E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP18;TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2026)09-3382-15

论文引用格式:Luo S T, Wang T Y, Yang S, Ni Q P and Shan S G. 2026. Memory-guided explicit prompting for speaker-adaptive lip reading. Journal of Image and Graphics, 31(9):3382-3396(骆嵩涛, 王天月, 杨双, 倪群平, 山世光. 2026. 记忆机制显式引导的说话人自适应唇语识别. 中国图象图形学报, 31(9):3382-3396)[DOI:10.11834/jig.250392]

记忆机制显式引导的说话人自适应唇语识别

骆嵩涛[1],王天月[1,2],杨双[1*],倪群平[3],山世光[1,2]

1. 中国科学院计算技术研究所智能算法安全全国重点实验室,北京　100190; 2. 中国科学院大学前沿交叉科学学院,北京　101408;
3. 天津712通信广播股份有限公司,天津　300462

摘　要:**目的**　唇语识别中说话人自适应方法通常依赖目标说话人的带标注数据进行微调,但在实际场景中常面临适应数据缺失的问题,为此提出一种基于记忆机制显式引导的说话人自适应方法,旨在不依赖适应数据以提升模型在跨说话人条件下的唇语识别能力。**方法**　通过显式构建静态视觉模式与动态视觉说话模式记忆库,记录训练阶段已见过的说话人的代表性特征,使模型在面对未见说话人时,通过与记忆库中的特征进行匹配与组合,生成具有个体区分性的表示,以达到说话人自适应的唇语识别。推理过程中,该记忆库将以提示形式嵌入到识别模型的注意力机制中,从而显式引导模型在编码阶段对输入的说话人特征进行自适应建模。**结果**　为了验证该方法对说话人自适应唇语识别的效果,基于大规模公开数据集LRS2(lip reading sentences 2)构造了一个新数据集LRS2-ID,通过按照说话人标签重新划分,确保测试集中的说话人未在训练集中出现过,从而更贴近真实应用中"零样本适应"的说话人泛化场景。在该挑战数据集上的跨说话人识别任务的实验结果,验证了方法的有效性与泛化能力。结果表明,在完全无适应数据的条件下,本文方法显著提升了模型对未见过的说话人的识别性能,且在个体差异显著或外观变化复杂的场景中亦表现出了良好的稳定性和普适性。**结论**　本文为提升唇语识别系统的跨个体泛化能力提供了有效思路和实践路径,解决了无适应数据情况下唇语识别中说话人的自适应问题。

关键词:唇语识别;说话人自适应;提示学习;记忆机制;静态视觉模式(SVP);动态视觉说话模式(DVP)

Memory-guided explicit prompting for speaker-adaptive lip reading

Luo Songtao[1], Wang Tianyue[1,2], Yang Shuang[1*], Ni Qunping[3], Shan Shiguang[1,2]

1. State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing　100190, China;
2. School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, Beijing　101408, China;
3. Tianjin 712 Communication & Broadcasting Co., Ltd., Tianjin　300462, China

Abstract: **Objective** Lip reading systems infer linguistic content by analyzing a speaker's lip movements, relying entirely on visual modality compared with traditional speech recognition, thus facing greater uncertainty due to missing modalities and temporal variations. Although recent advances in lip motion modeling and semantic alignment have driven technological progress, existing methods predominantly focus on generic modeling while neglecting the critical impact of individual differences. In practical applications, considerable performance variations across speakers remain a major bottleneck for cross-scenario generalization, making speaker adaptation a core challenge in lip reading research. Current speaker adaptation approaches fall into two categories: 1) generalization strategies without target speaker data, such as constructing

收稿日期:2025-08-18;**修回日期**:2026-01-08;**预印本日期**:2026-01-15

*通信作者:杨双　shuang.yang@ict.ac.cn

基金项目:国家自然科学基金项目(62276247, U24A20332)

Supported by:National Natural Science Foundation of China(62276247, U24A20332)

speaker-invariant representations or employing visual i-vectors for cross-speaker modeling, and 2) personalized strategies with limited target data, including fine-tuning, model pruning, parameter-efficient prompt learning, and contrastive learning. Recent studies have also explored vision-language collaborative mechanisms to enhance individual semantic and visual style perception. However, existing methods still inadequately address the explicit decoupling and utilization of speaker features under zero-resource conditions, particularly failing to distinguish the differential impacts of static appearance features and dynamic behavioral patterns on lip motion modeling. This study aims to tackle this zero-resource adaptation problem by proposing a novel speaker-adaptive lip reading method based on memory-guided prompting. The proposed method seeks to enhance the model's ability to generalize across unseen speakers without accessing any speaker-specific adaptation data. To evaluate systematically the performance of lip reading models in speaker generalization scenarios, this study constructs a new benchmark dataset, LRS2-ID, on the basis of the public LRS2 dataset. LRS2-ID not only preserves the diversity of the original corpus but also introduces speaker identity-controlled annotations to support separate modeling and evaluation of "seen/unseen speakers", providing a practical benchmark for assessing zero-shot adaptation capabilities. **Method** In this study, we introduce an explicit memory prompting framework that leverages structured visual speaker prototypes as external guidance during training and inference. Specifically, the framework constructs two types of speaker-specific memory banks: a static visual pattern memory and a dynamic visual speech pattern memory. Static visual patterns capture facial appearance and are obtained by averaging face features over multiple representative frames. Dynamic speech patterns are derived from frame-level features in the frequency domain, emphasizing interframe variations while suppressing static components. Negative samples are introduced for contrastive learning, where shuffled and reversed sequences generate temporally incorrect samples, training the model to identify correct speaker-specific dynamic patterns. Additionally, a gradient reversal layer is incorporated at the frame level to suppress the model's reliance on static appearance features during backpropagation, further enhancing dynamic pattern recognition. Static and dynamic patterns are clustered into prototype memory banks using K-means clustering, ensuring compactness and scalability. During inference, when encountering an unseen speaker, the model retrieves relevant prototypes from these memory banks based on similarity with the input sequence. The retrieved prototypes are then injected into the attention mechanism of a Conformer-based encoder as key and value components. A hierarchical injection strategy is adopted: Static prototypes are fused into shallow encoder layers to enhance appearance modeling, while dynamic prototypes are injected into deeper layers to guide temporal behavior modeling. This cross-layer prompting scheme allows the model to align and adapt to the unseen speaker's patterns dynamically without modifying model parameters or requiring supervised adaptation. **Result** Comprehensive experiments on the LRS2-ID dataset demonstrate the effectiveness of the proposed method. Results show that the method reduces the word error rate by over 4. 3% absolutely, without requiring target speaker adaptation data. Ablation studies confirm the contributions of each component: Static prompting alone aids stable appearance modeling, particularly for speakers with consistent visual traits, while dynamic prompting captures personalized speech behavior and benefits speakers with diverse visual conditions. The fusion of both prompt types, combined with structured memory organization, consistently enhances model robustness across varying speaker profiles. Per-speaker performance analysis reveals consistent improvements across most speakers, including those with considerable intervideo appearance changes or high articulation variability. Visualization of representative samples highlights the complementary strengths of static and dynamic memory prompts, particularly in challenging scenarios involving illumination shifts, pose variations, or expressive speech. **Conclusion** This study presents a novel memory-based prompting framework for speaker-adaptive lip reading in zero-resource conditions. By explicitly modeling static and dynamic visual speaker patterns and integrating them hierarchically into the attention mechanism of a Conformer-based encoder, the proposed method achieves robust generalization to unseen speakers without requiring adaptation data. Experimental results on the LRS2-ID benchmark confirm the method's effectiveness, stability, and practical value. This study advances structured prompting and memory-enhanced adaptation, paving the way for scalable and user-friendly lip reading systems capable of generalizing across diverse real-world users. Future work may explore finer-grained adaptive mechanisms and additional modalities to further enhance its performance.

Key words: lip reading; speaker adaptation; prompt learning; memory mechanism; static visual pattern(SVP); dynamic visual speech pattern(DVP)

0 引 言

唇语识别是通过分析说话人唇部的运动动态以准确推断其所表达的语言内容信息的技术。与传统语音识别相比,唇语识别完全依赖视觉模态,在推断言语内容信息时常面临遮挡、姿态变化、不均匀光照以及说话人发音差异等带来的输入层面的信息不确定(Wang等,2025)。近年来,虽然已有大量研究关注基于唇部运动的言语内容识别,一定程度上推动了唇语识别技术的发展,但大多数方法忽略说话人个体差异在识别过程中的关键影响。实际应用中,唇语识别模型对不同说话人的识别性能表现存在显著差异,已成为限制模型跨场景泛化能力的主要瓶颈。如何提升模型对不同说话人的适应能力、实现对未见过的说话人的自适应,已成为当前唇语识别研究的核心挑战之一(Luo等,2023)。

随着对个体差异建模需求的增强,说话人自适应学习逐渐成为唇语识别的重要研究方向。根据是否需要目标说话人数据,该类方法可划分为两类:一类为无目标说话人数据的泛化策略,通过构建说话人不变表示(Yang等,2020;Zhang等,2021)或引入视觉i-vector特征等跨说话人建模方法(Kandala等,2019),提升系统在未见说话人上的性能;另一类为有限的目标说话人数据条件下的说话人个性化策略,旨在模型稳定性与个体适应性间寻求平衡,相关方法包括微调(Kim等,2017)、模型剪枝(Zhao等,2021)、参数高效的提示学习(Kim等,2022,2025)与对比学习(黄奕洋,2024)等,如Kim等人(2022)结合卷积神经网络(convolutional neural network, CNN) padding与提示模块,实现了对个体特征的高效适配。Yeo等人(2025)还尝试引入视觉—语言协同机制,进一步增强模型对个体语义与视觉风格的感知能力。

在唇语识别领域,当模型面对未在训练集中出现过的说话人时,识别性能的显著下降已成为公认的挑战。这种现象通常被视为训练数据域与测试数据域之间的不匹配问题,即模型在训练阶段接触到的说话人表现的说话模式与测试阶段遇到的说话人的说话模式不同。本文将说话人在唇语过程中的表现模式分为静态模式和动态模式两大类,前者主要以说话人的唇形、嘴唇厚薄、下颚骨形状、口腔内部形状、牙齿、眼型和鼻型等不变特征为主,可从单帧图像中提取,无需进行时序层面的动态建模。静态模式主要反映说话人的固有生理结构,在说话过程中相对稳定,但不同人之间的静态模式差异性会对其表达言语内容的唇部动作过程产生影响,从而影响对不同人的唇语识别性能。与之相对的,说话人的动态模式则主要包括语速、发音习惯(口型为主)、口癖、眼动和表情等(崔鑫宇 等,2024)。该类特征无法从单帧图像中提取,需要通过分析说话过程中的面部动态获得,直接影响唇语动作的解析,如语速的变化会影响在同样时间内唇部动作的序列长短,进而影响模型的解析效果,发音习惯和口癖则反映了说话人在发音时的独特习惯,在发音时对唇形的动态表现过程产生直接影响,进而影响唇语识别性能。

综上,本文提出一种基于记忆机制显式引导的说话人自适应唇语识别方法,通过构建包含静态模式与视觉动态说话模式的结构化的说话人特征记忆空间,形成对测试时的未见说话人的鲁棒表达,然后以向唇语模型显式引入该说话人特征的提示信息的形式,提升模型对未见说话人的快速适应能力。具体而言,本文基于训练数据中的不同说话人的唇语模式,分别构建静态视觉模式与动态视觉说话模式的记忆库,记录训练阶段说话人的代表性特征,使模型在面对未见说话人时,通过与记忆库中的唇语模式进行匹配与组合,生成针对未见说话人的具有个体区分性的表示。该记忆库将以提示形式嵌入到模型的注意力机制中,从而显式引导模型在编码阶段对输入的说话人特征进行自适应建模。为了系统评估唇语识别模型在说话人泛化场景下的性能表现,本文基于公开数据集LRS2(lip reading sentences 2)构建了一个新的基准数据集LRS2-ID。一方面保留了原始语料库的多样性和挑战性;另一方面基于各数据样本的说话人身份标签,形成了完全未在训练集中出现过的未见说话人集合所构建的测试集,从而使模型的评估更加接近实际场景下零样本说话人的情况。这一新数据集有助于更全面地分析说话人个体差异对识别性能的影响,并为后续研究提供标准化的评测基准数据。

本文主要贡献如下:1)通过解耦说话人的静态与动态特征,分别构建基于静态视觉模式与动态视觉说话模式的记忆原型库,显著提升了模型在未见

说话人上的泛化能力；2）设计了一种分层提示注入机制，将静态视觉模式与动态视觉说话模式分别嵌入模型的浅层和深层注意力模块，实现了对不同层级说话人特征的高效融合与自适应建模；3）针对零样本的说话人自适应唇语识别任务，构建了一个新的具有挑战性的数据集 LRS2-ID，后续将公开该划分，为业界对该任务的评测提供测试基准。在该挑战数据集上的实验结果表明，本文方法能够在完全无适应数据的条件下显著提升模型对未见说话人的识别性能，且在个体差异显著或外观变化复杂的场景中表现出良好的稳定性和普适性。

1 方　法

基于记忆机制显式提示引导的说话人自适应方法，面向唇语识别中不同说话人的泛化问题，设计核心在于，在模型训练与推理过程中主动引入结构化的说话人模式信息，以显式提示的形式利用训练集中已见过的说话人特征来辅助构建未见说话人的特征，从而提升系统在面对不同说话人环境下的适应性和鲁棒性。

首先，基于唇部运动的时序变化特征，提出了动态视觉说话模式（dynamic visual speaking pattern，DVP）新概念，旨在展现建模个体在发音行为中的动态发音风格。与静态视觉外貌类似，唇语识别中同样存在具有区分性且可建模的动态模式，反映了说话人在发音方式、语速控制和面部动作节奏上的一致性特征。具体而言，每位说话人在发音过程中，其面部运动往往呈现出相对稳定的个体化特征，例如特定词语所对应的口型变化模式、停顿与过渡的表现方式等。这些随时间变化所展现出的个体化表达风格，即构成了本文所提出的动态视觉说话模式；与之对应的静态视觉模式（static visual pattern，SVP）则指个体在视觉外观层面的属性，如脸型、肤色和嘴唇轮廓等。

随后，通过在模型表征空间中引入基于训练数据习得的静态视觉模式与动态视觉说话模式，构建两个独立的记忆模块，分别存储由不同说话人的静态外观模式（如面部结构）和动态发音模式（如语速、口型轨迹）所对应的原型集合。

在推理阶段，模型通过计算当下测试时的输入模式与记忆库中各原型之间的相似性，从中选取相关提示向量，并显式嵌入至编码器结构的不同层级，从而引导模型在训练数据中的模式基础上来理解和适应当前输入所对应的个体差异。在该过程中，将在浅层网络中侧重注入静态提示，以感知面部外观等个体特征；在深层网络中引入动态提示，以加强对时序行为如发音节奏和唇动方式的建模能力。通过这种跨层级的显式提示注入机制，模型在识别过程中能够动态参考历史经验，实现高效的个体差异建模与跨说话人泛化。

总体来说，该方法的关键流程包括说话人个体特性建模、表示空间构建、表示空间融合、自适应模式交互以及其与唇语识别任务的适配等。首先，通过说话人个体特性建模，从输入视频中分离并建模个体的静态视觉模式（SVP）与动态视觉说话模式（DVP），形成对说话人双重特性的结构化描述；其次，在表示空间构建阶段，利用训练数据学习并构建两个独立的记忆模块，分别存储 SVP 和 DVP 的原型集合，形成结构化的先验知识库；然后，在表示空间融合过程中，模型在前向传播时动态检索记忆库，提取与当前输入最相关的静态与动态提示向量；进一步地，通过自适应模式交互机制，将这些提示向量按层级显式注入编码器网络（浅层注入静态提示，深层注入动态提示），实现对个体差异的精细化建模与上下文感知的自适应调整；最后，将该自适应表征用于唇语识别任务，在词汇或句子级别实现更鲁棒的跨说话人识别，体现其任务适配性与性能优势。图 1 为本文总体技术流程。

1.1 预备知识

给定输入的唇部视频序列 $\boldsymbol{X}=[x_1,x_2,\cdots,x_T]$，其中 T 表示视频帧数，唇语识别任务的目标是将其准确映射为对应的文本序列 $\boldsymbol{Y}=[y_1,y_2,\cdots,y_L]$，其中 L 为文本序列的长度。

本文方法的核心是构建两个结构化的记忆模块：SVPs 和 DVPs 记忆库，统一用 $\boldsymbol{D}=\{D_1,D_2,\cdots,D_{D_M}\}$ 来表示，D_M 表示记忆库中原型的数量，这些记忆库通过在训练数据上对所有说话人的特征进行聚类分析得到，作为丰富的先验知识库，在推理阶段为模型提供有效的提示信息。模型的整体处理流程可表示为

$$\hat{\boldsymbol{Y}}=f_{\mathrm{VSR}}(\boldsymbol{X};\boldsymbol{D}) \tag{1}$$

式中，f_{VSR} 表示唇语识别模型，其参数通过优化识别

损失函数进行学习。

1.2 说话人个体特性建模

为了有效建模唇语识别中说话人的个体差异，本文基于Luo等人(2025)提出的静态与动态视觉说话模式来综合表征说话人，从不同角度揭示说话人的视觉个体性，构成了本文说话人建模与泛化的基础。

1.2.1 静态视觉模式提取

为了有效建模说话人的静态视觉特征，采用预训练的ArcFace(Deng等，2019)模型作为基础特征提取器，如图1左下蓝色部分所示。对于每个说话人的视频数据，首先从视频帧中提取多帧人脸图像，然后通过ArcFace模型提取每帧图像的特征表示，最后采用多帧平均策略获得稳定可靠的静态视觉模式表示，具体为

$$z_s = \frac{1}{N}\sum_{i}^{N} f_{\text{ArcFace}}(\boldsymbol{I}_i) \tag{2}$$

式中，$\boldsymbol{I}_i$表示第i帧人脸图像，N为用于特征提取的总帧数，f_{ArcFace}代表ArcFace模型的特征提取函数。这种多帧平均策略有效避免了因单帧图像质量不佳或光照条件变化带来的特征不稳定问题。

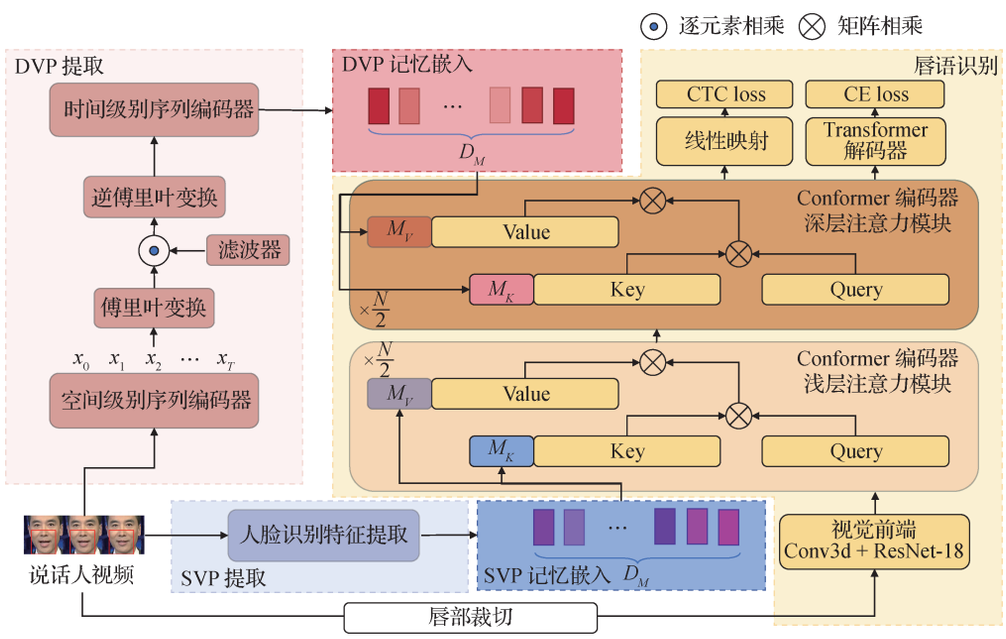


图1 技术路线框架

Fig. 1 Overall framework

1.2.2 动态视觉说话模式提取

为了提取动态视觉说话模式，本文采用一套完整的时序特征处理流程(如图1左上的DVP部分所示)。输入视频序列$\boldsymbol{X}$，首先采用2D-CNN结构的空间级别序列编码器提取初始帧级特征$\boldsymbol{H}=[h_1, h_2, \cdots, h_T]$。随后对初始帧特征施加离散傅里叶变换(discrete Fourier transform，DFT)，将其从时域转换为频域序列(李夫辰 等，2025)。通过高通滤波方式去除低频静态成分(对应于静态视觉模式)，突出帧间动态变化(对应于动态视觉说话模式)。接着，以原始特征序列作为正样本进行说话人分类，学习正常的动态发音模式；同时构造反转和打乱的序列作为负样本，破坏时间顺序并赋予均匀分布标签，使模型不将其误认为任何真实说话人；将负样本与原始正样本进行对比训练，促使模型学习区分自然发音节奏与非自然时序模式，从而增强对个体化动态特征的建模能力。最后，在帧级分类任务中引入梯度反转层(gradient reversal layer，GRL)，以抑制模型对静态外观特征的依赖，进一步提升其利用动态时序信息进行说话人识别的能力(Li等，2025)。动态视觉说话模式提取模块通过前述频域滤波、时域负样本增强与梯度反转抑制等方法，训练得到用于提取说话人动态模式的特征网络。考虑到在唇语识别阶段只需关注动态信息，提取模块中仅保留了频域滤波的解耦部分。当输入说话人视频时，其帧级特征经由频域滤波处理以抑制静态成分，然后通过逆

离散傅里叶变换重构为时域信号，从而生成描述个体动态发音特征的说话人动态说话模式z_d。

1.3 表示空间构建

为了将编码得到的静态说话模式和动态视觉说话模式进一步应用于模型的说话人自适应，本文设计了一个具有结构化记忆功能的模块，用以存储不同说话人的代表性视觉模式。如图1 SVP记忆嵌入（下方蓝色区域）与DVP记忆嵌入（上方红色区域）所示，分别存储静态模式和动态模式。这些记忆库不仅对训练数据中的说话人进行抽象建模，还具备良好的可泛化能力，能够在推理阶段为模型提供历史经验，以辅助新说话人的识别过程。

本文分别构建了静态视觉模式和动态视觉说话模式两个独立的记忆库，并将每个库中的模式组织为一组代表说话人视觉特征空间的“原型”，使模型在遇到新说话人时，可自动从记忆库中调用类似模式进行对比与补偿，从而提升跨个体的泛化能力。

考虑到视觉说话模式在不同说话人之间具有明显的离散分布特性，直接采用原始的说话人嵌入向量作为记忆条目将导致系统资源消耗巨大，且难以泛化到未见个体。本文采用聚类机制，将训练集中所有说话人的特征嵌入进行离线聚类，划分为若干类别，每一类代表一组在视觉属性上具有相似性的说话人。每一类的中心即为该原型类的表示，记忆库即由这些原型组成。优势表现在：1）显著压缩内存使用，避免随着训练说话人数的增加导致记忆规模不可控；2）通过保留代表性的类中心嵌入，提升原型的覆盖性与抽象表达能力；3）避免过拟合于特定个体特征，为未见说话人提供了与历史原型比对的可能，从而提升泛化能力。

具体而言，本文对所有训练说话人的动态视觉说话模式和静态视觉模式进行编码，并分别采用K-means聚类算法，将高维特征空间划分为D_M个聚类，每个聚类的中心点作为该类模式的典型代表。以构建静态视觉模式记忆库$\boldsymbol{D}=\{D_1,D_2,\cdots,D_{D_M}\}$为例，其中每个原型$D_i$通过K-means聚类得到，具体为

$$D_i=\frac{1}{\left|C_i\right|}\sum_{z_s\in C_i} z_s \tag{3}$$

式中，C_i表示第i个聚类。动态视觉说话模式记忆库采用相同的构建方式，既保证了特征空间的代表性，也提升了存储效率。动态记忆库与静态记忆库的构建在训练阶段离线完成，最终得到的两个记忆库将分别在后续的唇语识别过程中与提示学习结合，自适应未见过的说话人。

1.4 表示空间融合

为实现对未见说话人的高效自适应，本文将记忆模块中的说话人原型视为可调用的“提示”，并提出一种基于提示学习的表示空间融合机制。该机制在训练与推理阶段均与Conformer编码器的注意力模块进行交互，通过显式引入结构化先验知识，引导模型动态感知并补偿个体差异，从而提升跨说话人场景下的泛化能力。

图1中下部蓝色区域和中上部红色区域分别展示了基于静态视觉模式和动态视觉说话模式的提示方法的基本原理，及其与Conformer编码器中注意力机制交互的方式，详细算法原理如图2所示。

在跨说话人唇语识别中，由于说话人之间存在显著的个体差异（如口型、发音习惯和面部结构等），传统模型往往难以有效泛化到未见过的说话人。为解决这一问题，一种常用做法是采用基于大规模数据的预训练或特征归一化技术，但往往导致说话人特定信息的丢失，从而限制了模型的性能上限。本文受认知科学中人类记忆机制的启发，如在识别新面孔或理解新口音时，会不自觉地与已知模式进行匹配和联想。基于此，本文将说话人自适应建模为一个提示学习任务：即通过从历史经验中检索相似模式，并将其作为外部提示显式注入模型，使模型能够在不需要额外微调的情况下快速适应未见过的说话人，不仅提高了模型的泛化能力，还保留了说话人特定的语音特征，实现了在精度和泛化性之间的平衡。

从图2可以看出，记忆模块在模型中以统一形式接入注意力机制，无论输入为静态视觉模式还是动态视觉说话模式，都会经历相同的转换与交互过程。为了更清晰地展示具体机制，本文以静态视觉模式为例，说明记忆提示向量如何构建并融合到模型中。

将记忆库中的说话人原型视为一组可复用的提示模板。对于静态视觉模式记忆库$\boldsymbol{D}=\{D_1,D_2,\cdots,D_{D_M}\}$，首先通过线性变换将其投影到与注意力机制相匹配的表示空间，具体为

$$M_K=Concat(\boldsymbol{W}_{MK}D_1,\boldsymbol{W}_{MK}D_2,\cdots,\boldsymbol{W}_{MK}D_{D_M}) \tag{4}$$

$$M_V=Concat(\boldsymbol{W}_{MV}D_1,\boldsymbol{W}_{MV}D_2,\cdots,\boldsymbol{W}_{MV}D_{D_M}) \tag{5}$$

式中，$\boldsymbol{W}_{MK}$和$\boldsymbol{W}_{MV}$为可学习的投影矩阵，二者将静态视觉模式分别映射为键和值向量，确保记忆特征能

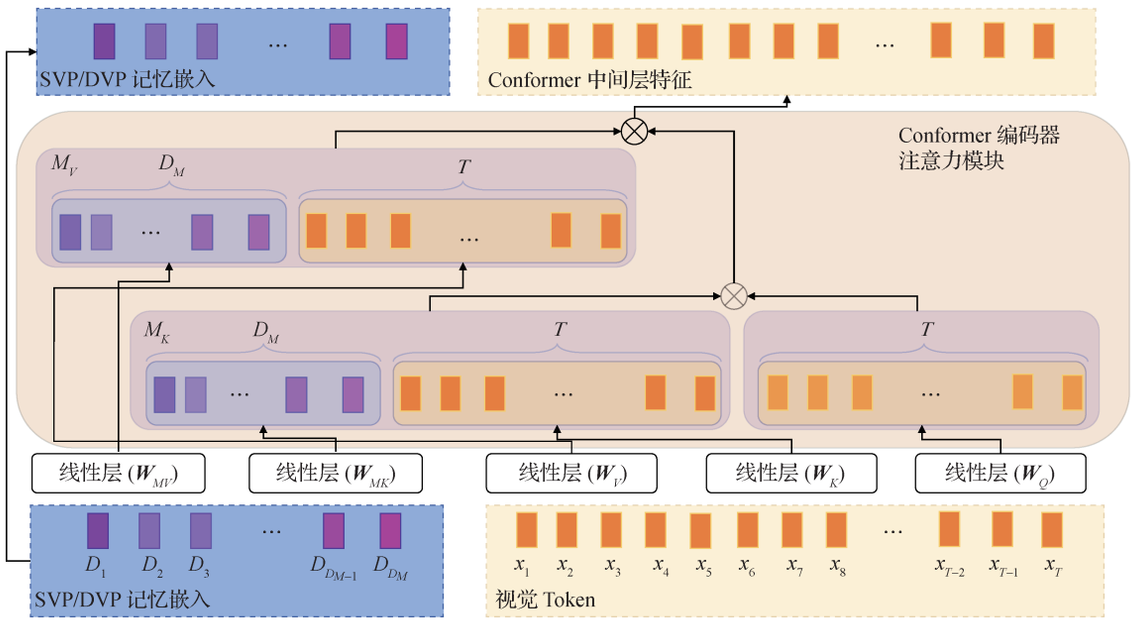


图2　记忆模块与注意力融合示意图

Fig. 2　Illustration of memory module and attention integration

够与当前输入特征在同一表示空间中进行交互，类似于神经科学中的模式完成机制，使系统能够基于部分信息激活相关的完整记忆表示。接着，将记忆键值对与原始输入键值进行拼接，形成扩展的注意力输入，具体为

$$\boldsymbol{K}_m = [K_x; M_k], \boldsymbol{V}_m = [V_x; M_v] \tag{6}$$

记忆增强的注意力计算最终表示为 Attention($\boldsymbol{Q}_x, \boldsymbol{K}_m, \boldsymbol{V}_m$)，其中 $\boldsymbol{K}_m$ 和 $\boldsymbol{V}_m$ 分别为拼接后的键和值向量，表示融合了记忆特征的键和值。$\boldsymbol{Q}_x\boldsymbol{K}_m^{\mathrm{T}}$ 的维度为 $T \times (D_M + T)$，表示输入查询与扩展键的相似度矩阵，经过 softmax 归一化后生成注意力权重。这些权重与 $(D_M + T) \times d_x$ 的 $\boldsymbol{V}_m$ 矩阵相乘，最终的输出维度还是 $T \times d_k$，与输入的 $\boldsymbol{X}$ 维度保持一致，确保了模型架构的兼容性和计算效率。

1.5　自适应模式交互

为了实现静态模式与动态模式的有效融合，本文提出了一种基于分层机制的静态模式与动态模式交互方法，即通过在模型的浅层和深层分别引入静态视觉模式和动态视觉说话模式，利用不同层级模式的优势进行信息融合，如图1。

设编码器总层数为 N，浅层与深层的分界层为 $N/2$。记忆模块中的静态视觉模式被接入至唇语识别模块中基于 Conformer 编码器的浅层注意力模块中，而记忆模块中的动态视觉说话模式则被引入至基于 Conformer 的编码器的深层注意力模块中。浅层特征通常不具备复杂的时序建模能力，但对于静态外观模式的识别较为有效，因此模型在浅层捕捉面部结构、肤色等静态信息，并为后续深层的时序建模打下基础。而深层具备更强的时序建模能力，动态视觉说话模式在此处能够帮助模型识别个体发音方式、语速与口型变化路径等细粒度行为模式。该分层接入策略使得提示特征能够充分发挥其层级作用，有效协同模型表征不同维度的说话人信息。

通过这种分层机制，模型在浅层强化静态建模，在深层加强动态建模，使得两类模式在各自最合适的语义层面中协同作用。最终，模型在处理未见说话人时，能够根据输入中所包含的不同维度提示特征实现快速而精准的适应，从而显著提升其在跨说话人场景下的鲁棒性与实用性。

1.6　唇语识别损失函数

本文模型利用连接时序分类（connectionist temporal classification，CTC）和交叉熵（cross entropy，CE）联合训练策略，总损失函数定义为

$$L_{\mathrm{VSR}} = \lambda L_{\mathrm{CTC}} + L_{\mathrm{CE}} \tag{7}$$

式中，L_{CTC} 为 CTC 损失，L_{CE} 为 Transformer 的自回归交叉熵损失，超参数 λ 取值为0.1。

2　LRS2-ID 数据集

为了验证本文方法对说话人自适应唇语识别问

题的效果，本文在唇语识别数据集LRS2（Afouras等，2022）基础上构造了一个专门针对说话人自适应调整的数据集：LRS2-ID。对LRS2的样本通过采用ArcFace提取人脸特征，然后利用DBSCAN（density-based spatial clustering of applications with noise）算法进行聚类，获得说话人标签，然后通过剔除训练集中出现的说话人，重新划分训练集与测试集。具体而言，LRS2-ID数据集通过剔除训练集中出现的说话人，并重新划分训练集和测试集，使最后的测试集中的说话人与训练集中的说话人没有交集，从而模拟了“零样本适应”的场景。在这种设置下，模型必须应对未见过的说话人，从而可以评测模型在跨说话人场景下的泛化能力。这种划分方式与原始LRS2数据集的划分方式不同，原始划分在训练集与测试集中可能包含了相同的说话人，会使模型在测试阶段容易识别已见说话人。本文构建的LRS2-ID数据集不仅保持了原始语料的多样性，还引入了面向说话人身份控制的标注方式，可以更好地评估模型在面对全新说话人时的识别能力。数据集信息如表1所示。为进一步分析说话人重叠情况，表2给出了测试集与训练集、验证集中的说话人的重叠分布信息。可以看出，测试集大部分说话人未在验证集和训练集中出现，其中超过93%的说话人从未出现在验证集中、62%以上未出现在训练集中，为说话人自适应与泛化能力的评估提供了良好基础。

表1　LRS2-ID数据集信息

Table 1　LRS2-ID dataset information

数据集	时长/h				说话人数			
	预训练	训练	验证	测试	预训练	训练	验证	测试
LRS2	195	29	–	0.5	–	–	–	–
LRS2-ID	194.92	27.79	0.66	0.57	15 271	7 411	297	456

注：“–”表示空。

为了评估说话人自适应机制，本文在测试集中划分并构建了一个代表性的说话人子集。即验证集（适应集）与测试集分别从原始数据集的训练集、验证集和测试集中选取，样本长度统一且较短，文本标签经过人工校对。为确保测试集中的说话人不会出现在训练集和预训练集中，对预训练集中的部分样本进行了剔除。选取10位在测试集中具备一定数据量的说话人，如表3所示，为每位说话人配置约3 min的适应集。该设置模拟了用户提供短时唇动片段即可快速完成模型适配的实际需求，如便携设备上的本地个性化服务。同时还对预训练集中该说话人的重叠数据进行了统计，用于分析预训练先验对自适应性能的影响。

表2　LRS2-ID测试集与其他集合说话人的重叠情况

Table 2　Overlap of speakers in the LRS2-ID test dataset with other datasets

数据集		测试集	
		说话人数	时长/h
验证集	未出现	426（93.42%）	0.485（85.09%）
	出现	30（6.58%）	0.085（14.91%）
训练集	未出现	286（62.72%）	0.190（33.33%）
	出现	170（37.28%）	0.380（66.67%）

3　实验与分析

3.1　实验设置

本文旨在探究并提示学习与显式记忆机制在未见目标说话人条件下的唇语识别泛化能力。因此，在实验过程中不使用目标说话人适应数据进行模型适应，而是直接对未见目标人进行测试。

在评价指标方面，采用词错误率（word error rate，WER）为评价指标，其定义为识别错误的词数（包括替换、插入和删除）与总词数的比率，计算式为

$$f_{\mathrm{WER}}=\frac{S+D+I}{Num}\times 100\% \tag{8}$$

式中，S为替换错误数（substitutions），即识别结果中被错误替换的词数；D为删除错误数（deletions），即参考文本中有但识别结果中被遗漏的词数；I为插入错误数（insertions），即识别结果中多出的、参考文本中没有的词数；Num为参考文本中的总词数。WER值越低表示模型性能越好，能够直观反映模型对说话人内容的准确理解能力（Zhang等，2024）。

采用Conformer（Gulati等，2020）作为基础编码器，该架构在卷积神经网络的高效局部特征提取与Transformer（Vaswani等，2017）的强大全局依赖建模能力之间取得了良好平衡。相比于纯Transformer架构，Conformer能够更有效地捕捉唇部序列中的局部时空模式（如口型细微变化）和长程时序依赖（如发音节奏），对于准确建模个体发音特性至关重要。同

表3　LRS2-ID数据集子集的说话人样本数量及时长对比

Table 3　Number of speaker samples and duration in different sets of the LRS2-ID dataset

ID	测试集		适应集		剔除前预训练集中的重叠部分	
	时长/min	样本数量	时长/min	样本数量	时长/min	样本数量
1（#42）	1.78	45	3	69	21.47	184
2（#145）	2.14	47	3	77	36.89	242
3（#504）	6.10	130	3.01	71	3.17	21
4（#1077）	3.97	74	3	55	0	0
5（#1596）	0.40	7	3.02	75	7.85	58
6（#1713）	0.37	10	3.02	75	7.30	51
7（#2597）	0.36	10	3.01	84	3.87	21
8（#2624）	12.08	272	3.01	72	0.11	1
9（#3112）	2.07	49	3.01	72	18.91	150
10（#4103）	5.82	122	3.02	61	21.34	180
总计	35.09	766	30.10	711	120.91	908

时，其分层结构设计与本文提出的分层提示注入机制高度契合，便于在浅层和深层分别引入静态与动态视觉模式提示。

所有实验均在LRS2-ID数据集上进行。模型的输入为96×96像素分辨率的唇部区域图像序列；训练过程中，对输入图像帧使用水平翻转与随机裁剪至88×88像素策略，以增强模型鲁棒性。考虑到输入序列长度往往存在差异，训练过程中采用动态批次（dynamic batching）策略，并将每个batch的总帧数控制在不超过1 800帧的上限内，以兼顾训练效率与内存利用。输入的序列帧率与原始视频帧率保持一致，从而确保时序层面建模的一致性。

在训练策略方面，本文引入课程学习（curriculum learning）机制，即按照由简到难的顺序组织训练样本，逐步提升模型的学习与泛化能力。即训练初期优先使用长度较短（小于4 s）的语句作为训练样本，在验证集上的WER趋于稳定（即收敛）后，再引入所有时长的句子进行完整训练。该策略有助于缓解长句子序列对模型优化过程造成的不稳定影响，同时也使模型能够更自然地适应唇语序列中复杂的时序模式。这一分阶段训练方式在唇语识别中较为常见（Ma等，2023），尤其适用于提升模型对长句、口型变化剧烈语句的建模能力。

本文模型在6张RTX 4090显卡上训练，采用Adam优化器，最大学习率设为1×10^{-3}，并结合One-Cycle学习率调度策略，以实现更快的收敛与更强的泛化能力。总训练轮数设定为75个epoch。此外，为提高模型在推理阶段的稳定性与性能表现，避免单一epoch模型的性能偏差，本文采用模型平均策略，即取最后3个epoch的模型参数的平均值进行最终模型的测试。

3.2　方法对比

为了验证本文方法对说话人自适应唇语识别任务的效果，本文对比了5种唇语识别方法：TM-Seq2Seq（Afouras等，2022）、CM-Seq2Seq（Ma等，2021）、AutoAVSR（Ma等，2023）、DVPs和SVPs（Luo等，2025），如表4所示。由表4可知，本文方法在跨说话人场景下达到了最优性能，词错误率（WER）为46.99%，相较当前最佳的说话人自适应方法DVPs（48.25%）提升了2.61%，证明了记忆机制与显式提示引导策略在建模说话人个体差异方面的优势。其次，说话人自适应方法整体优于通用唇语识别模型。SVPs、DVPs和本文方法均显著超越了通用唇语识别基线AutoAVSR（49.08%），表明专门针对说话人差异设计的自适应机制对提升模型泛化能力至关重要。

3.3　消融实验

3.3.1　课程学习策略与输入区域影响

为系统性评估不同说话人建模策略对唇语识别性能的影响，本文逐步评测引入输入区域限制、静态

表 4　不同方法在 LRS2-ID 数据集上的 WER

Table 4　WER of different methods on the LRS2-ID dataset

方法	年份	WER/%
TM-Seq2Seq	2018	57.02
CM-Seq2Seq	2021	53.53
AutoAVSR	2023	49.08
SVPs	2025	48.53
DVPs	2025	48.25
本文	2025	**46.99**

注：加粗字体表示最优结果。

提示、动态提示与记忆机制的效果，以从多个维度分析不同组件对模型泛化能力的贡献，如表 5 所示。

设置 a）中以完整人脸区域代替唇部区域作为输入，在 LRS2-ID 训练数据上完成学习过程后，得到的词错误率高达 76. 94%，凸显了使用全脸图像可能过多引入个体表观差异信息干扰的风险，尤其是静态外观因素（如肤色、发型、面部结构）会干扰模型对动态说话过程所表达的语义内容的建模，进而降低了模型对未见说话人的泛化能力。

需要指出的是，本文所提出的方法并非完全屏蔽人脸信息。在说话人建模阶段显式引入了全脸区域的信息提取模块，用于构建动态视觉说话模式和静态视觉模式，从而实现更全面的个体特征建模。因此，设置 a）不仅用于验证“是否使用人脸信息”所带来的影响，更关键的是作为控制变量，用于验证在本文整体框架下引入人脸编码模块是否真正通过本文的结构设计实现了有效利用，而非依赖“输入维度增加”带来的表面性能提升。该实验也从侧面说明，未经建模的原始人脸信息不仅无助于内容建模，反而可能引入冗余干扰，进一步突出了本文所提出的“模式分离建模策略”在提升模型性能与鲁棒性方面的必要性与有效性。

为了排除上述冗余表观信息对唇语任务的影响，设置 b）开始限定输入区域为唇部图像，并使用小于 4 s 的片段进行预训练。该设置将词错误率大幅降低至 51. 24%，显著优于设置 a）的全脸输入。这一结果验证了在说话人信息缺失的情形下，仅关注唇部区域更有助于模型从输入中学习到与语言内容直接相关的视觉说话模式，同时减少无关的静态特征的干扰。此外，该设置使用长度较短的 4 s 语句作为训练样本，使模型具备初步识别能力，为后续使用完整句子训练奠定了稳定的基础。该策略有助于缓解长序列对模型优化过程造成的不稳定影响，同时也使模型能够更自然地适应唇语序列中复杂的时序模式。

在设置 c）中，本文使用更长的 < 24 s 视频片段继续训练模型，以更贴近实际应用中完整语句的时长分布。该设置加载设置 b）的模型权重作为起始，遵从课程学习的训练策略：从较短、较简单的任务逐步过渡到更复杂的任务。从结果来看，该策略进一步提升了模型的识别性能，词错误率降低至 49. 08%，说明更长的上下文可以为模型提供更多语义连续性和发音模式，从而提升语言建模效果。同时，这一阶段的训练也构成了本文后续说话人自适应方法的通用基线（即本文方法“未引入任何个性提示或记忆机制”时的效果），为后续静态提示、动态提

表 5　在 LRS2-ID 数据集下的消融实验

Table 5　Ablation study on the LRS2-ID dataset

/%

设置	训练数据	加载权重	记忆机制	静态特征	动态特征	WER ↓	Δ WER ↑
a）	< 4 s 切片面部	无	×	×	×	76.94	0
b）	< 4 s 切片唇部	无	×	×	×	51.24	+25.70
c）	< 24 s 切片唇部	b）	×	×	×	49.08	+27.86
d）	< 24 s 切片唇部	b）	×	√	×	48.53	+28.41
e）	< 24 s 切片唇部	b）	×	×	√	48.25	+28.69
f）	< 24 s 切片唇部	b）	√	√	×	47.87	+29.07
g）	< 24 s 切片唇部	b）	√	√	√	**46.99**	**+29.95**

注：加粗字体表示各列最优结果；“√”和“×”分别表示使用和未使用对应模块；“↓”表示值越小越好，“↑”表示值越大越好。

示与记忆提示机制的验证提供了统一的比较基准。

3.3.2　记忆提示的引入与动静融合优势

引入静态视觉特征作为辅助信息形式与原始唇部帧级特征进行拼接融合直接送入编码器(实验结果见表5设置d))较于基线模型有了轻微提升,识别错误率(WER)从49.08%降到了48.53%,平均提升幅度为1.11%。尽管静态特征方法在某些情况下能够提供一定的性能改善,但其提升的幅度相对较小,表明在跨说话人适应方面,静态特征的表现有其局限性。相较之下,设置e)中加入动态视觉说话模式展现了明显更大的性能提升,与基线设置c)比较,总体提升幅度达到1.69%,表明动态特征能够更好地捕捉和利用说话人特有的面部动态信息,从而在不同说话人之间提供显著的性能提升;动态特征对于个体差异的适应性更强,能够显著提高模型的整体性能,尤其是在面对较大说话人差异时,动态特征的优势尤为突出。

静态模式与动态模式分别从面部结构和发音行为两个角度建模说话人个体性特征,具有天然的互补性。下面进一步验证对两类提示特征的融合方法的效果。具体而言,设置f)与设置g)分别在单独静态视觉模式与联合动态视觉说话模式基础上引入记忆提示模块,构建跨说话人的原型记忆空间。设置f)仅结合静态视觉模式与记忆提示机制;设置g)则融合静态与动态两类提示特征,并同时使用对应的双记忆库进行辅助提示学习。由表5可见,设置f)在设置d)的基础上引入记忆机制后,词错误率进一步下降至47.87%,说明记忆机制的引入有助于模型相对更稳定地应对未见说话人。更进一步地,设置g)融合静态视觉模式与动态视觉说话模式后,在记忆机制引导下模型取得了本节最低词错误率(46.99%),相比设置c)(无提示)整体降低约4.3%,验证了本文提出的提示与记忆机制协同建模在说话人适应场景下的有效性。

该性能的提升并非因为简单加入了更多的信息(如静态或动态特征),而是因为这些信息是通过本文设定的方式进行了有效的组织和融合:一方面,静态与动态提示特征通过对当前样本进行个体化建模,使模型能够快速捕捉未见说话人的模式;另一方面,记忆机制通过引入训练阶段聚类形成的原型表示,提供了对说话人空间的全局建模能力,在推理阶段起到结构性约束作用,从而进一步提升了模型的泛化能力。

LRS2-ID数据集子集逐个说话人实验结果柱状图如图3所示。从图3可以看到,大多数说话人均受益于f)(紫色柱)与g)(棕色柱)设置,且g)在多个说话人上的词错误率显著优于前几项设置,呈现出较强的一致性,再次验证了静态视觉模式和动态视觉说话模式在多说话人建模中的协同优势,以及记忆机制在未见说话人识别中的引导作用。

综上,本文提出的融合静态视觉模式与动态视觉说话模式的记忆引导机制,在无适应数据的前提下有效提升了模型的个体泛化能力,凸显出提示学习与记忆机制在唇语识别中协同建模的重要性。

3.3.3　特征可视化

为验证记忆模块中说话人原型特征的有效性,

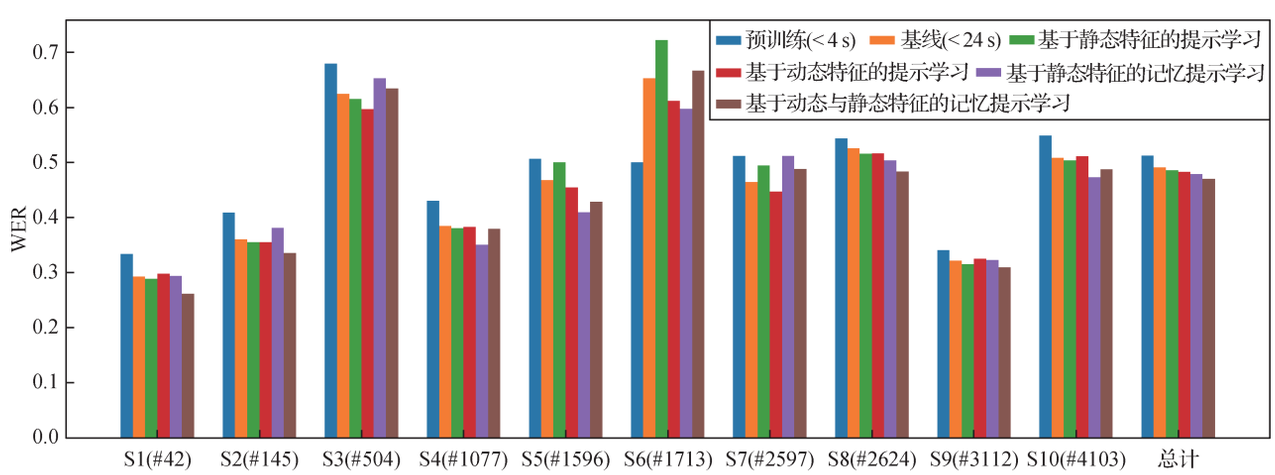


图3　LRS2-ID数据集子集逐个说话人实验结果

Fig. 3　Per-speaker experimental results on the LRS2-ID dataset

本文通过 t-SNE（t-distributed stochastic neighbor embedding）对聚类后的特征分布进行可视化分析（van der Maaten 和 Hinton，2008）。以静态视觉模式为例，图 4 为 100 个说话人特征聚类 t-SNE 可视化。

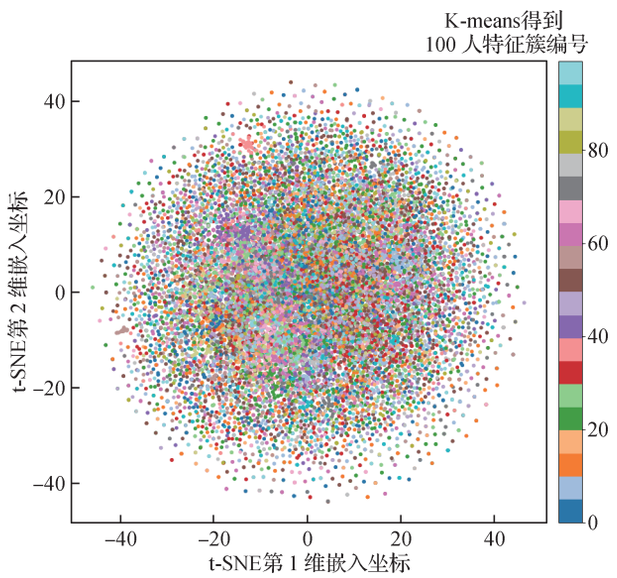


图 4　说话人特征聚类 t-SNE 可视化

Fig. 4　t-SNE visualization of speaker feature clustering

从图 4 可以看出，说话人原型在二维特征空间中呈现出良好的聚类结构和分离性，表明 SVP 特征能够有效区分不同说话人的个体差异。特征点在空间中形成连续分布且局部区域密集聚集，表明记忆库构建的合理性：各原型具有良好的代表性，能够覆盖主要的说话人模式；同时连续分布特性支持跨说话人泛化，使模型可通过检索相邻原型组合生成个性化提示。综上所述，特征空间可视化直观地表明了本文方法在说话人模式建模方面的有效性，为记忆机制和自适应方法提供了可视化依据。

3.3.4　不同说话人效果分析

图 3 和表 6 展示了在 LRS2-ID 测试集中 10 位未在训练集中出现过的说话人上的词错误率具体表现。从整体趋势来看，本文所提出的融合静态视觉模式与动态视觉说话模式的记忆提示方法（设置 g））在大多数说话人上均取得性能提升，尤其在个体差异显著的说话人上表现尤为突出。

为进一步分析提示特征的作用机制，图 5 展示了两位典型说话人的样本帧，分别为 S1（#42）与 S5（#1596）。从图 3 和表 6 可以看出，这两位说话人分别对应在静态视觉模式与动态视觉说话模式下性能提升最显著的两个案例。

S1（#42）是一个受益于动态特征提示的典型例子。如图 5（a）所示，该说话人在视频中呈现出较强的视觉外貌变化——包括多套服装、背景环境差异及光照条件变化等。由于静态视觉模式（如肤色、面部结构）在不同视频段落中存在较大变异，模型在仅使用静态提示（设置 d））时的适应效果受到限制（词错误率提升至 29.36%）。相比之下，动态视觉说话模

表 6　不同说话人在记忆提示学习方法下的 WER 结果比较

Table 6　Comparison of WER results for different speakers using the memory-prompted learning method

/%

ID	基线	基于 SVP 的记忆提示学习	基于 SVP 与 DVP 说话模式的记忆提示学习
S1（#42）	29.23	29.36（−0.44%）	26.15（+10.54%）
S2（#145）	35.99	38.06（−5.75%）	33.56（+6.75%）
S3（#504）	62.42	65.27（−4.57%）	63.41（−1.59%）
S4（#1077）	38.40	35.02（+8.80%）	37.94（+1.20%）
S5（#1596）	46.75	40.91（+12.49%）	42.86（+8.32%）
S6（#1713）	65.28	59.72（+8.52%）	66.67（−2.13%）
S7（#2597）	46.43	51.19（−10.25%）	48.81（−5.13%）
S8（#2624）	52.57	50.33（+4.26%）	48.31（+8.10%）
S9（#3112）	32.11	32.23（−0.37%）	30.88（+3.83%）
S10（#4103）	50.79	47.28（+6.91%）	48.77（+3.98%）
总计	49.08	47.87（+2.47%）	46.99（+4.26%）

注：括号里的“+/−”表示相对基线的 WER 提升/下降的相对百分比。

式更侧重于捕捉说话风格、发音节奏与口型变化路径等时间信息，受外貌因素影响较小。因此该说话人在融合动态提示后（设置e））取得显著性能提升，最终在静态与动态提示及记忆机制联合使用（设置g））后，词错误率降至26.15%，实现相对提升约10.5%。

S5（#1596）则是静态提示更有效的代表，该说话人整体视觉表现稳定，图5（b）展示了该说话人在不同说话视频中的样本帧。可以看到，该说话人在不同片段中外貌变化极小，始终保持类似的头部姿态、背景与服装特征（如大光头和统一色调）。在这种外观一致性高的场景下，静态视觉模式更容易精准建模其个体身份，从而为模型提供强有力的先验信息。在该设置下，词错误率从46.75%降至40.91%，获得超过12%的性能相对提升。相比之下，动态提示对于该说话人作用相对有限，表现略弱。



(a) S1 (#42) 说话人样本　　(b) S5 (#1596) 说话人样本

图5　LRS2-ID测试集说话人样本示例

Fig. 5　Speaker samples from the LRS2-ID test set ((a) speaker sample S1 (#42); (b) speaker sample S5 (#1596))

3.4　计算效率分析

为评估本文方法的实用价值，对模型的计算效率进行了分析。与基线模型相比，本文方法在保持高性能的同时仅引入轻微的计算开销。

1）内存占用方面。静态与动态记忆库（各100个原型）在推理前被预先加载为固定数据结构，仅占用约数百KB的显存。记忆提示投影矩阵$\boldsymbol{W}_{MK}$和$\boldsymbol{W}_{MV}$是仅有的新增参数，新增参数量约为0.1 M，相对于主流唇语识别模型（参数量通常为30 M～50 M），参数量增长低于0.5%。

2）推理速度方面。计算输入序列与记忆库中100个原型的相似度的计算量极小，因其基于预提取的紧凑特征（SVP/DVP向量），而非原始视频帧。在Conformer编码器的注意力模块中引入提示向量，会将注意力计算复杂度从$O(T^2)$略微增至$O(T(T+D_M))$。实测结果表明，相对于基线模型，完整方法的推理速度仅下降约5%～8%，仍能保持实时处理能力（>140 帧/s）。

与需要执行完整模型微调的传统自适应方法相比（Kim等，2022），本文方法以极小计算代价换取显著性能提升的设计，实现了“即插即用”式的自适应，在推理效率上具有显著优势。

4　结　论

本文提出了一种基于显式记忆提示机制的说话人自适应方法，旨在解决唇语识别任务中模型对未见说话人泛化能力不足的问题。通过构建静态视觉模式与动态视觉说话模式两类原型记忆库，模型检索最相似的提示模板，生成个性化提示向量，并将其按语义层级显式注入Conformer编码器的注意力模块：浅层注入静态提示以感知面部外观；深层注入动

态提示以建模发音节奏。该分层提示注入机制实现了对个体差异的精细化引导，在无需微调的情况下完成快速适应。大量实验结果表明，该方法在完全无适应数据的条件下可显著提升模型对未见说话人的识别性能，且在个体差异显著或外观变化复杂的场景中具备良好的稳定性和普适性，从而为提升唇语识别系统的跨个体泛化能力提供了有效思路和实践路径。

本文方法作为一种通用自适应范式，其核心思想具备良好的可迁移性。该方法不局限于唇语识别任务，其通过构建先验记忆库并利用提示引导模型快速适应未知样本的机制，可扩展至其他视觉语音任务，如语音驱动面部动画、音视频语音识别等，为相关领域的个性化建模研究提供了新的技术路径。

尽管本文说话人自适应的唇语识别方法中取得了一定进展，但仍存在一些有待进一步探索的方向。首先，目前方法仅基于单一视觉模态进行建模，未来可探索多模态协同建模的潜力，如研究如何利用少量音频数据作为辅助监督信号，通过跨模态对比学习来增强视觉说话人特征的判别力（Yeo 等，2024）；或构建一个视听联合的记忆原型空间，从而更鲁棒地感知说话人身份与发音内容。其次，当前的说话人适应机制主要集中在全局身份级别的调控，未来可向更细粒度的个性化建模深化，如探索对语速、地域口音、个人嘴型习惯等维度的解耦与自适应建模，使系统不仅能识别“谁在说话”，更能理解“如何说话”，从而在个性化语音合成与识别等任务中发挥更大价值。最后，推动方法走向实际应用也至关重要，包括研究更高效的记忆库压缩与检索策略以降低计算开销，以及探索面向真实场景的在线自适应机制，使模型能够根据用户交互动态优化，最终构建出更具实用性的唇语识别系统。

综上所述，本文从记忆建模的角度出发，为提升唇语识别系统在无适应数据条件下的跨说话人泛化能力提供了一种可行路径。未来工作将在多模态融合、更细粒度的个性化建模等方面继续拓展，进一步完善唇语识别系统在复杂多变说话人环境下的适应性能和实际可用性。

参考文献（References）

Afouras T, Chung J S, Senior A, Vinyals O and Zisserman A. 2022. Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44 (12): 8717-8727 [DOI: 10.1109/TPAMI.2018.2889052]

Cui X Y, He C, Zhao H K and Wang M L. 2024. Combining ViT with contrastive learning for facial expression recognition. Journal of Image and Graphics, 29(1): 123-133 (崔鑫宇, 何翀, 赵宏珂, 王美丽. 2024. 融合 ViT 与对比学习的面部表情识别. 中国图象图形学报, 29(1): 123-133) [DOI: 10.11834/jig.230043]

Deng J K, Guo J, Xue N N and Zafeiriou S. 2019. ArcFace: additive angular margin loss for deep face recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4685-4694 [DOI: 10.1109/CVPR.2019.00482]

Gulati A, Qin J, Chiu C C, Parmar N, Zhang Y, Yu J H, et al. 2005. Conformer: convolution-augmented transformer for speech recognition [2025-11-17]. https://arxiv.org/pdf/2005.08100.pdf

Huang Y Y. 2024. Lipreading Algorithms for Precision, Unseen Speaker Robustness, and Low-Resource Languages Adaptability. Xi'an: Xidian University (黄奕洋. 2024. 高精度、未知说话人鲁棒和低资源语种适应的唇语识别算法研究. 西安: 西安电子科技大学)

Kandala P A, Thanda A, Margam D K, Aralikatti R C, Sharma T, Roy S, et al. 2019. Speaker adaptation for lip-reading using visual identity vectors//Proceedings of Interspeech 2019. Graz, Austria: ISCA: 2758-2762 [DOI: 10.21437/Interspeech.2019-3237]

Kim M, Kim H and Ro Y M. 2022. Speaker-adaptive lip reading with user-dependent padding//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 576-593 [DOI: 10.1007/978-3-031-20059-5_33]

Kim M, Kim H I and Ro Y M. 2025. Prompt tuning of deep neural networks for speaker-adaptive visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47 (2): 1042-1055 [DOI: 10.1109/TPAMI.2024.3484658]

Kim T, Song I and Bengio Y. 2017. Dynamic layer normalization for adaptive neural acoustic modeling in speech recognition//Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA: 2411-2415 [DOI: 10.21437/Interspeech.2017-556]

Li F C, Gao S S, Liu Z, Zhang C M and Zhou Y F. 2025. Multimodal medical image fusion with progressive feature extraction and frequency domain information complementation. Journal of Image and Graphics, 30(5): 1510-1527 (李夫辰, 高珊珊, 刘峥, 张彩明, 周元峰. 2025. 渐进特征提取和频域信息补充的多模态医学图像融合. 中国图象图形学报, 30(5): 1510-1527) [DOI: 10.11834/jig.240509]

Li Y, Xue F, Li S J, Zhang J R, Yang S, Guo D, et al. 2025. Learning speaker-invariant visual features for lipreading [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2506.07572.pdf

Luo S T, Yang S, Shan S G and Chen X L. 2023. Learning separable hidden unit contributions for speaker-adaptive visual speech recog-

nition//Proceedings of the 34th British Machine Vision Conference. Aberdeen, Britain: BMVA

Luo S T, Yang S, Shan S G and Chen X L. 2025. Dynamic visual speaking patterns: you are the way you speak//Proceedings of the 19th IEEE International Conference on Automatic Face and Gesture Recognition. Tampa/Clearwater, USA: IEEE: 1-11 [DOI: 10.1109/FG61629.2025.11099168]

Ma P C, Haliassos A, Fernandez-Lopez A, Chen H L, Petridis S and Pantic M. 2023. Auto-AVSR: audio-visual speech recognition with automatic labels//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/ICASSP49357.2023.10096889]

Ma P C, Petridis S and Pantic M. 2021. End-to-end audio-visual speech recognition with conformers//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, Canada: IEEE: 7613-7617 [DOI: 10.1109/ICASSP39728.2021.9414567]

van der Maaten L and Hinton G. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, 9(86): 2579-2605

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017, #30

Wang T Y, Yang S, Shan S G and Chen X L. 2025. GLip: a global-local integrated progressive framework for robust visual speech recognition//Proceedings of the 36th British Machine Vision Conference. Sheffield, Britain: BMVC

Yang C Z, Wang S L, Zhang X X and Zhu Y. 2020. Speaker-independent lipreading with limited data//Proceedings of 2020 IEEE International Conference on Image Processing. Abu Dhabi, United Arab Emirates: IEEE: 2181-2185 [DOI: 10.1109/ICIP40778.2020.9190780]

Yeo J H, Kim C W, Kim H, Rha H, Han S, Cheng W H, et al. 2025. Personalized lip reading: adapting to your unique lip movements with vision and language//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI: #33026 [DOI: 10.1609/aaai.v39i9.33026]

Yeo J H, Kim M, Choi J, Kim D H and Ro Y M. 2024. AKVSR: audio knowledge empowered visual speech recognition by compressing audio knowledge of a pretrained model. IEEE Transactions on Multimedia, 26: 6462-6474 [DOI: 10.1109/TMM.2024.3352388]

Zhang Q, Wang S L and Chen G L. 2021. Speaker-independent lipreading by disentangled representation learning//Proceedings of 2021 IEEE International Conference on Image Processing. Anchorage, USA: IEEE: 2493-2497 [DOI: 10.1109/ICIP42928.2021.9506396]

Zhang Y H, Yang S, Shan S G and Chen X L. 2024. ES^3: evolving self-supervised learning of robust audio-visual speech representations//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 27059-27069 [DOI: 10.1109/CVPR52733.2024.02556]

Zhao Y Z, Ni C J, Leung C C, Joty S, Chng E S and Ma B. 2021. A unified speaker adaptation approach for ASR//Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 9339-9349 [DOI: 10.18653/v1/2021.emnlp-main.737]

作者简介

骆嵩涛,男,硕士研究生,主要研究方向为唇语识别。
E-mail: luosongtao18@mails.ucas.ac.cn
杨双,通信作者,女,副研究员,主要研究方向为唇语识别。
E-mail: shuang.yang@ict.ac.cn
王天月,女,博士研究生,主要研究方向为唇语识别。
E-mail: wangtianyue33@163.com
倪群平,男,工程师,主要研究方向为唇语识别。
E-mail: dandan171225@163.com
山世光,男,研究员,主要研究方向为视觉的情感计算和心理健康评估。E-mail: sgshan@ict.ac.cn