智能语音信息处理团队 19 篇论文被语音技术顶会 ICASSP 2026 接收

近日，ICASSP 2026 会议发出了审稿结果通知，语音及语言信息处理国家工程研究中心智能语音信息处理团队共 19 篇论文被会议接收，论文方向涵盖语音识别、语音合成、话者识别、语音增强、情感识别、声音事件检测等，各接收论文简介见后文。

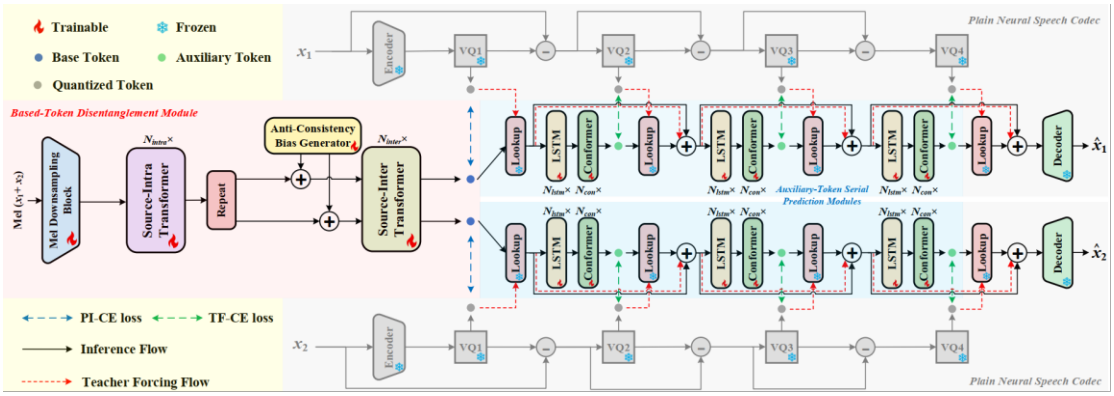
ICASSP （International Conference on Acoustics, Speech, and Signal Processing）是 IEEE 信号处理学会（Signal Processing Society）的学术年会，是全世界规模最大、最全面的声学、语音和信号处理及其应用方面的国际会议，也是语音技术领域最具影响力的顶级国际会议。本届会议以“Where Signals Meet Intelligence”为主题，内容涵盖语音识别、语音合成、语音增强、自然语言处理、机器学习等多个领域。

1. CodeSep: Low-Bitrate Codec-Driven Speech Separation with Base-Token Disentanglement and Auxiliary-Token Serial Prediction

论文作者：杜荟鹏，艾杨，江晓航，郑瑞晨，凌震华

论文单位：中国科学技术大学

论文简介：



本文聚焦语音分离与语音压缩相融合这一全新应用场景，旨在实现多说话人语音分离的同时，生成适用于高效传输与存储的离散表征，该技术可广泛应用于线上会议及对话存档等领域。针对这一应用需求，本文提出一种基于编解码器的联合优化模型 CodeSep，该模型可同时完成语音分离与低码率压缩两项任务。CodeSep 模型由三大核心模块构成：基于残差矢量量化器（residual vector quantizer，RVQ）的基础神经语音编解码器、基础表征序列解耦模块（base-token disentanglement，BTD）以及并行后续表征序列预测模块（auxiliary-token serial prediction，ATSP）。其工作流程如下：首先，基础表征序列解耦模块将混合语音的梅尔频谱分离为各说话人对应的基础表征序列；随后，并行后续表征序列预测模块对上述基础表征序列进行优化，通过序列预测生成后续表征序列；最终，编解码器的解码端会整合所有表征序列，完成分离语音波形的重构。在模型训练阶段，编解码器中的残差矢量量化器会提供监督信号，监督损失由置换不变交叉熵损失与基于教师强制策略的交叉熵损失共同构成。由于实际传输与存储过程中仅需处理基础表征序列，CodeSep 模型能够实现低码率压缩。实验结果表明，相较于现有基线模型，CodeSep 模型在仅 1 kbps 的超低码率下，仍能取得令人满意的语音分离效果。

论文资源：预印版网址 https://arxiv.org/abs/2601.12757

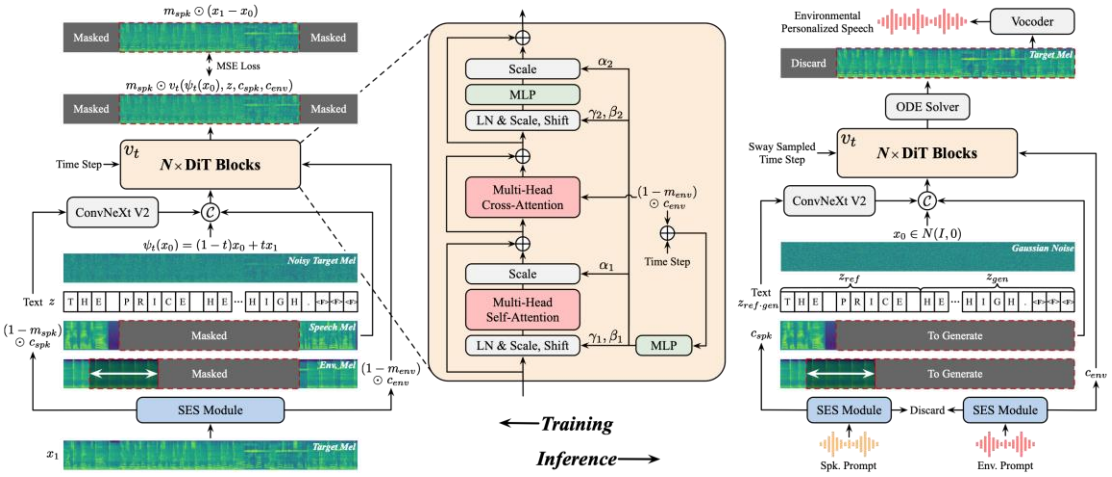
Demo 语音网页 https://redmist328.github.io/CodeSep/

2. DAIEN-TTS: Disentangled Audio Infilling for Environment-aware Text-to-Speech Synthesis

论文作者：鲁叶欣，顾宇，魏坤，杜荟鹏，艾杨，凌震华

论文单位：中国科学技术大学

论文简介：



本文提出了一种基于解耦音频填充的环境感知零样本语音合成模型 DAIEN-TTS，通过引入相互独立的话者提示音频与环境提示音频，实现对合成语音音色与背景环境的独立可控生成。DAIEN-TTS 以基于流匹配（flow matching, FM）的零样本语音合成模型 F5-TTS 作为骨干网络。在训练过程中，首先引入预训练的语音–环境分离（speech-environment separation, SES）模块，将环境语音解耦为干净语音与环境音频对应的梅尔谱表示。随后，对上述两类梅尔谱分别施加不同长度的随机跨度掩膜，以增强模型对推断阶段话者与环境音频提示时长不一致情形的适应能力。两类梅尔谱的未掩膜部分与文本嵌入共同作为条件，通过交叉注意力（cross attention, CA）模块注入至 Diffusion Transformer（DiT）块中，引导模型对原始环境语音频谱中的掩膜区域进行协同填充。在推断过程中，话者提示音频与环境提示音频分别经 SES 模块分离得到干净语音梅尔谱与环境音频梅尔谱，并作为话者条件与环境条件输入模型，使其能够在生成目标话者语音的同时续写时变背景环境。为进一步增强推断阶段的生成可控性，本文对话者条件与环境条件使用双重无分类器引导（dual classifier-free guidance, DCFG），并提出信噪比自适应策略，以保证合成语音与环境音频提示在能量分布上的一致性。实验结果表明，DAIEN-TTS 能够生成具有高自然度与可懂度的环境语音，且在音色与背景环境相似度上与对应的话者提示和环境提示保持高相似度。

论文资源：Demo 语音网页 https://yxlu-0102.github.io/DAIEN-TTS

3. Enhancing Noise Robustness for Neural Speech Codecs through Resource-Efficient Progressive Quantization Perturbation Simulation

论文作者：郑瑞晨，艾杨，杜荟鹏，戴礼荣

论文单位：中国科学技术大学

论文简介：

Algorithm 1 Progressive Probabilistic Top-K Sampling in RVQ

Require: Encoding feature $\mathbf{z}$, N VQ codebooks $\{\mathbf{e}_j^n\}_{n=1}^N$, number of top candidates K
Ensure: Quantized result $\hat{\mathbf{z}}$

for $l = N$ to 1 **do**
  $\hat{\mathbf{z}} = 0$, **residual** $= \mathbf{z}$
  for $n = 1$ to N **do**
    Compute distances $d_j = ||\textbf{residual} - \mathbf{e}_j^n||_2$
    if $n == l$ **then**
      Select top-K:
      $\{m_{c_1}, \ldots, m_{c_K}\} = \text{Top-K}_j \, \|\mathbf{z}_c - \mathbf{e}_j\|_2$
      Compute probabilities: $P_{m_i} = \frac{\exp(-d_{m_i}/\tau)}{\sum_{j=1}^{K} \exp(-d_{m_j}/\tau)}$
      Sample $m \sim \text{Categorical}(P_{m_1}, \ldots, P_{m_K})$
    else
      Select closest: $m = \arg\min_j d_j$
    end if
    $\hat{\mathbf{z}} = \hat{\mathbf{z}} + \mathbf{e}_m^n$, **residual** $=$ **residual** $- \mathbf{e}_m^n$
  end for
  Update gradients for components after l-th VQ.
end for

本文聚焦于神经语音编解码器在噪声环境下的鲁棒性问题。尽管现有的神经语音编解码器在干净语音重构上表现优异，但即便轻微的背景噪声也会导致量化过程产生的码字发生意外偏移（Codeword Shift），从而降低重构语音的质量。针对这一痛点，本文提出了一种无需依赖成对带噪-干净语音数据的资源高效型训练策略，通过直接在量化层模拟噪声扰动来增强语音编解码器模型的鲁棒性。该方法引入了两个核心机制：

1. 概率 Top-K 采样策略（Probabilistic Top-K Sampling）：替代了残差矢量量化（RVQ）中传统的确定性最近邻选择。该策略根据距离加权概率从前 K 个候选码本中进行采样，有效模拟了噪声引起的随机性变化 。
2. 渐进式训练方案（Progressive Training Scheme）：采用课程学习的思想，以可控的方式逐步引入扰动。训练从对音质影响较小的最后一个量化器开始，逐层向前推进至第一个量化器，确保了模型优化的稳定性 。

实验在 Encodec 和 WavTokenizer 两个先进的编解码器上进行验证。结果表明，该策略显著提升了语音编解码器在轻微背景噪声条件下的鲁棒性（例如在 15 dB SNR 环境下，经 Encodec 编解码的语音的 UTMOS 评分从 3.475 提升至 3.586），同时在干净语音上的编解码质量上也取得了提升。

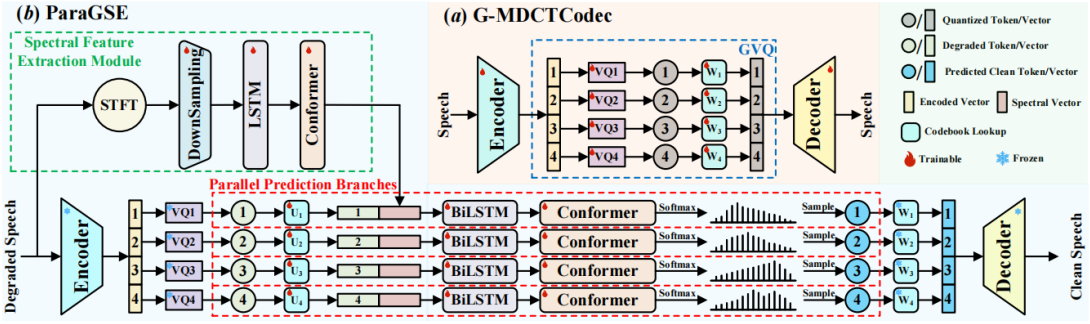
论文资源：预印版网址 https://arxiv.org/abs/2509.19025

4. ParaGSE: Parallel Generative Speech Enhancement with Group-Vector-Quantization-Based neural Speech Codec

论文作者：刘飞，艾杨

论文单位：中国科学技术大学

论文简介：



近年来，生成式语音增强受到广泛关注；然而，现有方法往往受制于模型复杂度过高、效率有限以及语音质量不够理想等问题。为解决这些挑战，本文提出了一种新的并行生成式语音增强（ParaGSE）框架，该框架基于一种采用

组向量量化（GVQ）的神经语音编解码器。该 GVQ 编解码器使用相互独立的多个向量量化器（VQ）生成彼此独立的离散 token，从而使 ParaGSE 能够高效地进行并行 token 预测。具体而言，ParaGSE 利用 GVQ 编解码器将退化语音编码为若干不同的 token；随后在以退化频谱特征为条件的并行分支中预测对应的干净 token；最后通过编解码器的解码器重建出干净语音。实验结果表明，在包含噪声、混响、带宽受限及其混合等多种失真条件下，ParaGSE 相比判别式与生成式基线方法都能稳定地产生更优的增强语音。此外，得益于 token 预测中的并行计算，ParaGSE 在 CPU 上的生成效率相较串行生成式语音增强方法提升约 1.5 倍。

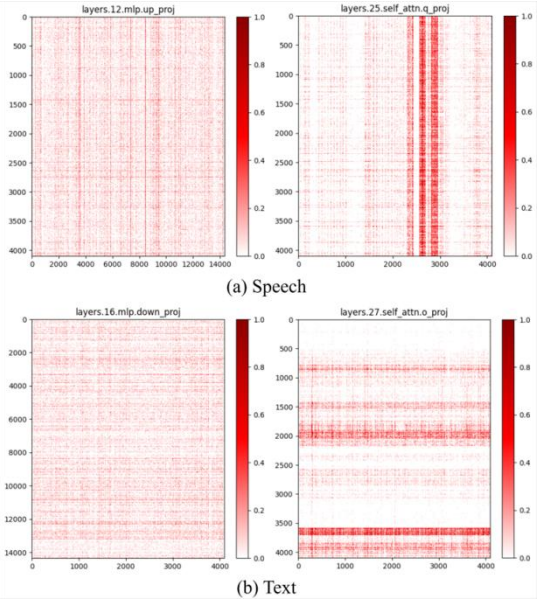
论文资源：Demo 语音网页 https://anonymity225.github.io/ParaGSE/

5. Understanding Textual Capability Degradation in Speech LLMs via Parameter Importance Analysis

论文作者：王超，郑瑞晨，艾杨，凌震华

论文单位：中国科学技术大学

论文简介：



将语音模态集成进大语言模型（LLM）显著扩展了其能力，但往往以削弱其核心文本能力为代价。这种退化现象限制了语音大模型充分挖掘其预训练文本知识的能力。本研究聚焦于语音大模型中广泛采用的编码器-适配器范式，深入剖析了该问题的底层机制。我们基于参数重要性评估提出分析框架，揭示了语音问答任务微调引发的文本参数重要性分布偏移现象：对文本推理至关重要的参数重要性层间分布遭到破坏。基于这一发现，我们探究了两种缓解策略：分层学习率（Layer-wise Learning Rate Scheduling）与低秩适配（LoRA），两者均旨在保持原始参数重要性分布。实验结果表明，这两种方法在保持文本能力方面均优于全量微调，同时还能提升下游口语问答任务的性能。此外，我们的分析将所提缓解策略的作用机制与大语言模型中的文本知识的结构特性相联系，为其有效性提供了理论依据。

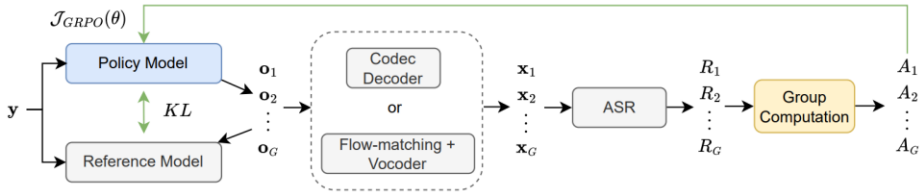
论文资源：预印版网址 https://arxiv.org/abs/2509.23755

6. GROUP RELATIVE POLICY OPTIMIZATION FOR TEXT-TO-SPEECH WITH LARGE LANGUAGE MODELS

论文作者：刘畅，胡亚军，高莹莹，张世磊，凌震华

论文单位：中国科学技术大学，科大讯飞，中国移动

论文简介：



本文提出了一种基于 GRPO 的方法，通过从现成的自动语音识别（ASR）模型中获取奖励信号，来提升基于大语言模型（LLM）的文本转语音（TTS）模型的性能。与以往针对 LLM-based TTS 的强化学习方法相比，我们的方法不需要专门的模型计算奖励或训练。此外，我们设计了一个复合奖励函数，将字符错误率（character error rate, CER）与从 ASR 模型中获得的负对数似然（negative log-likelihood, NLL）相结合，从而提供信息更丰富且更准确的奖励信号。我们将上述奖励信号用于 GRPO 微调预训练的 LLM-based TTS 模型，并评估其在零样本（zero-shot）TTS 任务中的表现。实验结果表明，该方法显著提升了 CosyVoice2 以及 Llasa 模型在中英日韩四个语种零样本合成语音的可懂度和自然度。消融实验和进一步的分析实验证明了结合这两种奖励成分的有效性。

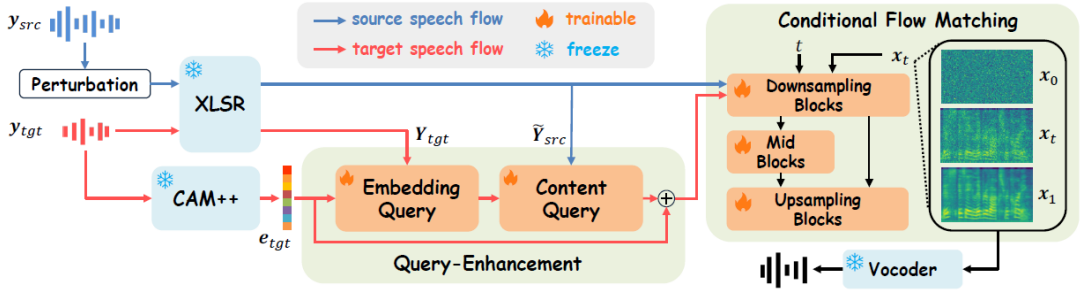
论文资源：预印版网址 https://arxiv.org/abs/2509.18798

7. QE-XVC: Zero-Shot Cross-Lingual Voice Conversion via Query-Enhancement and Conditional Flow Matching

论文作者：郭瀚杰，杜荟鹏，王诗明，江晓航，高莹莹，张世磊，凌震华

论文单位：中国科学技术大学

论文简介：



话者转换（Voice Conversion, VC）旨在修改语音的说话人身份，同时保留其语言内容。跨语言话者转换（Cross-lingual Voice Conversion, XVC）进一步实现了不同语言说话人之间的语音转换。现有的零样本 XVC 方法往往面临发音错误、说话人建模不足，或跨语言情境下源与目标对齐困难等问题。基于此，本文提出了 QE-XVC，一种结合了查询增强和条件流匹配（Conditional Flow Matching, CFM）的高效零样本 XVC 方法。具体而言，我们设计了一个查询增强模块，利用说话人验证（Speaker Verification, SV）模型的说话人嵌入向量以及内容表征依次作为查询，来增强由小波卷积提取的帧级说话人表征。随后，CFM 模型在内容和细粒度说话人表征的条件下生成转换后的声学特征，并通过声码器将其转化为语音波形。

实验结果表明，QE-XVC 在不引入任何量化或聚类操作的情况下，减少了潜在的发音错误，同时也提升了对源说话人韵律的保留效果。

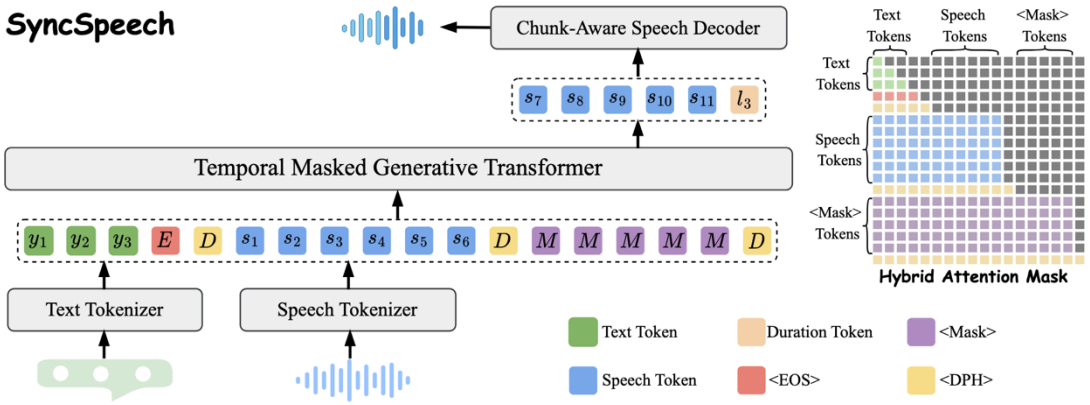
论文资源：语音样例 https://hjguo01.github.io/QE-XVC/

8. SyncSpeech: Efficient and Low-Latency Text-to-Speech based on Temporal Masked Transformer

论文作者：盛峥彦，杜志浩，张仕良，鄢志杰，陈丽萍

论文单位：中国科学技术大学

论文简介：



当前的语音合成模型面临着：自回归模型因语音 Token 帧率高而导致生成效率较低下，而非自回归模型因其时序上的无序特性而存在较高的首包延迟。为了弥合这一鸿沟，本文推出了基于时序掩码 Transformer 的低延迟高效语音合成模型 SyncSpeech，该模型实现了自回归模型的有序生成与非自回归模型的并行解码。在训练阶段，设计了随机截断的序列构建规则、相应的训练目标以及混合注意力掩码。此外，为了增强训练效率，我们引入了一种高概率掩码预训练策略，该策略同时也显著提升了模型的整体性能。在推理阶段，SyncSpeech 通过单步解码每个新到达文本标记（Token）所对应的所有语音标记，实现了极高的效率；同时，通过在接收到流式输入的第二个文本标记后立即开始生成语音，实现了极低的延迟。评估结果显示，SyncSpeech 在保持与主流自回归 TTS 模型相当的语音质量的同时，实现了首包延迟 5.8 倍的降低以及实时率 8.8 倍的提升。

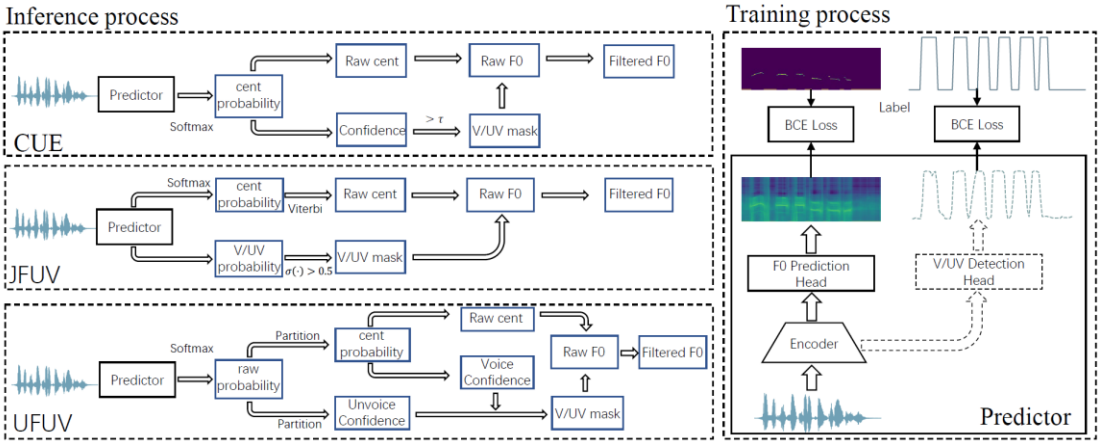
论文资源：Demo 语音网页 https://SyncSpeech.github.io/

9. JUND-F0: A NOVEL DEEP LEARNING FRAMEWORK FOR JOINT UNVOICED/VOICED DETECTION AND F0 ESTIMATION

论文作者：陈雨昂，冯锐，刘寅龙，胡郁，袁家宏

论文单位：中国科学技术大学

论文简介：



基频（F0）提取在语音信号处理中具有广阔的应用前景，但深度学习模型中常用的置信度驱动清音估计（CUE，confidence-driven unvoiced estimation）策略往往难以在适应性和显式清/浊音（V/UV）建模方面保持良好的平衡。本研究提出了一种名为 JUND-F0（联合清浊音检测与基频提取）的新型深度学习框架。首先，JUND-F0 集成了联合基频与清浊音学习（JFUV，Joint F0 and V/UV learning）和统一基频与清浊音检测（UFUV，Unified F0 and V/UV detection）两种创新策略。JFUV 通过将清浊音信息直接引入网络，实现了对语音活动和基频分布的同步学习。随后，UFUV 将清音帧定义为网络输出中的显式类别，在支持精确基频预测的同时，实现了端到端的清浊音判定。此外，该框架在平均绝对误差（MAE）、粗基频误差（GPE）、原始基频准确率（RPA）和清浊音决策误差（VDE）等指标上均表

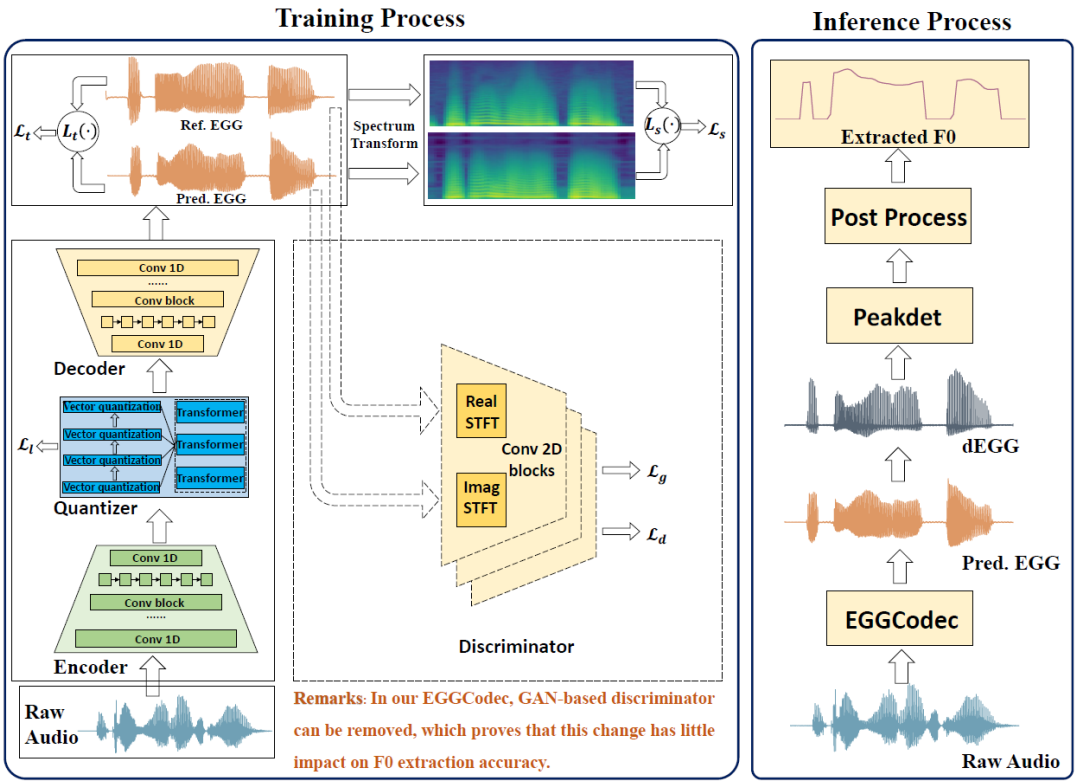
现优异。本研究表明，通过联合学习与统一检测策略，JUND-F0 有望成为稳健且具备高度自适应能力的基频分析工具。

10. EGGCODEC: A ROBUST NEURAL ENCODEC FRAMEWORK FOR EGG RECONSTRUCTION AND F0 EXTRACTION

论文作者：冯锐，陈雨昂，胡郁，袁家宏

论文单位：中国科学技术大学

论文简介：



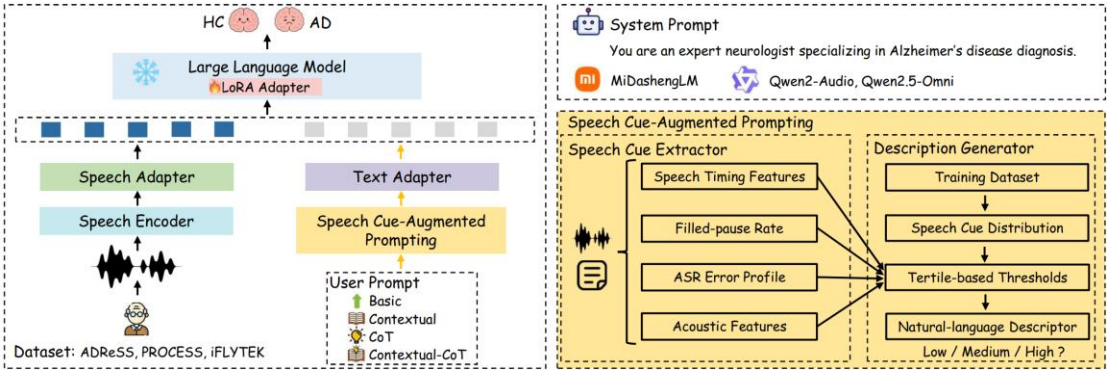
基频（F0）提取是语音信号处理的基础任务，但传统的直接提取方法受限于声带振动机制的复杂性和录音环境的多变性，往往难以保持良好的稳定性。本研究介绍了一种名为 EGGCodec 的稳健神经编解码器框架模型，专为电声门图（EGG）信号重建和基频（F0）提取而设计 。我们提出了一种多尺度频域损失函数，以捕捉原始和重建 EGG 信号之间的细微关系，并辅以时域相关性损失以提高泛化能力和准确性 。与直接从特征中提取 F0 的传统神经编解码器模型不同，EGGCodec 利用了与 F0 对应关系更紧密的重建 EGG 信号 。通过移除传统的 GAN 判别器，我们在不牺牲效率的前提下简化了 EGGCodec 的训练过程，且仅造成了可以忽略不计的性能下降 。在广泛使用的包含 EGG 的数据集上进行的训练和大量评估表明，EGGCodec 的表现良好，将平均绝对误差（MAE）从 14.14 Hz 降低至 13.69 Hz，并将清浊音决策误差（VDE）改善了 38.2% 。此外，大量的消融实验验证了 EGGCodec 各个组件的贡献 。

11. CROSS-LINGUAL ALZHEIMER’S DISEASE DETECTION WITH MULTIMODAL LLMS VIA SPEECH CUE-AUGMENTED PROMPTING AND INSTRUCTION TUNING

论文作者：刘寅龙，李远超，陈雨昂，何浏，冯锐，陈佳鑫，袁家宏

论文单位：中国科学技术大学，爱丁堡大学

论文简介：



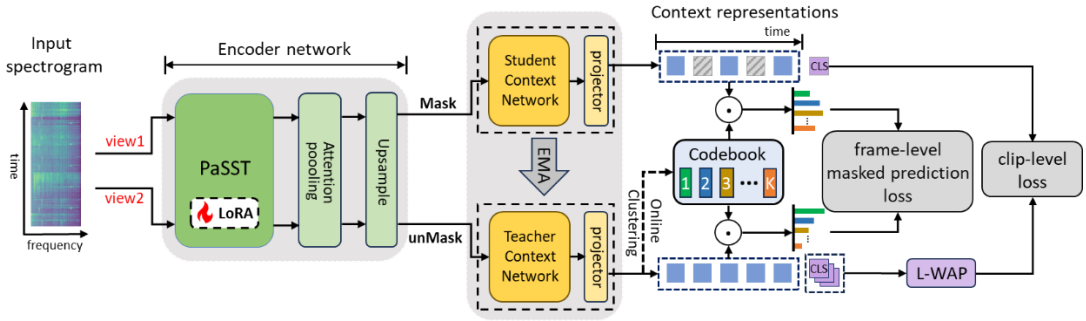
基于语音的阿尔茨海默病（AD）检测具有广阔的应用前景，但传统的监督学习模型往往难以在不同语言和数据集之间保持良好的泛化能力。本研究提出了一种利用多模态大语言模型（MLLMs）进行跨语言 AD 检测的新方法。首先，我们在零样本（Zero-shot）设置下，针对英语（ADReSS、PROCESS）和中文（iFLYTEK）三个 AD 数据集，系统评估了三种 MLLMs（MiDashengLM、Qwen2-Audio 和 Qwen2.5-Omni）在四种提示词（Prompt）下的表现。研究结果表明，上下文思维链提示（Contextual Chain-of-Thought Prompt）能够实现最强的零样本性能。随后，我们提出了语音线索增强提示（Speech Cue–Augmented Prompting, SCAP）方法，即在提示词中预置四类 与 AD 相关的语音线索的自然语言描述。实验结果显示，结合 SCAP 的 MLLMs 不仅提升了零样本表现，甚至超越了传统的监督学习模型。此外，我们探索了通过低秩自适应（LoRA）对 MLLMs 进行指令微调（Instruction Tuning）。结果发现，结合 SCAP 的指令微调进一步提升了域内（In-Domain，ID）准确率和跨域（Out-of-Domain，OOD）迁移能力，在 ID 和 OOD 两种设置下均优于所有监督学习基准模型。本研究表明，通过精心设计的提示策略与轻量化微调，MLLMs 有望成为稳健且具备语言无关性的 AD 检测工具。

12. A Task-Aware Dual-Level Self-Supervised Learning Method for Effective Sound Event Detection

论文作者：刘骏，顾庆，蔡鹏飞，江南，宋彦

论文单位：中国科学技术大学

论文简介：



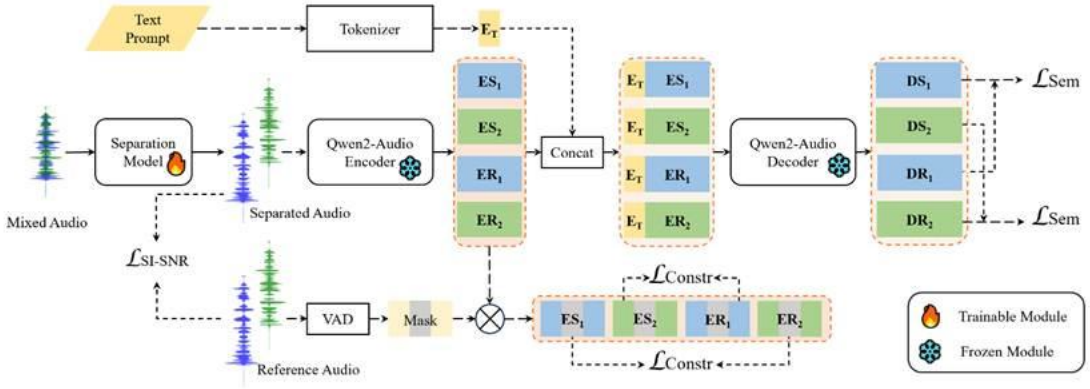
用于声音事件检测（SED）的半监督学习，以音频标注（AT）作为辅助任务，已被证明非常有效，而自监督学习（SSL）也显示出强大的潜力。因此，探索能够补充半监督学习的 SSL 方法十分具有潜力，但现有大多数 SSL 方法采用单级任务（例如片段级或帧级），与联合 SED–AT 范式不匹配。为了解决该问题，我们提出了一种任务感知的双级自监督学习方法，以匹配联合 SED-AT 半监督学习。具体来说，我们设计了一种基于 Transformer 的孪生网络，通过自蒸馏的方式联合利用针对 SED 设计的帧级任务和针对 AT 设计的片段级任务，有效地将知识迁移到下游任务，从而实现任务感知学习。此外，对于帧级目标，我们还在线学习了一个原型码本，为 SED 提供更有效的伪监督。在 DESED 数据集上进行的实验表明，我们的方法具有优越的性能，分别取得了 0.611/0.819 的 PSDS1/PSDS2。

13. Hierarchical Contrastive Learning with Speech Language Model for Separating Similar Speakers

论文作者：韩仪，陈航，王若愚，杜俊，牛树同，高建清，许苏魁

论文单位：中国科学技术大学，科大讯飞，安徽信息工程学院

论文简介：



当前语音分离技术在高度相似说话人场景下面临性能衰退。本文提出了 HCL-SLM（Hierarchical Contrastive Learning with Speech Language Model），一种用于解决高度相似说话人分离问题的创新框架。做法是引入 Qwen2-Audio 语音语言模型的分层语义表示，设计双对比损失机制：语义相似性损失使同一说话人的音频片段在语义空间中更接近；语义对比损失通过语音活动检测（VAD）排除静音帧并强制维持不同说话人片段的语义差异，从而有效减少分离片段错误交换问题。该框架不改变分离网络架构，仅通过添加语义监督损失来增强模型区分高度相似说话人的能力。针对多种最先进的盲源分离网络进行的广泛实验表明，作为即插即用的训练目标，HCL-SLM 能够稳定提高分离性能。

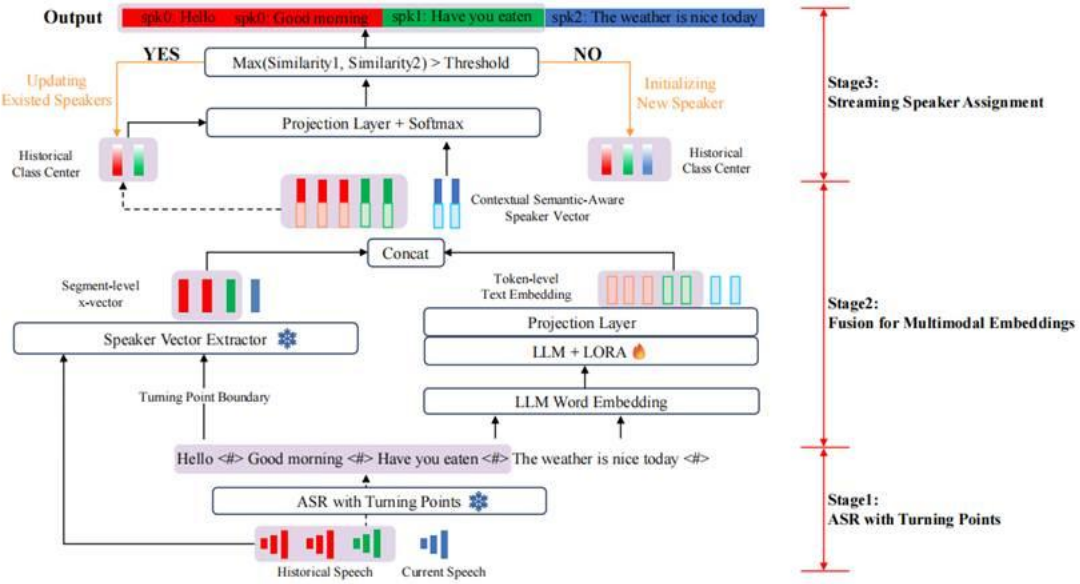
论文资源：开源代码：https://github.com/proudpie/HCL-SLM

14. INTEGRATING SPEAKER EMBEDDINGS AND LLM-DERIVED SEMANTIC REPRESENTATIONS FOR STREAMING SPEAKER DIARIZATION

论文作者：程天佑，奚昌凤，潘嘉，王若愚，陈航，韩江宇，Lukáš Burget，高建清，杜俊

论文单位：中国科学技术大学，科大讯飞，布尔诺理工大学

论文简介：



真实世界的会议对话通常呈现出同一说话人在其各次发言中的风格一致性，并且相邻语段之间往往具有很强的语义依赖关系。为利用这些特性，本文引入文本信息以辅助流式说话人日志。鉴于大语言模型（LLM）在长上下文理解方面的卓越能力，我们采用 LLM 进行长时语义建模。具体而言，我们首先使用自动语音识别模型，根据语音中的转折点将输入音频切分为句子级片段。对于每个句子，我们提取说话人嵌入，并利用 LLM 获取对应的 token 级文本表征。随后，

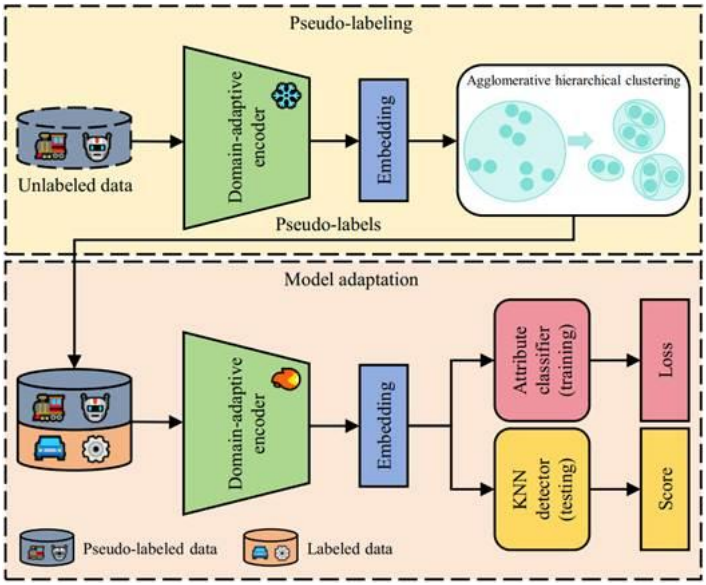
我们研究多种融合策略，将文本与音频特征进行融合，并将融合后的表征输入相似度度量模块，以判定每个句子的说话人身份。在自建的域内数据集以及公开的跨域 AISHELL-4 基准上的实验结果表明，LLM 生成的文本表征能够有效补充说话人嵌入，从而提升流式说话人日志性能，验证了所提方法在跨域场景下的有效性。

15. Improving Anomalous Sound Detection with Attribute-aware Representation from Domain-adaptive Pre-training

论文作者：方昕，钟桂锐，王青，褚繁，王磊，钱门贵，蔡明琦，吴江照，高建清，杜俊

论文单位：中国科学技术大学，科大讯飞，国家智能语音创新中心

论文简介：



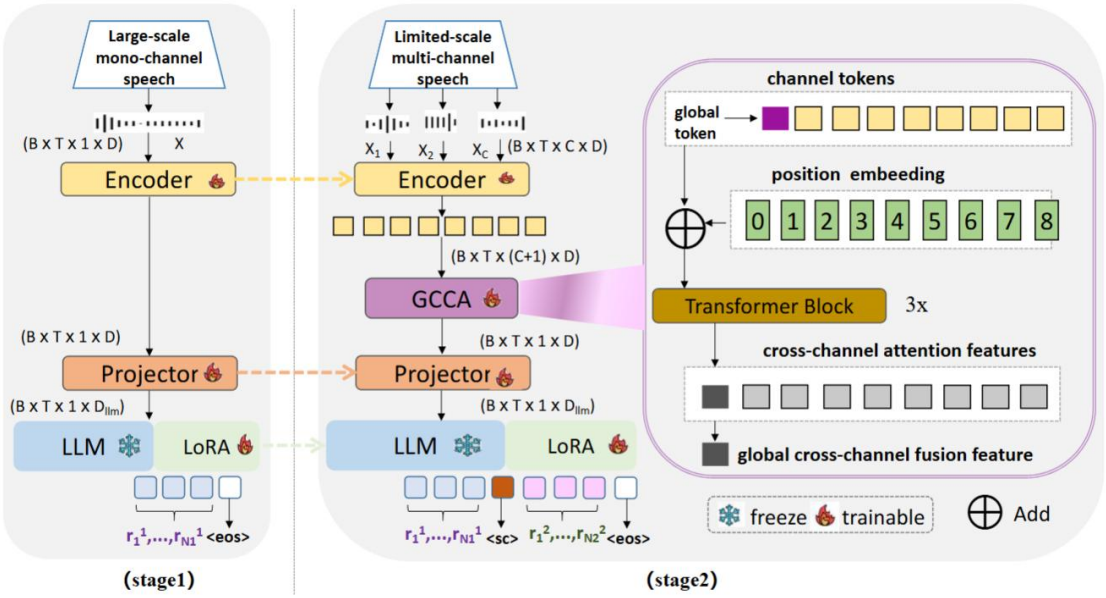
异常音频检测（ASD）通常被表述为一项机器属性分类任务，这是由于在常见场景中，只有正常数据可用于训练。然而，详尽地收集机器属性标签既费力又不切实际。为了应对属性标签缺失的挑战，本文提出了一种凝聚式层次聚类方法，该方法利用来自领域自适应预训练模型的表征来分配伪属性标签，这些表征有望捕捉到机器属性的特征。然后，我们通过有监督微调对预训练模型进行模型适应，以进行机器属性分类，从而取得了新的最优性能。在 2025 年声学场景与事件检测与分类（DCASE）挑战赛数据集上的评估表明，我们提出的方法取得了显著的性能提升，最终超越了我们之前在该挑战赛中排名第一的系统。

16. Advancing LLM-Based Multi-Channel Multi-Speaker speech recognition with Global Cross-Channel Attention and Sentence-Ordered First-In First-Out Serialized Output Training

论文作者：万根顺，刘丽娟，奚昌凤，陈航，于心迪，潘嘉，杜俊，叶中付

论文单位：中国科学技术大学，科大讯飞

论文简介：



近年来，自动语音识别（ASR）技术取得了显著进展，但在实际的多说话人对话场景中仍然面临诸多挑战。虽然多麦克风输入在多说话人场景下的性能优于单麦克风设置，但由于数据稀缺、复杂的声学环境以及跨通道依赖建模的局限性，多通道多说话人 ASR 仍然极具挑战性。

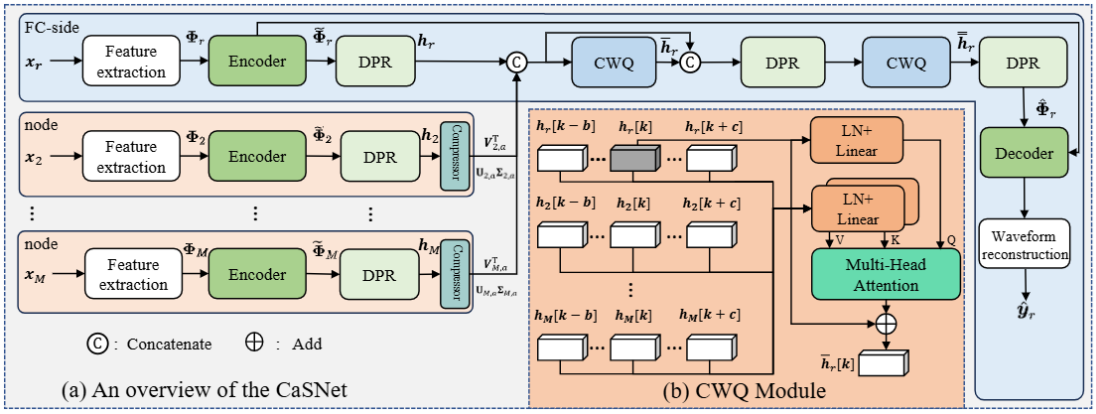
为了解决这些挑战，我们首次提出了一种将大型语言模型（LLM）集成到多通道多说话人 ASR 中的新型框架。首先，我们提出了一种两阶段训练策略——在大规模单通道数据上进行预训练，然后在多通道语料库上进行微调——以缓解数据稀缺问题并增强鲁棒性。其次，我们引入了一种句子顺序先进先出（FIFO）序列化输出训练（SOT）方法，以保持自然的语音顺序并提高语义连贯性。最后，我们设计了一种全局跨通道注意力机制，支持可变输入通道，无需填充。在 MISP-Meeting 数据集上的实验结果表明，我们的框架将基线 ASR 系统的字符错误率（CER）降低了 78.5%（单说话人）和 55.4%（重叠测试集），并且在不同的通道设置下表现出强大的泛化能力。

17. CaSNet: Compress-and-Send Network Based Multi-Device Speech Enhancement Model for Distributed Microphone Arrays

论文作者：蒋承乾，张结，严浩尹

论文单位：中国科学技术大学

论文简介：



分布式麦克风阵列是一种很有前景的新一代语音交互平台，但在嘈杂环境下仍需要语音增强来提升语音质量。现有的语音增强方法通常需要将所有设备采集的原始语音波形汇聚到融合中心，再设计多麦克风模型，这会带来较高的通信带宽和能耗开销。针对资源受限的分布式麦克风阵列，本文提出了一种压缩并传输网络。在该框架中，一个麦克风被指定为

融合中心和参考通道，其余设备则将采集到的原始语音数据编码为特征矩阵，并通过奇异值分解进行压缩，从而获得更加紧凑的表示。融合中心接收到各设备传输的特征后，首先基于参考通道通过跨窗口查询进行对齐，随后利用神经网络解码，生成具有空间一致性的增强语音。在多个数据集上的实验结果表明，与未进行压缩的情况相比，所提出的 CaSNet 能够在几乎不影响性能的前提下显著减少数据传输量。

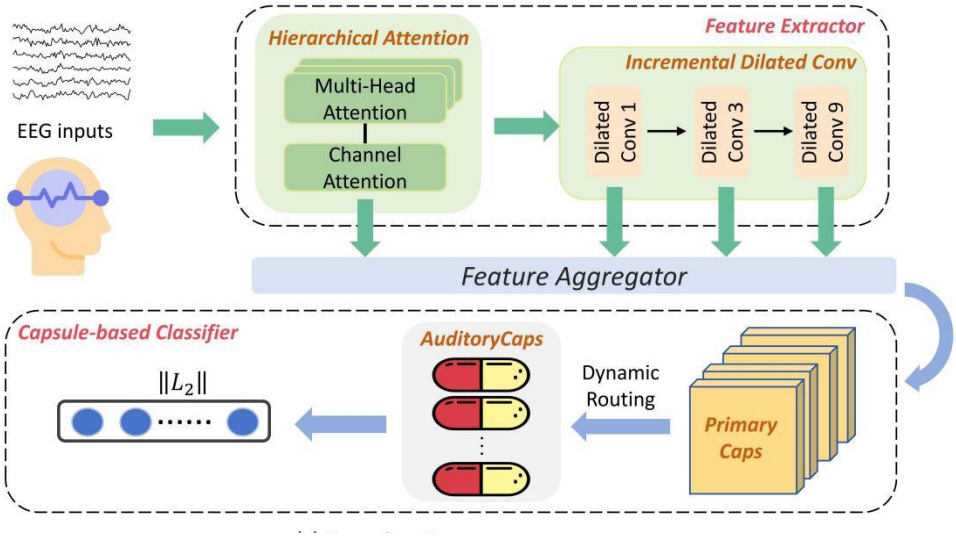
论文资源：开源代码：https://github.com/Jokejiangv/CaSNet

18. NeuroCapsNet: A Neuro-Inspired Capsule Network for Multi-Direction Auditory Spatial Attention Detection

论文作者：张育榜，张结，玉山·牙生江

论文单位：中国科学技术大学

论文简介：



基于脑电图（EEG）信号的听觉注意力解码（Auditory Attention Decoding, AAD）旨在识别听者当前关注的目标声源，以提升助听设备在多说话人场景中的言语理解能力。现有 AAD 方法多聚焦于二分类任务（如判断目标说话人位于听者左侧或右侧），当多个说话人处于相近空间方位时，其解码可靠性显著下降，推动了对更高空间分辨率听觉注意力解码方法的研究。为此，本文提出一种受神经机制启发的胶囊网络模型（NeuroCapsNet），用于多方向听觉空间注意力解码。该模型融合层次注意力机制与增量膨胀卷积单元，有效提取 EEG 信号中的听觉相关特征；并引入基于动态路由的胶囊分类器，模拟大脑在高低层级神经元之间对任务相关连接的选择性强化，从而更好地建模听觉注意相关的局部神经活动与全局空间方向表征之间的层级关系。在包含十个说话人方向的公开数据集上的实验结果表明，NeuroCapsNet 在试次内和留一交叉验证设置下均优于现有方法。

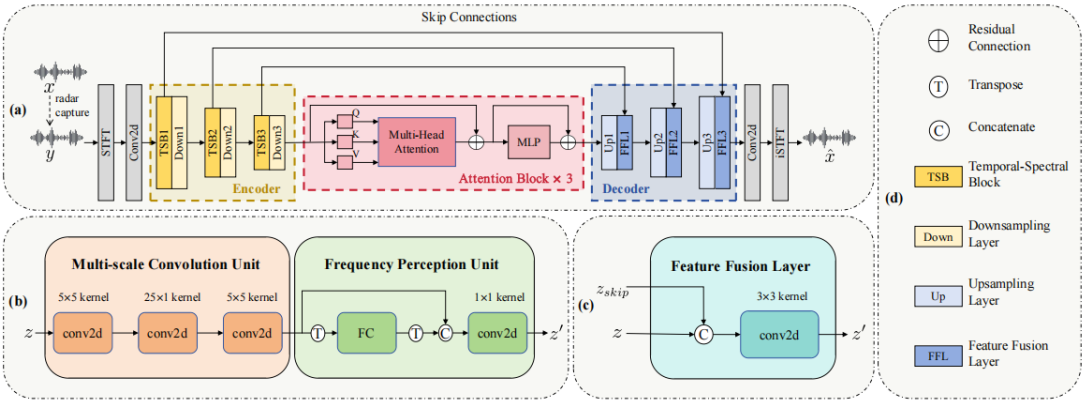
论文资源：开源代码：https://github.com/zyb6239/NeuroCapsNet

19. The USTC-NERCSLIP System for the RASE Challenge

论文作者：隋承展，张爽，张结

论文单位：中国科学技术大学，大连理工大学

论文简介：



雷达声学语音增强（RASE）挑战旨在从噪声和带宽受限的 FMCW 毫米波雷达信号中重建清晰、全带宽的语音，适用于声学噪声干扰或远距离场景。然而，传统方法往往难以有效处理噪声和受限带宽带来的挑战。针对这一问题，我们提出了时频融合网络（TSFNet），该网络集成多尺度特征提取和注意力机制，以增强语音的可理解性和质量。具体而言，TSFNet 采用编码器-解码器架构，其中编码器通过时序-频谱块进行多尺度分析，注意力瓶颈层建模全局依赖关系，解码器则利用上采样和特征融合层细化输出。在官方测试集上的实验结果表明，TSFNet 在 RASE 挑战中加权得分为 0.26，在所有队伍中排名第三，消融研究进一步证实了时序-频谱块、注意力机制等关键组件对性能的提升作用。