}, P_{max}) \in \mathbb{R}^{1\times 256}$. This passes through a simple Dense layer mapping the 256 to 128, then finally is used in our Tone classifier to predict the four Mandarin tones $y \in \{1, 2, 3, 4\}$.

Domain Discriminator and Loss Functions

To make CAT-Net maintain accuracy on unseen subjects, we attach a **domain-discriminator head** to the fused feature $\mathbf{f} \in \mathbb{R}^{128}$ during training. A *gradient-reversal layer* (Ganin et al. 2016) leaves the forward pass unchanged but flips the sign of the gradient, $\mathcal{R}_\lambda(\mathbf{f}) = \mathbf{f}$, $\frac{\partial \mathcal{R}_\lambda}{\partial \mathbf{f}} = -\lambda \mathbf{I}$ so that the discriminator learns to predict the subject label d while the backbone is forced toward **subject-invariant** representations. The network is trained with three losses (Algorithm 1, lines 8-15):

- **Focal loss** $\mathcal{L}_{\text{focal}}$ (Lin et al. 2017) on the tone logits, with $\gamma = 2$ and class-balancing vector $\boldsymbol{\alpha} = (0.2,\ 0.3,\ 0.2,\ 0.3)$. Our loss weighting strategy considers tone-level confusion patterns: more confused tones, such as tone 2 and tone 4, are emphasized to enhance discrimination, while less confused tones (e.g., tone 1 and tone 3) are still balanced to prevent bias.
- **Domain Discriminator** $\mathcal{L}_{\text{dom}}$ on the GRL-filtered branch, promoting invariance across the ten subjects.

The total objective is a weighted sum which improves leave-one-subject-out accuracy and tightens class clusters.

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + 0.05\,\mathcal{L}_{\text{dom}} + (0.2,\ 0.3,\ 0.2,\ 0.3)\,\mathcal{L}_{\text{cent}} \tag{4}$$

Experiment

Experiment Setup

Dataset Ten healthy native Mandarin-speaking adults (aged 24–35 years) participated in this study. To minimize potential confounding factors, all participants were instructed to abstain from consuming coffee, alcohol, or any other substances that might influence their neurophysiological states for 24 hours before the experimental session. Additionally, participants were required to: (1) maintain adequate sleep duration the night before testing; (2) thoroughly cleanse their scalp to ensure optimal electrode contact; and (3) review and sign informed consent documents. The study protocol was approved by our institutional ethics committee and strictly followed the ethical guidelines outlined in the Declaration of Helsinki (World Medical Association 2013).

Data Collection The experimental paradigm employed four phonologically distinct Mandarin tones, with each tonal category represented by 30 carefully selected Chinese characters that were balanced for lexical frequency and phonological characteristics. Each character were presented four times in a pseudorandomized sequence to prevent habituation effects, with complete counterbalancing of presentation order across all tonal categories (Fig. 2a). During each trial, characters were displayed centrally on the screen for 1.5 seconds, during which participants performed either silent or audible articulation according to the experimental condition, followed by a jittered inter-stimulus interval (1.5-3 seconds) to minimize anticipatory neural responses. The complete stimulus set was presented in both articulation modes, with all participants completing both conditions. Neural activity was recorded using a high-density 64-channel NeuSen W wireless EEG system (Neuracle Technologies, China) synchronized with electromyographic (EMG) recordings from five orofacial muscles (the right buccinator, right cervical trapezius, left buccinator, left cervical trapezius, and mentalis). All physiological signals were acquired at 1000 Hz sampling frequency. The experimental protocol continued until all 480 trials (4 tones × 30 characters × 4 repetitions) were completed in both articulation conditions.

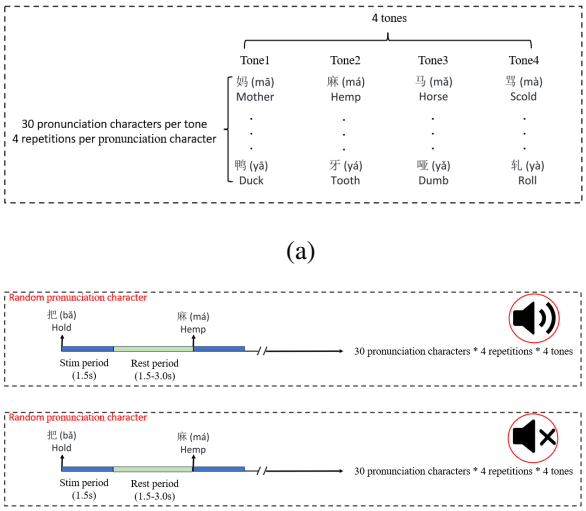

Figure 2: Experimental paradigm. (a) Stimulus design showing four tone categories, with thirty distinct pronunciation characters per tone category. Each pronunciation character was presented four times throughout the experiment. (b) Schematic representation of the experimental procedure, illustrating randomized character presentation in both silent and audible modes.

Preprocessing For EEG signal preprocessing, the raw signals from valid channels were first downsampled to 500 Hz to improve computational efficiency. A fourth-order Butterworth bandpass filter was applied for frequency band extraction, followed by a 50 Hz notch filter to eliminate power line interference. Common average referencing (CAR) was employed to reduce common-mode noise across channels. Subsequently, independent component analysis (ICA) was performed to identify and remove ocular (EOG) and muscular (EMG) artifacts. Epochs were extracted from (-2 s, 2 s) relative to stimulus onset, with baseline correction applied using the pre-stimulus interval (-2 to 0 s) to eliminate baseline drift. For EMG signals preprocessing, the EMG signals were downsampled to 500 Hz to match the EEG sampling rate. A fourth-order Butterworth bandpass filter was applied, followed by a 50 Hz notch filter to remove power line interference. Common noise was reduced using common average referencing (CAR). Data epochs were extracted from (-2 s, 2 s) relative to stimulus onset, and baseline correction was performed to minimize drift.

Channel Selection

To identify the most informative EEG channels contributing to tonal classification and improve training efficacy, we incorporated a Channel-Attention block in the early stage of the EEG branch. Inspired by the channel attention mechanism in CBAM (Woo et al. 2018), we apply a global average and max pooling–based gating module to adaptively re-weight each EEG channel. These weights reflect the relative importance assigned to each channel during the classification process and can be aggregated across samples to identify the most informative electrodes. After training, we visualized the normalized attention weights on a 2D topographic scalp map (Fig. 3). As shown, channels located in the frontal, central, and parietal regions exhibit stronger attention responses, suggesting greater contributions to tone classification. This finding is consistent with prior neurophysiological studies that have highlighted the involvement of the prefrontal and parietal cortices in tone and syllable perception (Ni et al. 2023), while our results further emphasize the potential role of central regions.

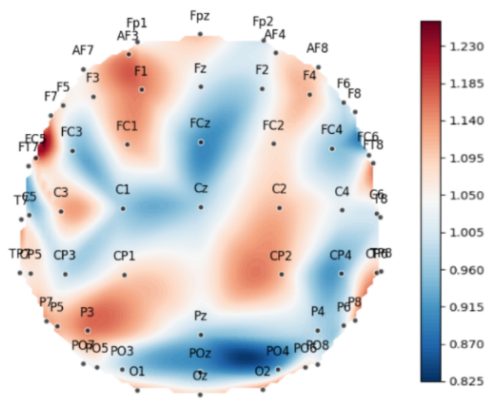

Figure 3: EEG channel weight calculated by Channel-Attention block.

Channel	Tone 1	Tone 2	Tone 3	Tone 4	Average	Kappa
5	99.42	79.60	85.17	82.36	86.92	0.8249
10	98.68	80.80	86.24	81.29	87.02	0.8251
20	98.67	83.64	87.29	83.10	88.08	0.8415
Full	99.12	83.99	88.25	84.67	88.45	0.8496

Table 1: Performance comparisons under different channel numbers(%). We list the ranked weighted channels in the Appendix B. Average denotes the average accuracy over K-fold cross-validation runs.

Based on this channel attention selection, we conduct experiments on the balance of channels and training efficacy, finding the least EEG channel to train without significantly lowering our prediction accuracy. In Table 1, we select the top five, ten, and twenty channels with the highest weights calculated by channel attention and compare them with the full channels' accuracy to investigate the balance between model efficacy and accuracy. Based on the results from Table 1, we observe that training with 20 channels reaches a balance between trimming less informative channels and high accuracy. Thus, subsequent experiments will primarily focus on models utilizing a reduced set of 20 EEG channels in the frontal, central, and parietal brain regions.

Method	Tone 1		Tone 2		Tone 3		Tone 4		Macro Avg	
	F	R	F	R	F	R	F	R	F	R
EEGNet	98.27	98.92	80.14	77.33	85.66	86.58	76.95	78.33	85.29	85.26
DeepConvNet	93.52	85.17	**83.91**	**85.18**	87.53	**91.17**	80.93	80.08	86.56	86.64
GAT	97.25	97.88	75.21	67.15	81.94	89.93	77.87	78.55	83.62	83.60
CATNet	**99.54**	**99.83**	81.45	80.36	**87.91**	87.58	**82.54**	**84.33**	**88.08**	**88.06**

Table 2: Comparison results on each tone's F-1 score and recall with three baseline methods with ranked highest average precisions. "F" and "R" represent "F1-score" and "Recall" respectively. The results are in the silent speech condition. All numbers are in %.

Method	Tone1	Tone2	Tone3	Tone4	Average	Kappa
ETE-CNN	71.15	69.16	46.59	48.18	57.33	0.3725
VLAAI	94.21	39.34	69.84	49.98	61.12	0.5515
EEGNet	98.10	83.15	84.75	75.62	85.29	0.8037
DeepConvNet	99.57	82.69	83.75	81.79	86.56	0.8125
FBCSP+SVM	87.30	41.01	39.04	37.24	51.65	0.3553
GAT	97.72	73.18	**88.56**	74.33	83.62	0.7888
DRDA	**99.71**	83.27	80.22	61.01	81.23	0.7802
EEG-Transformer	92.25	79.86	82.63	69.94	81.10	0.7477
CATNet	98.67	**83.64**	87.27	**83.10**	**88.08**	**0.8415**

Table 3: Tone precision with baseline comparisons in the silent speech condition. ETE-CNN represents End-to-End CNN. Average Precision and Kappa metrics are employed in the table. The best results in each column are highlighted in bold, and the second-best are underlined. The results are under 20 EEG channels and the silent speech condition.

Baseline Comparisons

K-Fold Cross Validation To validate the accuracy of our proposed EEG-EMG fusion model, we compared our algorithms in both silent and audio conditions against eight main state-of-the-art EEG decoding algorithms, including End-to-End CNN (Wang et al. 2023b), VLAAI (Accou et al. 2023), EEGNet (Lawhern et al. 2018), DeepConvNet (Schirrmeister et al. 2017), FBCSP+SVM (Ang et al. 2008; Cortes and Vapnik 1995), GAT (Song et al. 2023), DRDA (Zhao et al. 2021), and EEG-Transformer (Lee and Lee 2021). We reproduced these algorithms on our datasets and compared their performance with our proposed CATNet model. For algorithms that are not utilizing EMG signals, we concatenate EEG and EMG signals as input to the model. We use average precision and kappa values as metrics. To mitigate overfitting and ensure robust performance evaluation, we employed 5-fold cross-validation during model training and testing. For brevity, we present the results for the audio speech condition in Appendix C. The subsequent analysis centers on model performance under the silent speech condition.

As shown in Table 3, our proposed method achieves the highest average accuracy of 88.08% and kappa value of 0.8415 in the silent speech condition. For audio speech, we achieve 87.83% accuracy with kappa value of 0.8377. Both results outperform all baseline models. In particular, CATNet demonstrates superior performance on the challenging Tone 2 and Tone 4, which are traditionally difficult to distinguish due to their subtle frequency contours.

Table 2 presents F1 scores and recall for comprehensive evaluation. The three baselines are selected based on their high average precision in Table 3. CATNet achieves the highest F1 and recall for Tone 1 (F=99.54, R=99.83) and Tone 4 (F=82.54, R=84.33). Compared to DeepConvNet, CATNet improves Tone 4 recall (84.33 vs. 80.08) and achieves stronger macro-averaged performance (F1=88.08, R=88.06). While DeepConvNet shows slightly higher Tone 2 performance, CATNet achieves better precision (83.64 vs. 82.69), indicating fewer false alarms and more conservative predictions beneficial for high-stakes scenarios. The consistent performance across tones confirms that CATNet provides superior discriminative and generalizable representations for tone-level silent speech decoding.

Cross-Subject Calibration In general, EEG and EMG signals in BCI applications exhibit strong cross-subject variability, which poses significant challenges to model generalization on unseen individuals. To address this limitation, we conduct cross-subject evaluations to assess the transferability of our proposed model across different participants. We used a standard LOSO (Leave-One-Subject-Out) training strategy. We withheld data from a single subject as the test set and used the nine remaining subjects' data for training. Table 4 shows the results in silent speech conditions, and audio speech baseline comparisons are in Appendix C.

The results demonstrate our model's strong generalization ability under cross-subject evaluation, with accuracy dropping slightly **2.98%** from 88.08% to 85.10%. In contrast, previously strong-performing baselines exhibit greater deviation in accuracy, indicating a weakness in the model's performance on unseen subjects. For example, EEGNet and DeepConvNet experience accuracy drops of **7.5%** and **6.08%**, respectively. Notably, the accuracy for each subject, except S9 (0.84% lower), outperforms the second-best baseline by an average of **2.88%** across the remaining subjects, further confirming our significant superiority across all subjects. Similarly, our model accuracy dropped from 87.83% to 83.27% in the audio speech condition, while other baseline models also suffer from greater deviation. Furthermore, our model shows strong improvement in abnormal subjects such as S1(**6.8%** higher than the second-best baseline), S3(**6.25%** higher), and S4(**2.5%** higher), where baseline models struggle to generalize the learned features to abnormal individuals. Therefore, it can be shown that our pro-

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
ETE-CNN	38.12	30.00	71.67	36.67	70.21	83.13	49.38	70.00	25.42	54.79	52.94
VLAAI	46.53	50.26	56.72	60.46	65.43	63.56	62.96	55.87	62.38	60.45	58.47
EEGNet	25.83	79.17	75.00	78.96	79.37	81.25	92.92	82.92	95.42	87.08	77.79
DeepConvNet	41.67	73.75	76.46	82.50	84.58	86.88	90.42	85.83	**95.63**	87.08	80.48
FBCSP+SVM	38.37	43.36	48.31	50.24	52.06	57.91	57.49	42.93	55.83	51.02	49.75
GAT	43.75	78.75	71.04	79.79	82.29	90.42	89.58	87.29	92.08	90.42	80.54
DRDA	41.15	79.24	71.36	75.32	82.28	89.17	61.93	84.24	92.54	82.48	75.37
EEG-Transformer	27.50	53.33	63.33	76.25	66.25	63.96	81.87	82.29	86.04	84.38	68.52
CATNet	**53.33**	**79.58**	**82.71**	**85.00**	**87.08**	**91.67**	**95.00**	**89.17**	94.79	**92.71**	**85.10**

Table 4: Cross-subject classification performance on ten subjects. (Silent Speech Condition).

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
$CATNet_{w/o\ DD}$	48.96	80.83	82.92	84.17	87.29	92.50	93.54	89.38	94.79	92.50	84.69
CATNet	**53.33**	79.58	82.71	**85.00**	87.08	91.67	**95.00**	89.17	94.79	**92.71**	85.10

Table 5: Cross-subject ablation results. "DD": Domain Discriminator.

posed method is capable of handling individual differences and generating stable accuracy even with abnormal subjects' data.

Method	**CT**	**TF**	**FF**	**DD**	**P**	**R**	**F1**
$CATNet_{w/o\ ct}$		✓	✓	✓	77.63	77.62	77.00
$CATNet_{w/o\ BiLSTM}$	✓		✓	✓	78.42	78.42	78.41
$CATNet_{w/o\ fusionEMG}$		✓		✓	76.60	76.60	76.50
$CATNet_{w/o\ fusionEEG}$		✓		✓	76.46	76.46	76.35
CATNet	✓	✓	✓	✓	**88.08**	**88.08**	**88.06**

Table 6: Ablation studies in 5-fold training scenario. Legend: "CT": Cross-Attention, "TF": Temporal Feature with BiLSTM layer, "FF": Fusing Feature, "DD": Domain Discriminator. "w/o." is short for "without".

Ablation Study

We conducted ablation experiments to assess the contribution of each module in CATNet. The results of 5-fold validation and cross-subject generalization are presented in Tables 5 and 6, respectively. Our analysis focuses on two core aspects of CATNet: (1) the role of cross-attention in multimodal fusion, and (2) the impact of the domain discriminator (DD) on subject-level generalization.

Cross-Attention Mechanism Cross-attention allows the model to selectively focus on the relevant features between EEG and EMG. To assess its importance, we progressively removed the cross-attention module and its sub-components. As shown in Table 6, removing the entire cross-attention module leads to a substantial drop in precision (from 88.08% to 77.63%). Furthermore, when we isolate the effect of each directional attention (i.e., attention from EEG to EMG and vice versa), we observe comparable degradation, with both variants yielding precision scores around 76–77%. These results confirm that cross-attention is indispensable for capturing inter-modal dependencies to boost recognition accuracy.

Generalization Across Subjects To investigate the contribution of the domain discriminator (DD), we compare CATNet with and without the DD module under the cross-subject setting, as presented in Table 5. Overall, the absence of DD leads to a slight drop in average accuracy (from 85.10% to 84.69%), indicating that the model still maintains robust generalization capability even without domain supervision. However, substantial performance gains are observed on individual subjects with the DD module. In particular, for subject S1, accuracy improves significantly from 48.96% to 53.33%, highlighting DD's effectiveness in handling severe subject-specific variations. Since our dataset contains a relatively balanced distribution across subjects and employs a minimal-channel configuration that limits overfitting, the marginal domain gaps are already well-handled by the core network. Therefore, DD may yield greater benefits in low-resource scenarios or in datasets with greater heterogeneity across subjects or recording conditions.

Conclusion

In the paper, we propose the CATNet method that utilizes a deliberated mechanism, demonstrating the efficacy of a multimodal EEG-EMG fusion framework for tonal speech decoding, achieving high accuracy in both audible and silent speech conditions with minimal-channel configurations. Notably, CATNet maintains strong generalization ability in cross-subject evaluations—a challenging yet crucial setting for real-world BCI applications. These results establish CATNet as a solid baseline for tone-level bio-signal decoding and highlight its potential for low-resource, multimodal speech interfaces. In future work, we aim to generalize our framework to accommodate additional input modalities beyond EEG and EMG, and apply the framework to broader classification tasks. We also aim to validate the generalizability of our architecture across more diverse datasets, further exploring its scalability and potential for real-world deployment in low-resource, multi-modal BCI systems.

Acknowledgments

This work was supported by Shanghai Yang Fan Foundation (No. 24YF2730700).

References

Accou, B.; Vanthornhout, J.; hamme, H. V.; and Francart, T. 2023. Decoding of the Speech Envelope from EEG Using the VLAAI Deep Neural Network. *Scientific Reports*, 13: 812.

Andreou, D.; and Poli, R. 2016. A Feasibility Study of EEG-Based Brain-Computer Interfaces Using Temporal Derivative Features. In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016)*, 210–217.

Ang, K. K.; Chin, Z. Y.; Zhang, H.; and Guan, C. 2008. Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2390–2397.

Barrientos Rojas, S. J.; Ramírez-Valencia, R.; Alonso-Vázquez, D.; Caraza, R.; Martinez, H. R.; Mendoza-Montoya, O.; and Antelis, J. M. 2022. Recognition of grammatical classes of overt speech using electrophysiological signals and machine learning. In *2022 IEEE 4th International Conference on BioInspired Processing (BIP)*, 1–6.

Bharali, A.; et al. 2024. Enabling the Translation of Electromyographic Signals Into Speech. *SN Computer Science*, 5(1094).

Brumberg, J. S.; Nieto-Castanon, A.; Kennedy, P. R.; and Guenther, F. H. 2010. Brain-computer interfaces for speech communication. *Speech communication*, 52(4): 367–379.

Chen, T. Y.-H.; Chen, Y.; Soederhaell, P.; Agrawal, S.; and Shapovalenko, K. 2025. Decoding EEG Speech Perception with Transformers and VAE-based Data Augmentation. arXiv:2501.04359.

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine Learning*, 20(3): 273–297.

Dash, D.; Ferrari, P.; and Wang, J. 2021. Role of Brainwaves in Neural Speech Decoding. In *2020 28th European Signal Processing Conference (EUSIPCO)*, 1357–1361.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59): 1–35.

Janke, M.; and Diener, L. 2017. EMG-to-Speech: Direct Generation of Speech From Facial Electromyographic Signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25: 2375–2385.

Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. EEGNet: A Compact Convolutional Neural Network for EEG-Based Brain–Computer Interfaces. *Journal of Neural Engineering*, 15(5): 056013.

Lee, Y. E.; and Lee, S. 2021. EEG-Transformer: Self-attention from Transformer Architecture for Decoding EEG of Imagined Speech. *CoRR*, abs/2112.09239.

Lee, Y.-E.; and Lee, S.-H. 2022. EEG-Transformer: Self-attention from Transformer Architecture for Decoding EEG of Imagined Speech. In *2022 10th International Winter Conference on Brain-Computer Interface (BCI)*, 1–4.

Li, H.; Wang, M.; Gao, H.; Zhao, S.; Li, G.; and Wang, Y. 2023. Hybrid Silent Speech Interface Through Fusion of Electroencephalography and Electromyography. In *Proceedings of Interspeech 2023*, 1184–1188.

Li, M.; Pun, S. H.; and Chen, F. 2021. Mandarin Tone Classification in Spoken Speech with EEG Signals. In *2021 29th European Signal Processing Conference (EUSIPCO)*, 1163–1166.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 4768–4777.

Ni, G.; Xu, Z.; Bai, Y.; Zheng, Q.; Zhao, R.; Wu, Y.; and Ming, D. 2023. EEG-based assessment of temporal fine structure and envelope effect in Mandarin syllable and tone perception. *Cerebral Cortex*, 33(23): 11287–11299.

Phinyomark, A.; Quaine, F.; Charbonnier, S.; Serviere, C.; Tarpin-Bernard, F.; and Laurillau, Y. 2013. EMG Feature Evaluation for Improving Myoelectric Pattern Recognition Robustness. *Expert Systems with Applications*, 40(12): 4832–4840.

Saha, P.; and Fels, S. 2019. Hierarchical Deep Feature Learning for Decoding Imagined Speech from EEG. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 10019–10020. AAAI Press.

Schirrmeister, R. T.; Springenberg, J. T.; Fiederer, L. D. J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; and Ball, T. 2017. Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization. In *Human Brain Mapping*, volume 38, 5391–5420.

Sharon, R. A.; Narayanan, S. S.; Sur, M.; and Murthy, A. H. 2020. Neural Speech Decoding During Audition, Imagination and Production. *IEEE Access*, 8: 149714–149729.

Song, Y.; Zheng, Q.; Wang, Q.; Gao, X.; and Heng, P.-A. 2023. Global Adaptive Transformer for Cross-Subject Enhanced EEG Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31: 2767–2777.

Tang, C.; Gao, S.; Li, C.; Yi, W.; Jin, Y.; Zhai, X.; Lei, S.; Meng, H.; Zhang, Z.; Xu, M.; Wang, S.; Chen, X.; Wang, C.; Yang, H.; Wang, N.; Wang, W.; Cao, J.; Feng, X.; Smielewski, P.; Pan, Y.; Song, W.; Birchall, M.; and Occhipinti, L. G. 2025. Wearable intelligent throat enables natural speech in stroke patients with dysarthria. arXiv:2411.18266.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, X.; Li, M.; Li, H.; Pun, S. H.; and Chen, F. 2023a. Cross-Subject Classification of Spoken Mandarin Vowels and Tones with EEG Signals: A Study of End-to-End CNN with Fine-Tuning. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 535–539.

Wang, X.; Li, M.; Li, H.; Pun, S. H.; and Chen, F. 2023b. Cross-Subject Classification of Spoken Mandarin Vowels and Tones with EEG Signals: A Study of End-to-End CNN with Fine-Tuning. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 535–539.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19.

World Medical Association. 2013. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA*, 310(20): 2191–2194.

Wu, J.; Zhang, Y.; Xie, L.; Yan, Y.; Zhang, X.; Liu, S.; An, X.; Yin, E.; and Ming, D. 2022. A novel silent speech recognition approach based on parallel inception convolutional neural network and Mel frequency spectral coefficient. *Frontiers in Neurorobotics*, 16: 971446.

Zhang, J.; Xu, Q.-T.; Ling, Z.-H.; and Li, H. 2024. Sparsity-Driven EEG Channel Selection for Brain-Assisted Speech Enhancement. arXiv:2311.13436.

Zhang, Y.; Cai, H.; Wu, J.; Xie, L.; Xu, M.; Ming, D.; Yan, Y.; and Yin, E. 2023. EMG-Based Cross-Subject Silent Speech Recognition Using Conditional Domain Adversarial Network. *IEEE Transactions on Cognitive and Developmental Systems*, PP: 1–1.

Zhao, H.; Zheng, Q.; Ma, K.; Li, H.; and Zheng, Y. 2021. Deep Representation-Based Domain Adaptation for Nonstationary EEG Classification. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2): 535–545.

Zhu, Y.; Zhuang, F.; and Wang, D. 2019. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press. ISBN 978-1-57735-809-1.

Appendix

A. Hyperparameter Settings

Hyperparameter	Value
Optimizer	Adam
Learning Rate	1e-3
Loss Function	Focal Loss (γ=2.0, α=(0.2, 0.3, 0.2, 0.3))
Loss Weights	class: 1.0, domain: 0.05
Dropout Rate	0.4
Batch Size	64
Epochs	50
Cross Validation	5-Fold
LSTM Hidden Units	128
Early Stopping Patience	10
LR Scheduler	ReduceLROnPlateau (factor=0.5, patience=5, min_lr=1e-5)
EEG/EMG Channels Used	20 and 5

Table A1: Hyperparameter settings used for all experiments.

B. Experiment Details

EEG and EMG Channel Placements

We collected EEG and EMG data from participants, and the corresponding electrode placement is illustrated in the figure below.

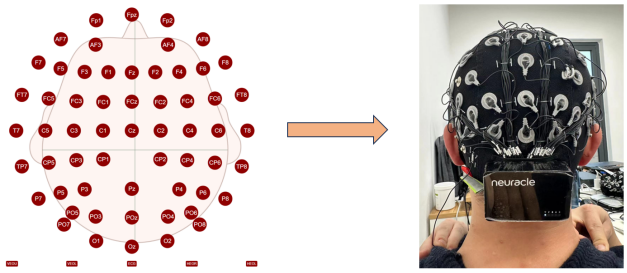

Figure B1: Spatial distribution of EEG electrode placement across the scalp.

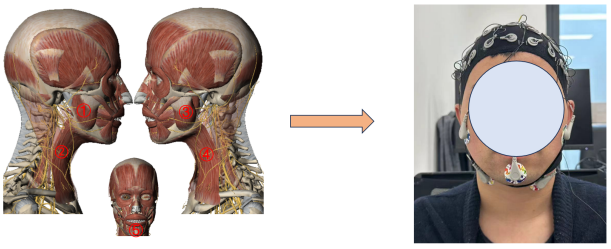

Figure B2: Locations of EMG sensor placement for muscle activity recording. (1. Right Buccinator. 2. Right Cervical Trapezius. 3. Left Buccinator. 4. Left Cervical Trapezius. 5. Mentalis.)

Visualization of Temporal Features

To better understand how to structure the model, we extract and visualize the temporal features of EEG and EMG. For brevity, we present one EEG channel and one EMG channel and their temporal features under both audio and silent speech conditions. From the graph, it can be observed that Tone 2 and Tone 4 exhibit similar EEG and EMG temporal patterns, which explains why our model and all baseline models struggle to differentiate these two tones.

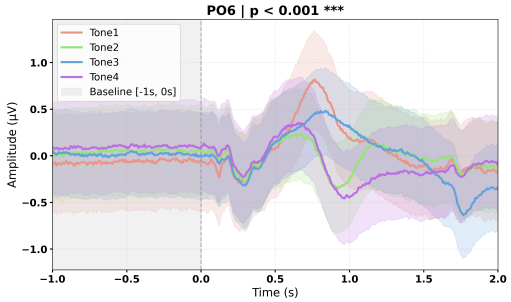

Figure B3: EEG channel PO6 temporal features in audio speech condition.

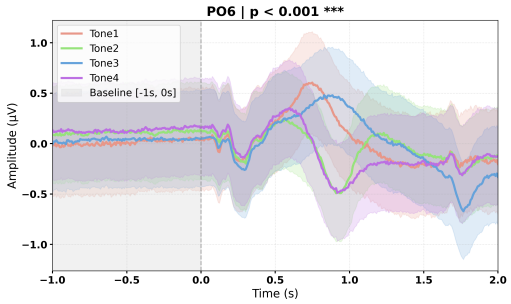

Figure B4: EEG channel PO6 temporal features in silent speech condition.

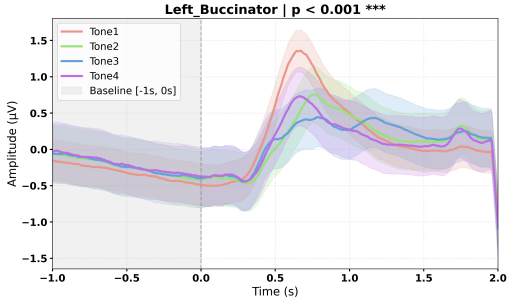

Figure B5: EMG channel Left Buccinator temporal features in audio speech condition.

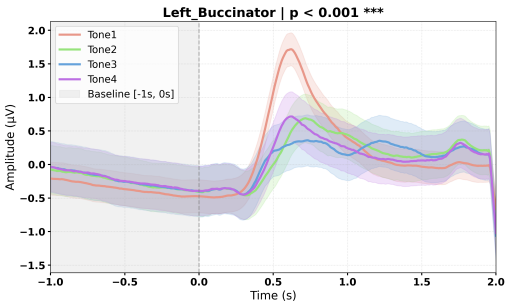

Figure B6: EMG channel Left Buccinator temporal features in silent speech condition.

Channel Weights Based On Channel Attention

Channel	Weight	Channel	Weight	Channel	Weight	Channel	Weight
AF7	1.1566	CP4	1.0700	FC5	1.0229	P3	0.9864
P03	1.1390	FT8	1.0648	T8	1.0223	FC2	0.9815
P06	1.1390	CP3	1.0555	P4	1.0199	FC3	0.9814
F7	1.1242	P07	1.0532	POz	1.0094	AF4	0.9760
FC6	1.1193	C3	1.0531	P7	1.0077	F2	0.9607
Fp1	1.1134	P05	1.0514	TP8	1.0077	PO8	0.9602
CP6	1.1100	F6	1.0487	C2	1.0058	Fz	0.9541
FC4	1.1097	P8	1.0430	CP1	1.0056	Oz	0.9507
F1	1.0793	C4	1.0401	FCz	1.0022	CP2	0.9477
PO4	1.0772	AF8	1.0343	Fp2	0.9944	TP2	0.9474
O1	1.0745	AF3	1.0329	T7	0.9943	F8	0.9404
O2	1.0724	Cz	1.0303	Pz	0.9910		
FT7	1.0718	P6	1.0251	F4	0.9878		

Table B1: EEG channel importance weights sorted in descending order. The table is viewed from left to right.

C. Baseline Comparison in Audio Speech Condition

Baseline Models

End-to-End CNN (Wang et al. 2023b) A convolutional neural network that directly maps raw EEG signals to class labels through stacked 1D convolutional layers and global average pooling.

VLAAI (Accou et al. 2023) A modular architecture composed of repeated blocks integrating convolutional layers, layer normalization, and GRUs, designed for decoding EEG sequences.

EEGNet (Lawhern et al. 2018) A compact convolutional neural network that combines temporal convolution, depthwise spatial filtering, and separable convolutions for efficient EEG feature extraction.

DeepConvNet (Schirrmeister et al. 2017) A deep network with four convolutional-max pooling blocks designed to capture low- to high-level EEG signal features progressively.

FBCSP+SVM (Ang et al. 2008; Cortes and Vapnik 1995) A traditional pipeline using filter bank common spatial patterns (FBCSP) to extract band-power features followed by a support vector machine classifier.

GAT (Song et al. 2023) A transformer-based network utilizing multi-head self-attention and cross-channel temporal fusiony.

Method	Tone 1		Tone 2		Tone 3		Tone 4		Macro Avg	
	F	R	F	R	F	R	F	R	F	R
EEGNet	98.59	99.08	75.56	75.25	86.46	85.92	74.51	74.92	83.78	83.79
DeepConvNet	96.42	**99.92**	79.48	76.17	87.93	87.08	80.30	81.50	86.03	86.17
GAT	97.51	98.00	68.72	62.04	76.92	83.33	73.58	75.25	79.54	79.57
CATNet	**99.50**	99.67	**81.03**	**80.08**	**88.76**	**89.17**	**81.97**	**82.42**	**87.82**	**87.83**

Table C1: Comparison results on each tone's f-1 score and recall with three baseline methods with ranked highest average precisions. "F" and "R" represent "F1-score" and "Recall" respectively. The results are in audio speech condition. All number are in %.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
ETE-CNN	25.00	50.62	48.54	59.38	25.00	46.80	25.00	25.00	49.58	69.37	42.44
VLAAI	51.25	55.23	55.67	54.80	61.73	62.11	64.23	72.14	72.32	64.22	61.34
EEGNet	55.00	68.75	67.92	70.21	80.83	79.17	89.17	83.54	94.79	81.46	77.08
DeepConvNet	49.79	58.33	72.29	80.63	83.75	80.83	85.21	79.79	87.71	85.21	76.35
FBCSP+SVM	33.95	41.04	46.87	43.33	47.50	54.16	53.95	48.95	56.66	53.95	48.04
GAT	51.67	64.79	67.29	71.88	85.00	85.21	85.00	84.17	86.88	**90.00**	77.18
DRDA	48.71	56.00	72.94	76.12	78.14	76.09	82.94	79.23	87.71	84.23	74.24
EEG-Transformer	49.79	47.92	57.29	74.79	76.04	80.21	79.58	67.08	55.21	58.54	64.65
CATNet	**56.04**	**72.08**	**75.42**	**80.63**	**92.08**	**90.00**	**91.87**	**90.62**	**94.79**	89.17	**83.27**

Table C2: Cross-subject classification performance on ten subjects.

DRDA (Zhao et al. 2021) A domain-adversarial architecture combining a shared feature extractor with a domain discriminator and gradient reversal layer.

EEG-Transformer (Lee and Lee 2021) A transformer based network with self-attention applied to EEG signals, enabling efficient modeling of long-range temporal and spatial dependencies for brain-computer interface tasks.

K-Fold Results

In this subsection, we run all the baselines and our proposed CATNet under the audio speech condition. The table is shown below in Table C3. As shown in Table C3, our proposed CATNet achieves the best performance across most tones and metrics, with an average tone precision of 87.83% and a Cohen's Kappa of 0.8377, surpassing all baselines. Notably, CATNet achieves best precision on two out of four tones, second-best precision on Tone 3, and maintains competitive results on Tone 2. These results demonstrate the effectiveness of our EEG-EMG architecture under audible speech conditions.

We then choose three baseline methods with best accuracy and investigate their F-1 score and Recall in Table C1. As presented in Table C1, CATNet achieves the highest F1-score and recall across all four tones in the audio speech condition, except for Tone 1 with slight less recall score. Compared to EEGNet (83.78% / 83.79%) and GAT (79.54% / 79.57%), CATNet shows a significant performance gain, indicating its superiority in tone-level classification.

Method	Tone1	Tone2	Tone3	Tone4	Average	Kappa
ETE-CNN	83.58	35.53	30.69	60.23	42.44	0.2325
VLAAI	94.42	54.64	64.28	56.62	66.81	0.5578
EEGNet	98.10	75.88	87.00	74.11	83.79	0.7837
DeepConvNet	93.16	<u>83.09</u>	**88.79**	<u>79.13</u>	<u>86.17</u>	<u>0.8154</u>
FBCSP+SVM	87.23	43.61	41.02	39.27	52.63	0.3681
GAT	97.10	82.51	71.95	73.53	79.57	0.7270
DRDA	<u>99.21</u>	**91.45**	61.24	58.94	77.72	0.7025
EEG-Transformer	97.15	64.93	73.45	66.48	75.69	0.6758
CATNet	**99.34**	82.00	<u>88.36</u>	**81.53**	**87.83**	**0.8377**

Table C3: Tone precision with baseline comparisons in audio speech condition. ETE-CNN represents End-to-End CNN. Average Precision and Kappa metrics are shown in the table. The best results in each column are highlighted in bold, and the second-best are underlined. The results are using 20 EEG channels.

CATNet's also shows great stability on the most challenging tone pairs. For Tone 2, CATNet achieves an F1 score of 81.03%. Similarly, for Tone 4, CATNet delivers the highest F1 score of 81.97% and recall of 82.42%, demonstrating its effectiveness in handling acoustically similar tonal variations that have historically posed difficulties for existing approaches.

In contrast, while models like GAT and DeepConvNet perform competitively on some tones, they fall short in generalization—especially on tones with subtle acoustic differences. These results collectively showcase CATNet's effectiveness in capturing both phonetic clarity and tonal variation, setting a new benchmark for EEG-based tone classifi-

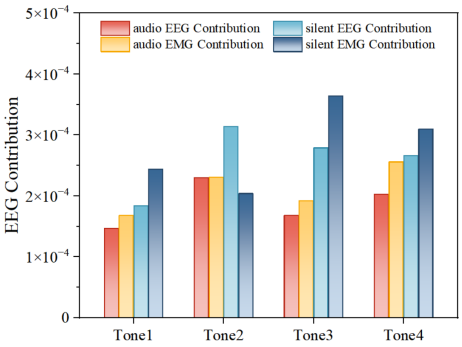

Figure D1: SHAP Value Interpreted on each tone.

cation under audio speech conditions.

Cross Subject Results

As shown in Table C2, CATNet achieves the highest average accuracy of 83.27% in the audio speech condition across all ten subjects, outperforming all baseline models by a significant margin. The second-best models, GAT (77.18%) and EEGNet (77.08%), lag behind by over 6 percentage points, highlighting CATNet's superior generalization ability across individuals.

Moreover, CATNet demonstrates remarkable stability and consistency, achieving the highest accuracy on all 10 subjects. Notably, for subjects S5–S9, all accuracies exceed 90%, while other baselines fails to maintain high accuracies on unseen subjects, indicating robust and scalable performance across varied participants.

D. Model Interpretability with SHAP Values

To interpret the decision-making process of our model, we employ SHAP (SHapley Additive exPlanations) values, a unified framework for explaining model predictions based on cooperative game theory (Lundberg and Lee 2017). Our analysis focuses on two key questions: (1) How does the model shift its reliance between audible and silent speech conditions? and (2) What is the relative importance of EEG and EMG modalities?

As shown in Fig. D1, EMG features generally contribute more than EEG features across both speech conditions, particularly in Tone 1, Tone 3, and Tone 4. This dominance suggests that muscle activity plays a primary role in tone classification, regardless of whether speech is audible or silent. Interestingly, Tone 2 presents an exception: EEG features contribute more than EMG features in the silent condition, highlighting the model's increased reliance on neural signals when muscular cues are insufficient or ambiguous.

These results demonstrate that our model dynamically adjusts its modality weighting based on both speech condition and tone-specific difficulty. While EMG captures more discriminative cues in relatively easier tones (e.g., Tone 1 and Tone 3), EEG serves as a compensatory modality for more challenging tones like Tone 2, especially under silent speech scenarios.