RC-MEMNET: A REGIONAL CONTEXT MEMORY NETWORK FOR CROSS-REGIONAL WU VOWEL RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-regional Wu vowel recognition aims to identify shared vowel categories in unseen regions. Existing methods often suffer from incomplete intra-regional feature representation, insufficient inter-regional dependency modeling, and the scarcity of fine-grained cross-regional Wu vowel resources. To address these issues, we propose a Regional Context Memory Network, termed RC-MemNet, which consists of a Regional Prompt Learning (RPL) module and a Region-Guided Memory Adapter (RGMA) module. Specifically, in the RPL module, learnable regional prompts are introduced to encode region-specific acoustic characteristics through a shared encoder, providing a more complete representation of intra-regional features. Additionally, in the RGMA module, a persistent memory mechanism is introduced to propagate acoustic knowledge across source regions, supporting effective modeling of inter-regional dependencies. To increase the diversity of fine-grained cross-regional Wu vowel resources, we construct WuVowelSet, which covers 45 survey locations across 14 Wu dialect regions. Comprehensive experiments are conducted on our presented WuVowelSet dataset. The experimental results demonstrate that RC-MemNet achieves state-of-the-art performance on WuVowelSet, which is conducive to advancing low-resource cross-regional speech recognition.

1 INTRODUCTION

Cross-regional Wu vowel recognition is a branch of phoneme recognition Glocker et al. (2023); Xu et al. (2022); Lee et al. (2025) that aims to identify shared vowel categories in unseen regions. It provides a fundamental step toward understanding and processing low-resource Wu varieties. Wang et al. (2026); Yue et al. (2025); Yue & Hu (2018). The key challenge is to characterize regional acoustic variation across regions while maintaining category discriminability.

Existing methods generally fall into two categories: phonological knowledge-driven methods and data-driven methods. Phonological knowledge-driven methods use acoustic-phonetic features, phoneme inventories, and articulatory attributes, which provide interpretable linguistic priors Li et al. (2020); Glocker et al. (2023). Data-driven methods learn phoneme representations directly from speech signals Baevski et al. (2020); Hsu et al. (2021a); Chen et al. (2022); Radford et al. (2023) and capture complex acoustic variation across speech conditions, constituting a major paradigm in current research Shi et al. (2024); Boito et al. (2024); Yadav et al. (2024); Ma et al. (2025).

Data-driven methods can be broadly divided into general discriminative methods and condition-adaptive methods. While general discriminative methods learn a shared mapping from acoustic observations to vowel categories, their representations show low sensitivity to region-specific acoustic variation Hsu et al. (2021b); Sun et al. (2018). In contrast, condition-adaptive methods condition phonetic representations with auxiliary information, enabling more comprehensive modeling of such variation Li et al. (2018); Yoo et al. (2019); Winata et al. (2020); Kashiwagi et al. (2024). Among condition-adaptive methods, prompt-based methods encode condition information in learnable prompts to modulate pretrained speech representations, which have become mainstream Chang et al. (2024); Li et al. (2024b); Kashiwagi et al. (2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recently, prompt-based methods have demonstrated strong performance. MAS-LoRA Bagat et al. (2025) employs a mixture of accent-specific LoRA experts to adapt a pretrained speech model to different accents. However, this adaptation paradigm represents each region with a shared condition, obscuring fine-grained intra-regional differences among vowel realizations. This is one limitation. WhisAID Wang & Wang (2026) extracts accent representations from accented speech using a pretrained Whisper encoder and an accent-identification objective. Although such methods characterize each region holistically, they leave corresponding vowel realizations disconnected across regions, weakening inter-regional acoustic dependency modeling. This is another limitation. Furthermore, existing fine-grained cross-regional Wu vowel resources remain insufficient for characterizing vowel variation across Wu regions Wang et al. (2026); Yue et al. (2025); Yue & Hu (2018). This is also a limitation.

To address these limitations, we propose a Regional Context Memory Network (RC-MemNet), which consists of a Regional Prompt Learning (RPL) module and a Region-Guided Memory Adapter (RGMA) module. In the RPL module, regional prompts interact with acoustic features to capture fine-grained intra-regional differences among vowel realizations. In the RGMA module, a shared persistent memory connects category-related acoustic knowledge across source regions, strengthening inter-regional dependency modeling. To enrich fine-grained cross-regional Wu vowel resources, we also construct WuVowelSet, covering 45 survey locations across 14 Wu dialect regions. The main contributions are summarized as follows:

- We propose a Regional Context Memory Network, namely RC-MemNet, for cross-regional Wu vowel recognition. The RPL module is presented to improve intra-regional feature representation through interactions between regional prompts and acoustic features, enabling the model to capture fine-grained variations among vowel realizations within each region.
- To strengthen acoustic dependency modeling across source regions, we propose RGMA with category-aware persistent memory shared across source regions. It consolidates transferable acoustic knowledge among regional realizations of shared vowel categories.
- We construct WuVowelSet, a fine-grained Wu vowel dataset containing vowel annotations from 127 native speakers across 45 survey locations and 14 Wu dialect regions. Comprehensive experiments demonstrate that RC-MemNet achieves state-of-the-art performance on WuVowelSet.

2 Related Work

2.1 Cross-Regional Speech Recognition

Cross-regional speech recognition studies how speech knowledge learned from source regions generalizes to unseen regions. Related research has addressed pronunciation variation across languages, dialects, and accents through shared phonetic representations and condition-specific adaptation.

Cross-lingual phoneme models exploit phonetic structures shared across languages. Allosaurus Li et al. (2020) uses a universal phone inventory with language-specific phoneme mappings, while Allophant Glocker et al. (2023) incorporates articulatory attributes into cross-lingual phoneme recognition. For dialect and accent variation, previous studies have explored multi-dialect modeling, adaptive acoustic modeling, and cross-accent adaptation Li et al. (2018); Yoo et al. (2019); Winata et al. (2020). Recent approaches further introduce language prompts, accent-specific LoRA experts, and accent-aware representations for efficient adaptation and robust ASR Kashiwagi et al. (2024); Bagat et al. (2025); Wang & Wang (2026).

Overall, existing methods promote generalization through shared phonetic structures or condition-aware adaptation. However, they mainly characterize pronunciation variation using global condition representations or condition-specific parameters. Fine-grained acoustic variation within regions and explicit dependencies among regional realizations of shared vowel categories remain insufficiently modeled for cross-regional Wu vowel recognition.

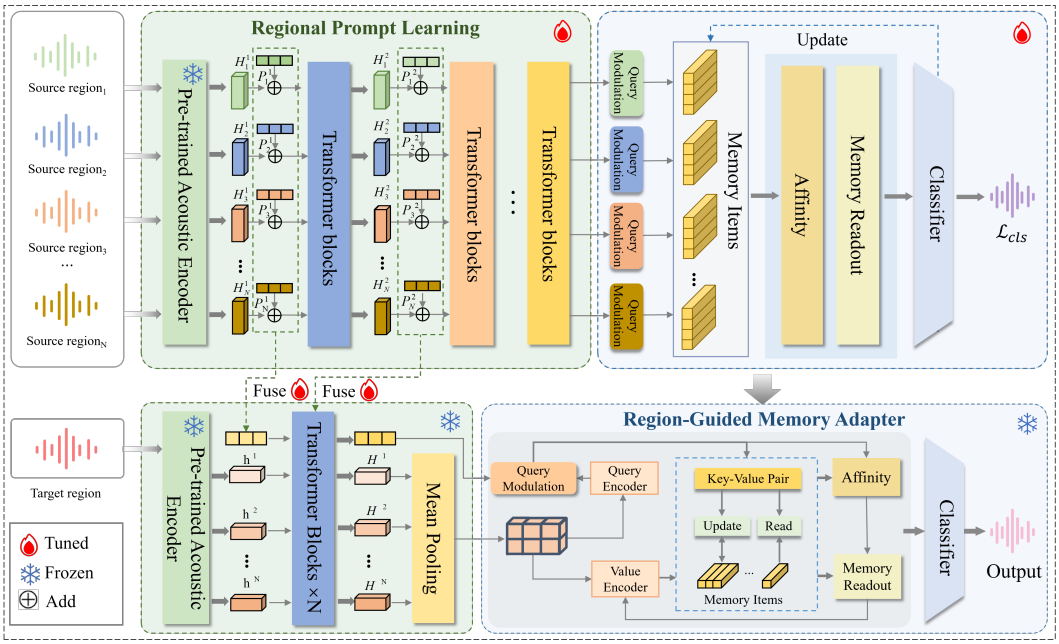

Figure 1: Overall architecture of RC-MemNet. The sample-adaptive regional context c_i^r guides regional representation learning and persistent-memory retrieval. During source-region learning, the memory follows a region-level read-before-write update.

2.2 Memory-Augmented Speech Modeling

Memory-augmented speech models enhance acoustic representations by retaining long-range context or retrieving relevant evidence from stored examples. Existing approaches mainly employ sequence memories, explicit memory banks, or non-parametric datastores.

Sequence-memory models such as FSMN and DFSMN Zhang et al. (2015; 2018) introduce learnable memory structures for long-range acoustic dependencies, while subsequent methods incorporate persistent or explicit memory into attention-based and streaming ASR You et al. (2020); Wu et al. (2020); Shi et al. (2021); Carvalho & Abad (2023). Another line of work retrieves acoustic evidence from non-parametric datastores, as in kNN-CTC and speaker-aware kNN adaptation Zhou et al. (2024); Li et al. (2024a).

Overall, existing memory-based methods primarily organize information along temporal context or retrieve instance-level evidence according to representation similarity. They are not explicitly structured around phonetic categories shared across source regions Zhao et al. (2021). Consequently, they provide limited support for consolidating category-level acoustic knowledge across regions and modeling dependencies among regional realizations of shared vowel categories.

3 Method

The overall framework of RC-MemNet is illustrated in Fig. 1. It mainly consists of the RPL module and the RGMA module. RPL introduces layer-specific regional prompts into the shared acoustic encoder to model fine-grained intra-regional acoustic variation. RGMA uses the resulting regional context to access a category-aware persistent memory shared across source regions, strengthening inter-regional acoustic dependency modeling. For an unseen region, RC-MemNet adapts the composition scores of the source-region prompts while keeping the persistent memory read-only. The regional and memory representations are finally fused for vowel classification.

3.1 REGIONAL PROMPT LEARNING

To improve intra-regional representation, RPL associates each source region with a learnable regional prompt. The prompt is injected into each encoder layer to interact with acoustic features and encode region-specific characteristics.

Let x_i^r denote the i-th input from source region $r \in \{1, \ldots, R\}$, where R is the number of source regions. The acoustic front end Φ produces the initial acoustic tokens $H_i^{r,0} = \Phi(x_i^r)$. For each source region r, RPL learns a regional prompt $P_r \in \mathbb{R}^{S_p \times d_p}$, where S_p is the number of prompt slots and d_p is the prompt dimension.

At the l-th encoder layer, the corresponding layer-specific prompt is generated as

$$T_r^l = \mathrm{LN}\left(\mathrm{Proj}_p^l(P_r)\right), \tag{1}$$

where $\mathrm{Proj}_p^l(\cdot)$ applies a slot-wise prompt projection into the hidden space of the l-th layer, and $\mathrm{LN}(\cdot)$ applies layer normalization along the feature dimension.

The projected prompt tokens are concatenated with the current acoustic tokens and jointly processed by the corresponding encoder layer:

$$[\widehat{T}_i^{r,l}; H_i^{r,l}] = E^l\left([T_r^l; H_i^{r,l-1}]\right), \quad l = 1, \ldots, L, \tag{2}$$

where E^l is the l-th encoder layer, L is the number of encoder layers, and $[\,;\,]$ denotes token-wise concatenation. $H_i^{r,l}$ and $\widehat{T}_i^{r,l}$ are the acoustic-token and prompt-token outputs, respectively.

After the layer-wise prompt–acoustic interaction, the acoustic representation and regional context are obtained as $h_i^r = \mathrm{Pool}(H_i^{r,L})$ and $c_i^r = \mathrm{LN}(\mathrm{Pool}(\widehat{T}_i^{r,L}))$, respectively. Since $\widehat{T}_i^{r,L}$ is jointly produced with the acoustic tokens of the current input, c_i^r varies across vowel segments and provides a sample-adaptive regional context. They are fused as

$$u_i^r = F_p\left([h_i^r \| c_i^r]\right), \tag{3}$$

where $\mathrm{Pool}(\cdot)$ aggregates the token sequence, F_p is the regional feature fusion layer, and $[\cdot\|\cdot]$ denotes feature-wise concatenation. The regional representation u_i^r is subsequently fused with the memory representation retrieved by RGMA for vowel classification, while c_i^r modulates memory access.

For an unseen region t, RPL constructs its regional prompt by adaptively combining the source-region prompts:

$$P_t = \sum_{r=1}^{R} \frac{\exp(a_{t,r})}{\sum_{s=1}^{R} \exp(a_{t,s})} P_r, \tag{4}$$

where $a_{t,r}$ is the composition score associated with source region r. The composition scores are shared across prompt slots and layers, and the composed prompt follows the same projection and interaction process as source prompts.

3.2 REGION-GUIDED MEMORY ADAPTER

To model inter-regional dependencies, RGMA introduces a category-aware persistent memory shared across source regions. The regional context guides memory access, while retrieved memory provides acoustic knowledge accumulated across regions. During source-region learning, RGMA first retrieves acoustic knowledge from the existing cross-regional memory state and then incorporates current-region acoustic evidence through category-restricted sparse assignment.

Let $y_i^r \in \{1, \ldots, C\}$ denote the vowel-category label of x_i^r, where C is the number of shared vowel categories. The persistent memory is represented as $\mathcal{M} = (K, V)$, where $K, V \in \mathbb{R}^{CJ \times d_m}$. Here, J is the number of memory items assigned to each category and d_m is the memory dimension. The memory keys K define memory addresses, while the memory values V store acoustic evidence aggregated across source regions. Let $v_{c,j}$ denote the value of the j-th memory item in category c. Multiple memory items within each category accommodate diverse regional realizations of the same vowel.

RGMA first projects the acoustic representation into the memory space as $q^r_{i,0} = W_q h^r_i + b_q$. The regional context then modulates this base query:

$$q^r_i = \mathrm{LN}\left([1 + \gamma(c^r_i)] \odot q^r_{i,0} + \beta(c^r_i)\right), \tag{5}$$

where W_q and b_q are the query-projection parameters, $\gamma(\cdot)$ and $\beta(\cdot)$ generate feature-wise scaling and shifting terms, and $\odot$ denotes element-wise multiplication.

Let $\bar{q}^r_i$ denote the ℓ_2-normalized query and $\bar{K}$ the row-wise ℓ_2-normalized memory keys. The memory representation is retrieved as

$$m^r_i = \mathrm{softmax}\left(\frac{\bar{q}^r_i \bar{K}^\top}{\tau_a}\right) V, \tag{6}$$

where τ_a is the addressing temperature. This label-free read operation covers all CJ memory items.

After memory reading, the label $y^r_i = c$ restricts writing to the memory items assigned to category c. The sparse assignment weights are computed as

$$\boldsymbol{\alpha}^r_{i,c} = \mathrm{TopKMask}_{K_w}\left[\mathrm{softmax}\left(\frac{\bar{q}^r_i \bar{K}^\top_c}{\tau_w}\right)\right], \tag{7}$$

where $\bar{K}_c$ contains the normalized memory keys of category c, τ_w is the write temperature, and K_w is the number of retained memory items, with $1 \le K_w \le J$. $\mathrm{TopKMask}_{K_w}(\cdot)$ retains the largest K_w weights and sets the others to zero without renormalization.

The modulated query determines the memory address, while the base query preserves the acoustic content of the write feature, $z^r_i = \mathrm{sg}[\mathrm{Norm}(q^r_{i,0})]$, where $\mathrm{sg}(\cdot)$ denotes stop-gradient and $\mathrm{Norm}(\mathbf{x}) = \mathbf{x}/(\|\mathbf{x}\|_2 + \epsilon)$. Here, $\epsilon > 0$ ensures numerical stability. Let $\mathcal{I}_{r,c}$ denote the indices of category-c samples involved in the current region-level update. The evidence assigned to the j-th memory item is aggregated as

$$\widehat{z}_{r,c,j} = \frac{\sum_{i \in \mathcal{I}_{r,c}} \alpha^r_{i,c,j} z^r_i}{\sum_{i \in \mathcal{I}_{r,c}} \alpha^r_{i,c,j} + \epsilon}, \tag{8}$$

where $\alpha^r_{i,c,j}$ is the j-th component of $\boldsymbol{\alpha}^r_{i,c}$. Its write strength is $\omega_{r,c,j} = \sum_{i \in \mathcal{I}_{r,c}} \alpha^r_{i,c,j}/(|\mathcal{I}_{r,c}| + \epsilon)$.

For each memory item, RGMA maintains a cumulative feature state $S_{c,j}$ and a cumulative write state $N_{c,j}$. The feature state is updated as

$$S_{c,j} \leftarrow S_{c,j} + \omega_{r,c,j} \widehat{z}_{r,c,j}. \tag{9}$$

The write state follows $N_{c,j} \leftarrow N_{c,j} + \omega_{r,c,j}$. When $N_{c,j} > 0$, the memory value is $v_{c,j} = \mathrm{Norm}(S_{c,j})$; otherwise, it retains its previous value. The memory keys are optimized through the classification loss, while the memory values follow this explicit region-level update. RGMA combines label-free global reading with category-restricted sparse writing, enabling shared memory retrieval while preserving category-specific regional evidence. The cumulative update consolidates acoustic evidence across source-region updates, producing a shared persistent memory for unseen-region inference.

3.3 Learning and Adaptation

RC-MemNet adopts a common classification-loss formulation for source-region learning and unseen-region adaptation. The regional representation u^r_i and memory representation m^r_i are fused by F_f, and the prediction distribution is

$$\widehat{\pi}^r_i = \mathrm{softmax}\left(\mathrm{Cls}\left(F_f([u^r_i \| m^r_i])\right)/\tau_c\right),$$

where $\mathrm{Cls}(\cdot)$ is the vowel classifier and τ_c is the classification temperature.

For source region r, the classification loss is

$$\mathcal{L}^r_{\mathrm{cls}} = -\frac{1}{|\mathcal{B}_r|} \sum_{i \in \mathcal{B}_r} \log \widehat{\pi}^r_{i,y^r_i}, \tag{10}$$

where $\mathcal{B}_r$ denotes the samples involved in the current source-region update and $\widehat{\pi}^r_{i,y^r_i}$ is the predicted probability of the ground-truth category.

Source regions					*Unseen regions*				
Region	Loc.	Spk.	Seg.	Kept	Region	Loc.	Spk.	Seg.	Kept
Dangtu	3	9	11,932	9,280	Qingyang	4	11	14,583	11,931
Shitai	2	13	16,908	15,582	Tongling	2	8	12,977	7,679
Xuancheng	1	3	3,978	3,978	Jingxian	5	15	19,925	18,600
Fanchang	3	7	9,282	9,282	Nanling	4	11	14,583	14,583
Huangshan	5	16	21,212	21,212	Ningguo	4	4	5,304	2,652
Gaochun	2	7	9,237	7,911	Chizhou	5	16	21,205	17,230
Lishui	4	4	5,302	5,302					
Wuhu	1	3	3,977	3,977					
Total: 45 locations, 127 speakers, 170,405 segmented, 149,199 retained									

Table 1: Regional statistics and evaluation split of WuVowelSet before and after quality control. Loc., Spk., and Seg. denote survey locations, native speakers, and segmented vowel samples, respectively.

During source-region learning, $\mathcal{L}_{\text{cls}}^{r}$ is used to optimize the learnable components introduced by RC-MemNet, including the regional prompts, prompt and query projections, memory keys, query-modulation functions, feature fusion layers, and classifier. The memory values follow the explicit region-level update described above.

During unseen-region adaptation, $\mathcal{L}_{\text{cls}}^{t}$ is computed on the labeled support samples, and only the composition scores $a_{t,r}$ in equation 4 are optimized. All remaining source-trained parameters stay fixed, and the persistent memory is read-only. This restriction preserves the source-trained representations and enables parameter-efficient adaptation under limited supervision.

4 Experiments

4.1 Dataset

To address the scarcity of fine-grained cross-regional Wu vowel resources, we present WuVowelSet, containing 149,199 quality-controlled vowel segments from 127 native speakers across 45 survey locations and 14 Wu dialect regions. The dataset statistics are summarized in Table 1, and all segments are resampled to 16 kHz. The original annotations contain 28 fine-grained IPA vowel realizations, which are normalized into nine shared canonical vowel categories for cross-regional recognition: $\{a, e, i, o, u, y, O, @, E\}$, while retaining the original IPA transcription for each segment. Source-region data are used for training, while unseen-region data are reserved for adaptation and evaluation, with no overlap in regions or speakers between the two partitions. Annotation details and quality-control procedures are provided in the supplementary material.

4.2 Experimental Setup

Extensive experiments are conducted on WuVowelSet following the cross-regional Wu vowel recognition protocol. RC-MemNet is first trained on the source regions and then adapted and evaluated on the unseen regions. For each unseen region, k labeled support samples per vowel category are selected from speakers disjoint from those in the query set and used for unseen-region adaptation. Across the 100 repetitions, each method uses a fixed source-trained state and the same query set, while only the support set is resampled. All methods use identical support-set indices in each repetition, and query samples are used only for evaluation. RC-MemNet takes 80-dimensional log-Mel features extracted from 16 kHz recordings as input and employs a frozen multilingual Whisper-Base encoder Radford et al. (2023). The main comparison and ablation studies use $k = 4$. Source-region learning uses AdamW Loshchilov & Hutter (2019) with a learning rate of 3×10^{-4}, a weight decay of 10^{-4}, and a batch size of 64; each source region contributes at most 3,000 cumulative optimization steps. The source-region order is randomly shuffled at the beginning of each epoch using the training seed of 42.

Method	Qingyang	Tongling	Jingxian	Nanling	Ningguo	Chizhou
SVM (1995)	24.99 ± 2.30	27.31 ± 2.51	20.91 ± 1.96	24.69 ± 2.36	26.84 ± 2.59	28.38 ± 3.40
XGBoost (2016)	19.38 ± 2.20	21.15 ± 2.44	16.36 ± 1.92	19.36 ± 2.32	19.88 ± 2.79	22.23 ± 3.76
LightGBM (2017)	17.45 ± 1.96	19.22 ± 2.35	15.02 ± 1.70	17.73 ± 2.24	17.90 ± 2.28	20.12 ± 3.39
Allophant (2023)	40.28 ± 4.78	39.54 ± 4.95	26.01 ± 5.03	40.24 ± 5.22	33.39 ± 3.18	43.80 ± 4.58
wav2vec 2.0 (2020)	38.11 ± 3.60	34.83 ± 4.09	27.86 ± 4.67	35.36 ± 4.95	26.98 ± 2.51	38.10 ± 4.45
HuBERT (2021)	46.70 ± 3.98	45.96 ± 5.49	33.42 ± 4.31	49.75 ± 4.51	38.13 ± 2.72	52.60 ± 3.39
WavLM Base (2022)	37.89 ± 4.74	38.51 ± 5.44	25.24 ± 3.67	40.57 ± 5.34	30.44 ± 3.56	41.35 ± 4.40
XLS-R (2022)	24.36 ± 3.45	24.35 ± 4.48	20.27 ± 3.50	23.97 ± 3.50	22.33 ± 3.13	25.16 ± 3.85
Whisper (2022)	52.26 ± 3.16	51.46 ± 4.34	38.36 ± 4.05	50.66 ± 3.62	43.40 ± 3.93	50.83 ± 4.20
MR-HuBERT (2024)	52.61 ± 2.61	53.86 ± 4.53	38.84 ± 4.67	53.06 ± 4.25	46.32 ± 2.54	56.44 ± 2.69
mHuBERT-147 (2024)	50.74 ± 4.24	47.63 ± 5.81	36.87 ± 5.06	50.23 ± 5.76	39.70 ± 3.06	51.27 ± 4.82
MS-HuBERT (2024)	54.03 ± 2.37	53.50 ± 4.25	39.06 ± 4.27	54.80 ± 3.66	44.71 ± 3.49	57.58 ± 2.62
SpeechPrompt (2024)	58.41 ± 1.42	60.02 ± 1.67	45.93 ± 1.18	60.90 ± 1.39	55.74 ± 1.24	60.71 ± 1.58
MAS-LoRA (2025)	$\underline{67.84} \pm 2.41$	$\underline{65.35} \pm 3.03$	$\underline{54.97} \pm 0.41$	$\underline{62.07} \pm 3.30$	$\underline{64.76} \pm 2.76$	$\underline{66.29} \pm 2.46$
WhisAID (2026)	49.59 ± 4.02	49.68 ± 3.56	47.39 ± 3.92	46.09 ± 4.27	47.82 ± 4.00	46.25 ± 4.20
RC-MemNet	$\mathbf{72.40} \pm 0.42$	$\mathbf{69.65} \pm 0.36$	$\mathbf{64.99} \pm 2.68$	$\mathbf{68.56} \pm 0.51$	$\mathbf{66.90} \pm 1.90$	$\mathbf{72.77} \pm 0.42$

Table 2: Accuracy (%) on six unseen regions under the 4-shot setting. Results are reported as mean $\pm$ standard deviation over 100 repetitions. The best and second-best mean accuracies in each column are shown in bold and underlined, respectively.

Variant	Qingyang	Jingxian	Chizhou
Main modules			
w/o RPL and RGMA	58.26	44.36	56.83
w/o RGMA	61.80	46.92	62.85
w/o RPL	65.30	52.87	67.95
RPL components			
w/o layer-specific prompt projection	68.92	50.45	69.68
w/o regional feature fusion	70.56	50.78	72.57
w/o adaptive prompt composition	65.72	49.70	68.18
RGMA components			
w/o persistent memory & query modulation	61.80	46.92	62.85
w/o query modulation	69.80	48.63	71.73
RC-MemNet	**72.40**	**64.99**	**72.77**

Table 3: Ablation studies of the main modules and their components under the 4-shot setting.

During unseen-region adaptation, only the prompt-composition scores in Eq. 4 are optimized for 30 steps with a learning rate of 5×10^{-2}. All remaining source-trained parameters are fixed, and the persistent memory is read-only. Results are reported as mean $\pm$ standard deviation over 100 support-set samplings under this fixed experimental configuration, and the mean accuracy across the unseen regions is used for overall comparison. Further implementation details are provided in the supplementary material.

4.3 Comparison with Representative Methods

As shown in Table 2, three main observations can be made. (i) RC-MemNet achieves the highest accuracy on all six unseen regions. Its mean accuracy across the unseen regions reaches 69.21%, compared with 63.55% for MAS-LoRA, the strongest competing baseline, yielding an improvement of 5.66 percentage points. Specifically, RC-MemNet outperforms MAS-LoRA by 4.56, 4.30, 10.02, 6.49, 2.14, and 6.48 points on Qingyang, Tongling, Jingxian, Nanling, Ningguo, and Chizhou, respectively. The consistent gains across all six regions indicate that the overall improvement is not driven by a single regional condition. (ii) Across the compared method families, RC-MemNet consistently outperforms the phonological knowledge-driven methods and both the general discriminative and condition-adaptive data-driven methods. The 10.02-point gain on Jingxian further indicates

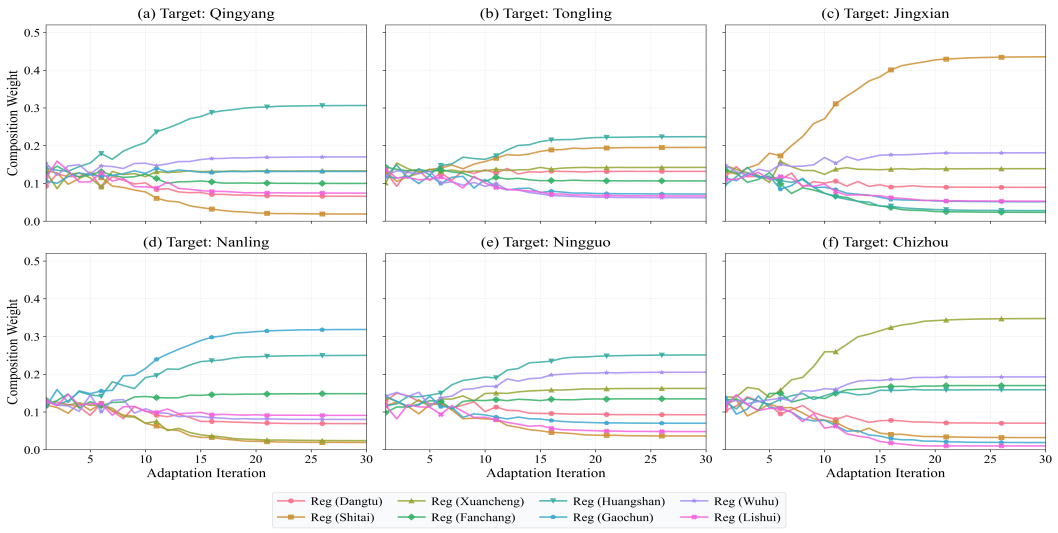

Figure 2: Evolution of source-prompt composition weights during adaptation. Each curve denotes one source-region prompt.

the benefit of RC-MemNet in a region where the compared methods achieve relatively low performance. (iii) These gains demonstrate the effectiveness of the overall RC-MemNet framework. The individual and combined contributions of RPL and RGMA are examined in the following ablations.

4.4 Ablation Studies

To evaluate the contribution of each design in RC-MemNet, we conduct ablation studies on three unseen regions: Qingyang, Jingxian, and Chizhou, with the results summarized in Table 3.

Overall components. We first evaluate the individual and combined contributions of RPL and RGMA. The variant without RPL removes the regional prompts and accesses the persistent memory using the acoustic-only base query. Retaining only RPL improves accuracy over the model without either module by 3.54, 2.56, and 6.02 percentage points on Qingyang, Jingxian, and Chizhou, respectively. Retaining only RGMA improves accuracy over the model without either module by 7.04, 8.51, and 11.12 points on the three regions. The complete RC-MemNet further outperforms the stronger single-module variant by 7.10, 12.12, and 4.82 points, respectively. It achieves the highest accuracy on all three regions, demonstrating the complementary contributions of RPL and RGMA.

RPL component analysis. With RGMA retained, we evaluate three components of RPL: layer-specific prompt projection, regional feature fusion, and adaptive prompt composition. Removing them decreases accuracy by 3.48/14.54/3.09, 1.84/14.21/0.20, and 6.68/15.29/4.59 points on Qingyang, Jingxian, and Chizhou, respectively. These results verify the effectiveness of RPL components, especially adaptive source-prompt composition.

As shown in Fig. 2, source prompts initialized with uniform weights gradually form distinct mixtures across target regions, indicating target-dependent prompt composition rather than fixed aggregation.

RGMA component analysis. With RPL retained, we evaluate persistent memory and regional-context-based query modulation within RGMA. The first variant removes both persistent memory and query modulation. The second retains persistent memory but replaces the regional-context-modulated query with the acoustic-only base query.

Compared with the variant without persistent memory and query modulation, introducing persistent memory with an acoustic-only query improves accuracy by 8.00, 1.71, and 8.88 percentage points on Qingyang, Jingxian, and Chizhou, respectively. Regional-context-based query modulation provides further gains of 2.60, 16.36, and 1.04 points. Since the latter two variants share the same persistent memory and differ only in query construction, their performance difference demonstrates the contribution of regional-context-based query modulation.

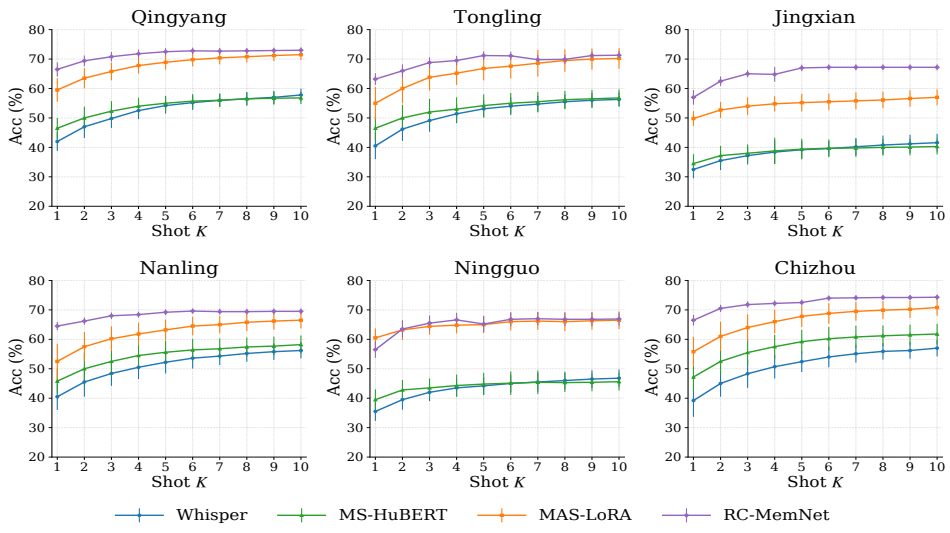

Figure 3: Impact of the number of shots on cross-regional Wu vowel recognition in unseen regions.

Further analysis of RGMA is provided in the supplementary material. The ablation results above verify that the proposed memory read and write mechanisms effectively support category-aware cross-regional knowledge aggregation.

4.5 Few-Shot Generalization

To evaluate RC-MemNet with different amounts of labeled data, we vary the number of support samples k per vowel category from one to ten. We compare RC-MemNet with Whisper, MS-HuBERT, and MAS-LoRA, with the results shown in Fig. 3, where the error bars denote the standard deviation.

RC-MemNet consistently achieves the strongest overall performance in low-shot settings. Its average accuracy increases from 62.27% at $k = 1$ to 69.21% at $k = 4$, and further reaches 70.62% at $k = 10$. As more labeled samples become available, RC-MemNet generally maintains the best performance, although it becomes comparable to MAS-LoRA on Tongling and Ningguo at larger support sizes. The smaller improvement beyond $k = 4$ indicates that RC-MemNet obtains most of its adaptation gains from limited unseen-region supervision, supporting the effectiveness of the proposed prompt-composition-based adaptation.

5 Conclusion

We propose RC-MemNet, a regional context memory network for cross-regional Wu vowel recognition. RPL derives sample-adaptive regional context for intra-regional feature representation, while RGMA uses category-aware persistent memory to model acoustic dependencies across source regions. We also construct WuVowelSet with 149,199 vowel segments from 127 native speakers across 14 Wu dialect regions. Experiments on WuVowelSet demonstrate that RC-MemNet achieves state-of-the-art performance.

AI Use Statement

Generative AI tools were used for language editing, structural revision, and LaTeX formatting assistance. All technical content, experimental design, results, and conclusions were checked by the authors.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020.

Rapha"el Bagat, Irina Illina, and Emmanuel Vincent. Mixture of LoRA experts for low-resourced multi-accent automatic speech recognition. In *Interspeech 2025*, pp. 1143–1147, 2025. doi: 10.21437/Interspeech.2025-775.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. mHuBERT-147: A compact multilingual HuBERT model. In *Interspeech 2024*, pp. 3939–3943, 2024. doi: 10.21437/Interspeech.2024-938.

Carlos Carvalho and Alberto Abad. Memory-augmented conformer for improved end-to-end long-form ASR. In *Proceedings of Interspeech 2023*, pp. 2218–2222, 2023. doi: 10.21437/Interspeech.2023-893.

Kai-Wei Chang, Haibin Wu, Yu-Kai Wang, Yuan-Kuei Wu, Hua Shen, Wei-Cheng Tseng, Iu-Thing Kang, Shang-Wen Li, and Hung-Yi Lee. SpeechPrompt: Prompting speech language models for speech processing tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3730–3744, 2024. doi: 10.1109/TASLP.2024.3436618.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113.

Kevin Glocker, Aaricia Herygers, and Munir Georges. Allophant: Cross-lingual phoneme recognition with articulatory attributes. In *Proceedings of Interspeech 2023*, pp. 2258–2262, 2023. doi: 10.21437/Interspeech.2023-772.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021a. doi: 10.1109/TASLP.2021.3122291.

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. In *Interspeech 2021*, pp. 721–725, 2021b. doi: 10.21437/Interspeech.2021-236.

Yosuke Kashiwagi, Hayato Futami, Emiru Tsunoo, Siddhant Arora, and Shinji Watanabe. Rapid language adaptation for multilingual E2E speech recognition using encoder prompting. In *Interspeech 2024*, pp. 2900–2904, 2024. doi: 10.21437/Interspeech.2024-702.

Sangmin Lee, Woojin Chung, and Hong-Goo Kang. LAMA-UT: Language agnostic multilingual ASR through orthography unification and language-specific transliteration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24393–24401, 2025. doi: 10.1609/aaai.v39i23.34617.

Bo Li, Tara N. Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yonghui Wu, and Kanishka Rao. Multi-dialect speech recognition with a single sequence-to-sequence model. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4749–4753, 2018. doi: 10.1109/ICASSP.2018.8461886.

Shaojun Li, Daimeng Wei, Hengchao Shang, Jiaxin Guo, ZongYao Li, Zhanglin Wu, Zhiqiang Rao, Yuanchang Luo, Xianghui He, and Hao Yang. Speaker-smoothed kNN speaker adaptation for end-to-end ASR. In *Proceedings of Interspeech 2024*, pp. 2390–2394, 2024a. doi: 10.21437/Interspeech.2024-1873.

Song Li, Yongbin You, Xuezhi Wang, Ke Ding, and Guanglu Wan. Enhancing multilingual speech recognition through language prompt tuning and frame-level language adapter. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 10941–10945, 2024b. doi: 10.1109/ICASSP48485.2024.10446990.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W. Black, and Florian Metze. Universal phone recognition with a multilingual allophone system. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8249–8253, 2020. doi: 10.1109/ICASSP40776.2020.9054362.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and Xie Chen. Speech recognition meets large language model: Benchmarking, models, and exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24840–24848, 2025. doi: 10.1609/aaai.v39i23.34666.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

Jiatong Shi, Hirofumi Inaguma, Xutai Ma, Ilia Kulikov, and Anna Sun. Multi-resolution HuBERT: Multi-resolution speech self-supervised learning with masked unit prediction. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kUuKFW7DIF.

Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, and Michael L. Seltzer. Emformer: Efficient memory transformer based acoustic model for low-latency streaming speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6783–6787, 2021. doi: 10.1109/ICASSP39728.2021.9414560.

Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie. Domain adversarial training for accented speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4854–4858, 2018. doi: 10.1109/ICASSP.2018.8462663.

Chengyou Wang, Mingchen Shao, Jingbin Hu, Zeyu Zhu, Hongfei Xue, Bingshen Mu, Xin Xu, Xingyi Duan, Binbin Zhang, Pengcheng Zhu, Chuang Ding, Xiaojun Zhang, Hui Bu, and Lei Xie. WenetSpeech-Wu: Datasets, benchmarks, and models for a unified chinese wu dialect speech processing ecosystem. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27999–28011. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1395.

Xintong Wang and Ye Wang. Joycent: Diffusion-based accent TTS without accented phone prediction. *arXiv preprint arXiv:2606.16417*, 2026.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, Peng Xu, and Pascale Fung. Learning fast adaptation on cross-accented speech recognition. In *Proceedings of Interspeech 2020*, pp. 1276–1280, 2020. doi: 10.21437/Interspeech.2020-45.

Chao Wu, Yongqiang Wang, Yangyang Shi, Ching-Feng Yeh, and Frank Zhang. Streaming transformer-based acoustic models using self-attention with augmented memory. In *Proceedings of Interspeech 2020*, pp. 2132–2136, 2020. doi: 10.21437/Interspeech.2020-2079.

Qiantong Xu, Alexei Baevski, and Michael Auli. Simple and effective zero-shot cross-lingual phoneme recognition. In *Proceedings of Interspeech 2022*, pp. 2113–2117, 2022. doi: 10.21437/Interspeech.2022-60.

Hemant Yadav, Sunayana Sitaram, and Rajiv Ratn Shah. MS-HuBERT: Mitigating pre-training and inference mismatch in masked language modelling methods for learning speech representations. In *Interspeech 2024*, pp. 5053–5057, 2024. doi: 10.21437/Interspeech.2024-1978.

Sanghyun Yoo, Inchul Song, and Yoshua Bengio. A highly adaptive acoustic model for accurate multi-dialect speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5716–5720, 2019. doi: 10.1109/ICASSP.2019.8683705.

Zhao You, Dan Su, Jie Chen, Chao Weng, and Dong Yu. DFSMN-SAN with persistent memory model for automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7704–7708, 2020. doi: 10.1109/ICASSP40776.2020.9053568.

Xinyi Yue, Lizhi Miao, and Jiahao Ding. Research on wu dialect recognition and regional variations based on deep learning. *Applied Sciences*, 15(18):10227, 2025. doi: 10.3390/app151810227.

Yu Yue and Fang Hu. Vowels and diphthongs in hangzhou wu chinese dialect. In *Proceedings of Interspeech 2018*, pp. 207–211, 2018. doi: 10.21437/Interspeech.2018-1225.

Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Lirong Dai, and Yu Hu. Feedforward sequential memory networks: A new structure to learn long-term dependency. *arXiv preprint arXiv:1512.08301*, 2015.

Shiliang Zhang, Ming Lei, Zhijie Yan, and Lirong Dai. Deep-FSMN for large vocabulary continuous speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5869–5873, 2018. doi: 10.1109/ICASSP.2018.8461404.

Yingzhu Zhao, Chongjia Ni, Cheung-Chi Leung, Shafiq Joty, Eng Siong Chng, and Bin Ma. A unified speaker adaptation approach for asr. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9339–9349. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.737.

Jiaming Zhou, Shiwan Zhao, Yaqi Liu, Wenjia Zeng, Yong Chen, and Yong Qin. kNN-CTC: Enhancing ASR via retrieval of CTC pseudo labels. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 11006–11010, 2024. doi: 10.1109/ICASSP48485.2024.10447075.

SUPPLEMENTARY MATERIAL

A ADDITIONAL DETAILS OF WUVOWELSET

This section provides additional information on vowel-category normalization, annotation and quality control, participant protection, data availability, and geographic coverage.

A.1 IPA NORMALIZATION AND RECOGNITION CATEGORIES

WuVowelSet retains 28 fine-grained vowel realizations transcribed using the International Phonetic Alphabet (IPA). For cross-regional recognition, these realizations are normalized into nine shared canonical vowel categories, {a, e, i, o, u, y, ɔ, ə, ɛ}. The normalization groups fine-grained realizations into broader canonical categories according to their overall articulatory similarity in the IPA vowel space, considering vowel height, backness, and lip rounding. The original IPA transcription is retained for each segment and is not replaced by the normalized task label.

Table A.1 gives the complete normalization from the 28 fine-grained IPA vowel realizations to the nine recognition categories.

The normalization defines a shared label space for cross-regional classification rather than replacing the original phonetic annotations. In particular, fine-grained regional differences remain available through the original IPA transcriptions, whereas the normalized labels provide common recognition targets across regions.

A.2 ANNOTATION AND QUALITY CONTROL

All vowel segments were resampled to 16 kHz. Two trained annotators independently performed temporal segmentation and IPA transcription, achieving 86.7% exact agreement on the original IPA vowel labels (Cohen's $\kappa = 0.82$), 96.9% agreement after mapping to the nine recognition categories, and 92.1% temporal-segmentation agreement when both boundaries were required to differ by no more than 20 ms; a third expert adjudicated the remaining disagreements in segment boundaries and vowel labels. Segments were excluded when their boundaries could not be determined reliably, severe noise obscured the target vowel, overlapping speech prevented reliable transcription, the vowel structure remained unclear, or the label could not be resolved after expert review. This procedure retained 149,199 vowel segments from 170,405 initially segmented samples, using the same criteria across all regional subsets.

A.3 PARTICIPANT PROTECTION AND DATA AVAILABILITY

Prior informed consent was obtained from all participants for research use and public release of the resulting vowel data. Direct personal identifiers were removed, speakers are represented by anonymous codes, and geographic information is retained only at the survey-location and regional levels used in the analysis.

Upon publication, the vowel-level audio samples, annotations, and predefined regional partition will be released as WuVowelSet for research use. The metadata will include anonymous speaker codes, survey locations, regional labels, original vowel transcriptions, task labels, and the experimental partition. These procedures provide the data basis for the protocol described next.

A.4 GEOGRAPHIC COVERAGE

WuVowelSet covers 45 survey locations across 14 Wu dialect regions. It contains 149,199 quality-controlled vowel segments collected from 127 native speakers.

B TRAINING, EVALUATION, AND REPRODUCTION DETAILS

This section specifies the common evaluation protocol, the source-learning and target-adaptation procedure of RC-MemNet, and the common fairness controls used for method comparison. Source and target domains refer respectively to the source and unseen regions defined in the main paper.

Task category	Original IPA realizations
/i/	/i/, /ɪ/, /ɨ/
/y/	/y/, /ʏ/, /ʉ/
/u/	/u/, /ʊ/, /ɯ/
/e/	/e/, /ø/
/o/	/o/, /ɤ/
/ə/	/ə/, /ɔ/, /ɵ/, /ɜ/, /˜/, /ɐ/
/ɛ/	/ɛ/, /œ/, /æ/
/ɔ/	/ɔ/, /ʌ/, /ɒ/
/a/	/a/, /ɶ/, /ɑ/

Table A.1: Normalization of the 28 fine-grained IPA vowel realizations into the nine shared recognition categories. The mapping is defined according to overall articulatory similarity in the IPA vowel space.

B.1 Data Partition and Evaluation Protocol

RC-MemNet is trained on the predefined source domains, while the source-validation split contains no target-domain samples. For each target domain, the query set is fixed. In every repetition, $k = 4$ labeled support samples per vowel category are drawn from speakers disjoint from those in the query set. All methods use identical support indices, and query samples and labels are used only for evaluation.

For methods requiring checkpoint selection, the checkpoint with the highest source-validation Macro-F1 is selected once and fixed across 100 support-set samplings. Results are reported as mean $\pm$ standard deviation, and the overall result is the unweighted average of the target-domain means. Under this shared protocol, the following implementation details define the update scope of RC-MemNet.

B.2 RC-MemNet Implementation

The multilingual Whisper-Base encoder remains frozen. During source learning, gradients update the source prompts, prompt and query projections, context-modulation networks, memory keys, fusion layers, and classifier; memory values and cumulative states are modified only by the explicit no-gradient writing rule.

Each source domain has a cumulative budget of at most 3,000 predictive parameter updates. After every effective visit to a source domain, one no-gradient memory-writing pass is performed without consuming this budget. Domains that exhaust their budgets receive no further predictive updates, and source learning terminates when the global source-validation stopping criterion is met or all budgets are exhausted. During target adaptation, all source-trained parameters and the persistent memory remain fixed; only the uniformly initialized composition scores $\{a_{t,r}\}_{r=1}^{R}$ are optimized on the support set and then fixed for query evaluation. The corresponding configuration is summarized in Table B.1.

Table B.1 separates the fixed architecture and optimization settings from the evaluation controls. With the proposed method fully specified, we next summarize how the baselines are kept comparable.

B.3 Baseline Reproduction and Fairness Controls

All compared methods use the same retained segments, recognition categories, source/target partition, source-validation split, fixed query sets, and speaker-disjoint support sampling with identical support indices in each repetition. Neural checkpoints are selected using source-validation data and remain fixed; non-neural models are fitted on source data and refitted on labeled support features. Query labels are never used for training, adaptation, memory updating, or model selection.

For the neural baselines, we follow a common source-training and few-shot evaluation protocol while preserving the adaptation mechanism of each method. The backbone, source-trainable pa-

Item	Setting
Input and encoder	16-kHz vowel segments; 80-dimensional log-Mel spectrograms; multilingual Whisper-Base frozen during source learning and target adaptation.
RPL	One base prompt per source domain; $S_p = 8$, intermediate hidden dimension $d_h = 256$, and prompt dimension $d_p = 128$. Only composition scores are updated at the target stage.
RGMA	$d_m = 128$, $J = 4$, $K_w = 1$, $\tau_a = 0.2$, $\tau_c = 0.2$, memory fusion coefficient 0.5; memory values and cumulative states initialized to zero.
Source learning	AdamW; learning rate 3×10^{-4}; weight decay 10^{-4}; batch size 64; at most 3,000 cumulative predictive updates per source domain; source-domain order shuffled at the beginning of each epoch.
Memory and selection	One no-gradient writing pass after every effective source-domain visit; unassigned memory items remain unchanged. The checkpoint with the highest source-validation Macro-F1 is retained.
Target adaptation	Composition scores initialized uniformly and optimized with AdamW for 30 steps at learning rate 5×10^{-2}; all other parameters fixed and memory read-only.
Evaluation	4-shot by default; 100 speaker-disjoint support samplings; fixed query sets and identical support indices across methods.
Environment	Ubuntu 22.04.1; AMD EPYC 9354; 125 GiB RAM; one NVIDIA RTX 5090 with 32 GiB; Python with PyTorch, NumPy, pandas, scikit-learn, and librosa.

Table B.1: Principal implementation and evaluation settings of RC-MemNet.

rameters, optimization settings, and target-stage adaptation procedure of the strongest adaptation baselines are specified below.

For MAS-LoRA, we use Whisper-Base as the acoustic backbone. The Whisper encoder is kept fixed except for its top two encoder layers and normalization layers, while region-specific low-rank experts with rank $r = 16$ and the vowel classifier are optimized on the source-region data. Source-stage training uses AdamW with a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, a batch size of 64, and a training budget of 3,000 optimization steps. For an unseen target region, the outputs of the source-region experts are averaged to obtain the target representation. The $k = 4$ labeled support samples per vowel category are used to construct ℓ_2-normalized class prototypes, and query samples are classified by cosine similarity to these prototypes.

For WhisAID, we also use Whisper-Base as the acoustic backbone. The convolutional feature-extraction layers are kept fixed, while the remaining encoder parameters together with the region, speaker, and vowel prediction heads are optimized on the source-region data. Its source-stage objective combines vowel classification, region classification, and adversarial speaker classification, with the speaker-loss weight and gradient-reversal coefficient both set to 0.05. Source-stage training uses AdamW with a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, a batch size of 64, and a training budget of 3,000 optimization steps. At the target stage, the encoder representation is combined with the predicted region representation, and the same $k = 4$ labeled support samples per vowel category are used to construct ℓ_2-normalized class prototypes. Query samples are then classified by cosine similarity to these prototypes.

The corresponding source-trainable parameter counts and target-stage adaptation mechanisms are summarized in Table B.2.

B.4 Parameter and Adaptation Cost

Source-trainable parameters denote optimizer-updated parameters during source learning; frozen backbones and support-derived prototypes are excluded. Table B.2 reports this common counting rule together with the target-domain adaptation mechanism.

RC-MemNet reduces source-trainable parameters by 96.57% relative to WhisAID and 95.83% relative to MAS-LoRA, while restricting target adaptation to a small set of prompt-composition vari-

Method	Source-trainable	Target adaptation
WhisAID	71,257,339	Support-derived prototypes
MAS-LoRA	58,671,113	Support-derived prototypes
RC-MemNet	2,444,147	Composition scores, 30 steps

Table B.2: Source-trainable parameters and target-domain adaptation.

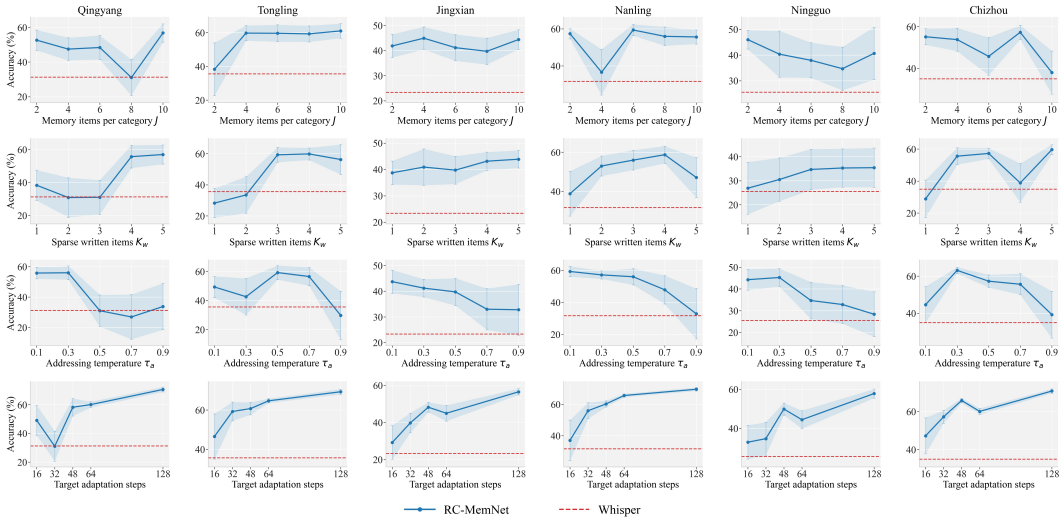

Figure C.1: Hyperparameter sensitivity of RC-MemNet under the 4-shot setting. Rows denote hyperparameters and columns denote target domains. Blue curves, red dashed lines, and shaded bands indicate the mean Accuracy of RC-MemNet, the supervised Whisper baseline, and the standard deviation over repeated support-set samplings, respectively.

ables. The following section examines the resulting behavior from complementary experimental perspectives.

C Additional Experimental Analyses

This section analyzes hyperparameter sensitivity, source-prompt composition dynamics, and category-level classification behavior under the common 4-shot protocol defined in Section B.

C.1 Hyperparameter Sensitivity

Figure C.1 varies one of four principal hyperparameters at a time under the same support/query protocol, while keeping all non-varied hyperparameters at their default settings. Rows denote the examined hyperparameters, columns denote the target domains, and each panel reports classification Accuracy. The blue curves report the mean Accuracy of RC-MemNet over repeated support-set samplings, while the horizontal red dashed lines denote the supervised Whisper baseline under the same 4-shot protocol. The shaded bands indicate the standard deviation over repeated support-set samplings.

As shown in Fig. C.1, sensitivity varies across target domains, yet RC-MemNet remains competitive over a broad range of settings and exceeds the supervised Whisper baseline in most comparisons. The absence of widespread abrupt degradation indicates relatively stable behavior under moderate hyperparameter changes. This analysis characterizes model behavior without introducing target-domain-specific tuning. We next examine the adaptation process that produces the target-domain prompt.

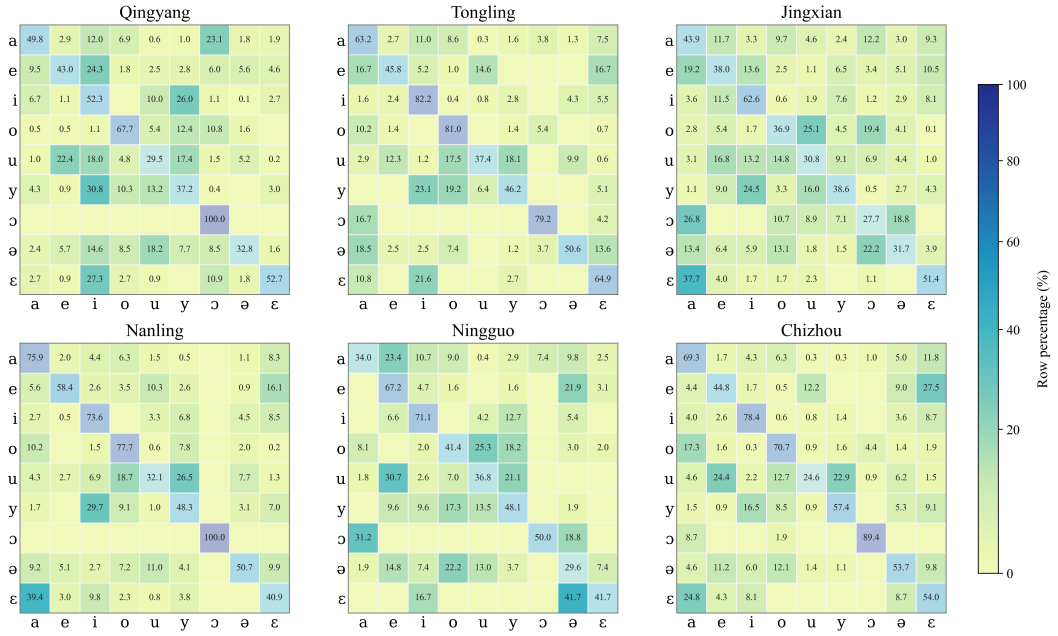

Figure C.2: Row-normalized confusion matrices (%) of RC-MemNet on the target domains under the 4-shot setting, aggregated over 100 support-set samplings. Rows denote ground-truth categories, columns denote predicted categories, and panel titles give the region names.

C.2 Category-Level Classification Analysis

Figure C.2 reports row-normalized confusion matrices under the same 4-shot protocol. For each target domain, confusion counts are accumulated over 100 evaluations on the fixed query set and then normalized independently within each ground-truth row. Each row therefore sums to 100%, assigning equal visual weight to every category.

As shown in Fig. C.2, the target domains exhibit distinct category-level error structures. Because each row is normalized independently, the diagonal cells directly reflect the recognition rate of the corresponding ground-truth category, whereas the off-diagonal cells show how its remaining errors are distributed. Jingxian has the broadest off-diagonal distribution, consistent with its lower overall performance. The other target domains retain more mass near the diagonal, while their detailed error patterns remain different. The matrices therefore show that the regional differences reported by Accuracy are not caused by one shared confusion pattern; each target domain presents a different combination of well-recognized and difficult vowel categories.

Table C.1 complements this visual evidence with Macro-F1 and Weighted-F1. Macro-F1 averages the category-level F1 scores uniformly, while Weighted-F1 weights them by the number of query samples in each category. Both metrics are computed independently for each support sampling, and the reported means and standard deviations are calculated over the 100 samplings. The regional mean is then obtained as the unweighted average of the region-level means, ensuring that every target domain contributes equally to the summary result.

Table C.1 quantitatively supports the confusion-matrix observations. Chizhou obtains the highest Macro-F1 and Weighted-F1 scores of 65.36% and 72.36%, respectively, followed by Qingyang with 64.46% and 71.82%. Jingxian obtains the lowest scores of 42.84% and 53.27%, which are 22.52 and 19.09 percentage points below those of Chizhou. This difference is consistent with the broader off-diagonal pattern in Fig. C.2 and identifies Jingxian as the most challenging target domain at the category level.

The reported standard deviations further characterize sensitivity to the sampled support set. Ningguo has comparatively larger deviations than the other target domains, indicating that its 4-shot perfor-

Table C.1: Comparative Macro-F1 and Weighted-F1 results (%) on six unseen regions under the 4-shot setting.

Model	Metric	Qingyang	Tongling	Jingxian	Nanling	Ningguo	Chizhou
Whisper	Macro-F1	46.53±2.81	44.09±3.72	25.29±2.67	45.52±3.25	38.74±3.51	45.65±3.77
	Weighted-F1	51.84±3.13	51.55±4.35	31.44±3.32	50.19±3.59	42.83±3.88	50.54±4.18
MS-HuBERT	Macro-F1	48.10±2.11	45.84±3.64	25.75±2.81	49.24±3.29	39.90±3.11	51.72±2.35
	Weighted-F1	53.60±2.35	53.59±4.26	32.02±3.50	54.29±3.63	44.12±3.44	57.26±2.61
MAS-LoRA	Macro-F1	60.40±2.15	56.00±2.60	36.24±0.27	55.77±2.97	57.80±2.46	59.54±2.21
	Weighted-F1	67.30±2.39	65.46±3.04	45.06±0.34	61.49±3.27	63.91±2.72	65.92±2.45
RC-MemNet	Macro-F1	64.46±0.32	59.68±0.56	42.84±0.38	61.60±0.52	59.71±1.87	65.36±0.52
	Weighted-F1	71.82±0.36	69.77±0.34	53.27±0.35	67.92±0.41	66.02±2.14	72.36±0.37

mance is more affected by support-example selection. This variation is reported directly rather than being hidden by the regional mean.

Weighted-F1 exceeds Macro-F1 in every target domain. The differences range from 6.31 to 10.43 percentage points, with a regional-mean gap of 7.92 points. Frequency-weighted evaluation therefore emphasizes better-recognized categories with more query samples, whereas Macro-F1 gives equal importance to all categories regardless of their query-set frequencies. Reporting both metrics prevents the overall classification result from being interpreted only through the dominant categories.

RC-MemNet obtains regional means of 58.94% Macro-F1 and 66.86% Weighted-F1. Taken together, the row-normalized confusion matrices and the two complementary F1 measures provide a category-level account that is consistent with the regional Accuracy differences reported in the main paper.