Technical Report 1

T237 **FlySpeech**

He Wang, Lei Xie

NPU-ASLP团队CNVSRC 2024视觉语音识别系统描述

成员：**王贺**、谢磊

单位：西北工业大学音频语音与语言处理研究组（ASLP@NPU）

目录

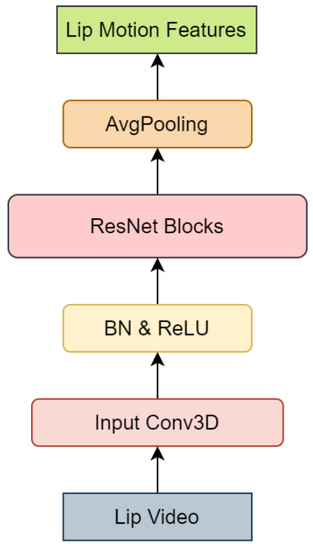
概述

- **视觉语音识别（VSR）**
  - 将说话人的连续图像信息（视频）自动识别为相应文本的技术
  - 语音识别（ASR）系统性能在复杂的声环境中会有明显降低
  - 可以应用于无声语言接口或帮助有语言障碍的人等
- **第二届中文连续视觉语音识别挑战赛（CNVSRC 2024）**
  - 旨在探索录音棚阅读和互联网上演讲场景下的大词汇量中文连续视觉语音识别（LVCVSR）
  - 数据集：CNCVS（300h+）、CNVSRC-Single （80h+）、CNVSRC-Multi （25h+）
  - 赛道设置：单说话人与多说话人两个任务、分别又设固定（fixed）和开放（open）两个赛道
- **方案简介与竞赛成绩**
  - Enhanced ResNet3D视频前端 + E-Branchformer编码器 + Bi-Transformer解码器 + 多系统融合策略
  - 单说话人测试集CER 30.4679% （fixed赛道第一、open赛道第二）
  - 多说话人测试集CER 34.2955% （fixed赛道第一、open赛道第一）

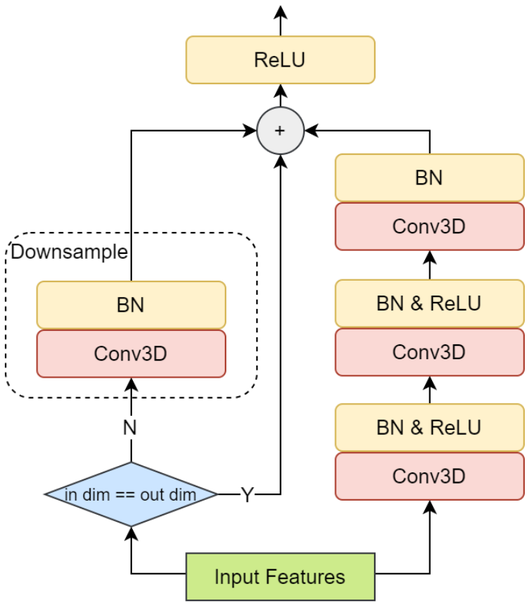
目录

Enhanced ResNet3D视觉前端

- 整体模型设计参考于经典图像分类网络ResNet
- 二维卷积模块 -> 三维卷积（Conv3D）模块
- Enhanced ResNet3D◎ ◎ 前端（图a）由Input Conv3D、ResNet3D堆叠块以及最后的平均池化模块◎ 成
  - Input Conv3D将输入的视频特征映射到更高的维度
  - 每个ResNet3D◎ 由几个基本◎ （图b）◎ 成，◎ 些基本◎ 由堆叠的Conv3D和BatchNorm◎ 成，构建◎ ◎ 特征建模的基本◎ 元
  - 最◎ 的平均池化模块负责对建模后的视频特征的高、◎ 度◎ 度进行平均，将原先的三维特征◎ ◎ ◎ 一维

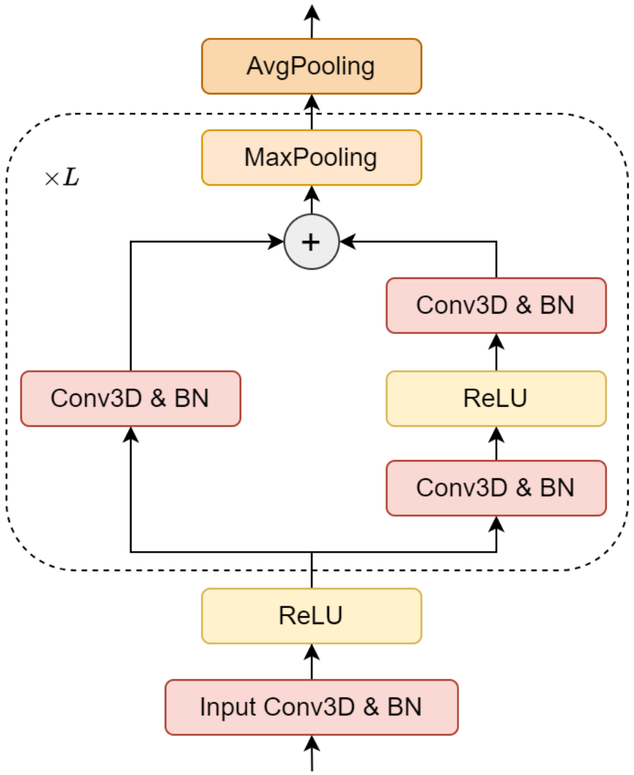


(a) Enhanced ResNet3D Visual Front-end

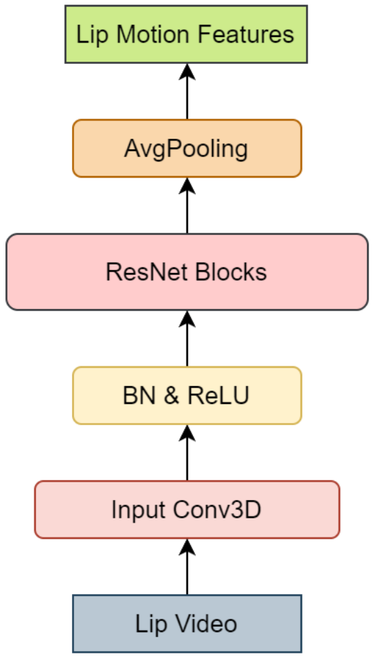


(b) Enhanced ResNet3D Basic Block

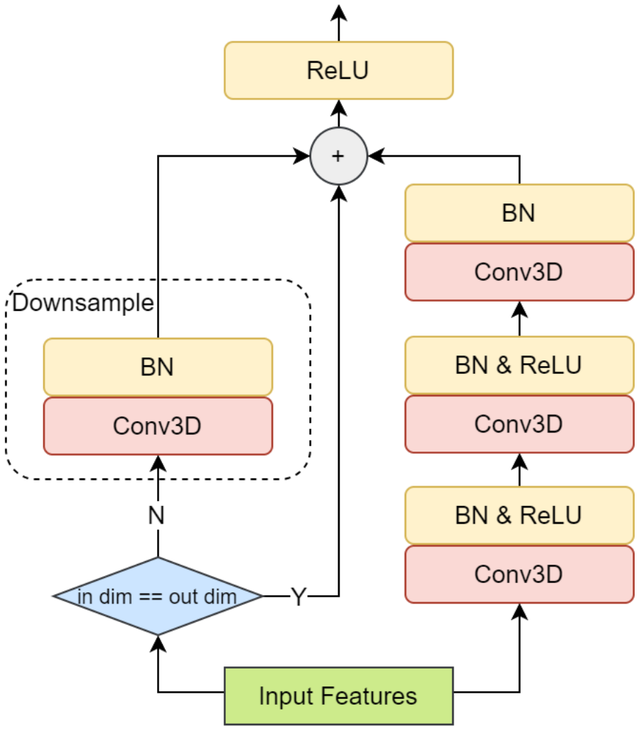
ResNet3D VS Enhanced ResNet3D



ResNet3D Visual Front-end



(a) Enhanced ResNet3D Visual Front-end



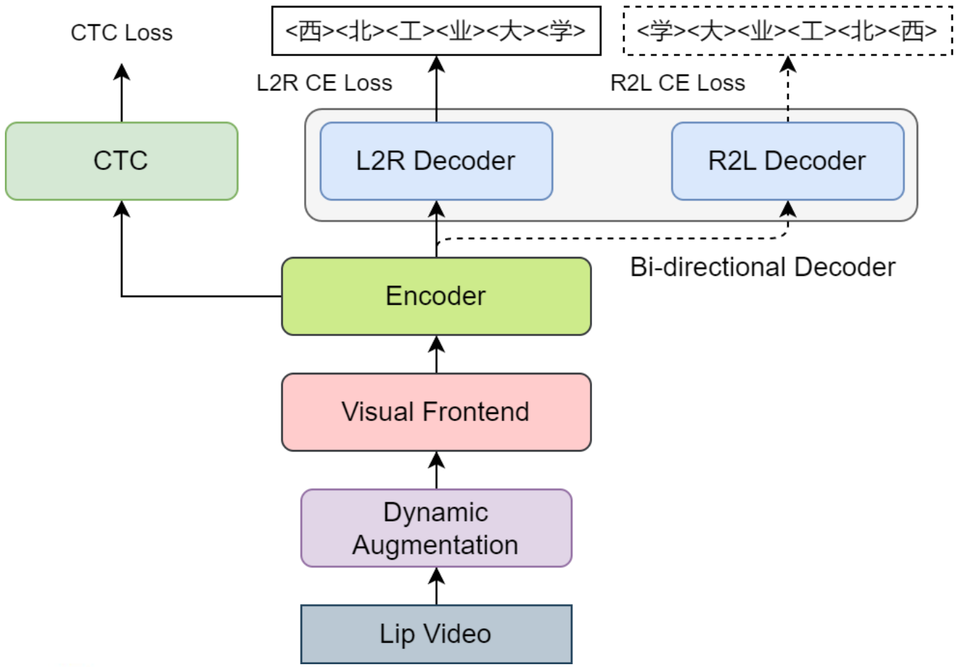
(b) Enhanced ResNet3D Basic Block

视觉语音识别系统

- 整体结构follow经典的端到端Joint CTC/Attention框架
- 主要由四部分组成：视觉前端、编码器、CTC层、解码器
  - Enhanced ResNet3D视觉前端用于提取视频特征
  - 编码器进一步建模视频特征；共采用了Confomer、Branchformer以及E-Branchformer三种编码器
  - CTC层接受编码器输出计算CTC Loss辅助编码器训练
  - 解码器由两部分组成，分别为接受正序文本输入的L2R Decoder以及倒序文本输入的R2L解码器，两者均为Transformer Decoder
- 模型损失函数

$$loss = \lambda loss_{ctc} + (1-\lambda)(\alpha loss_{ce}^{r2l} + (1-\alpha) loss_{ce}^{l2r})$$

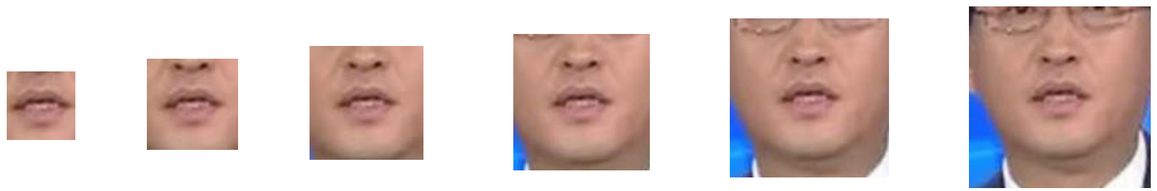
- 其中λ、α均为可调节参数，均设置为0.3



目录

数据集与数据处理

- **数据集**
  - 单说话人模型：CN-CVS + CNVSRC-Single
  - 多说话人模型：CN-CVS + CNVSRC-Multi
- **数据处理**
  - 静态数据增广：0.9、1.0、1.1三倍变速
  - 动态数据增广：训练过程中随机进行旋转、水平翻转、灰度化以及色彩抖动
  - 多尺寸数据构建



~~size 48~~ ~~size 64~~ size 80 size 96 size 112 size 128

CN-CVS
sample n001_00001_001

模型配置与系统构建流程

- **实验配置**
  - 所有系统构建均基于WeNet开源工具
  - Enhanced ResNet3D视频前端共有4个ResNet3D Block，分别含有3, 4, 6, 3个ResNet Basic Blocks，特征长度依次为32, 64, 128, 256（参考ReNet50的相关配置）
  - E-Branchformer编码器共有12层，特征维度为256，注意力头为4，1024维Feed Forward
  - Bi-Transformer L2R Decoder和R2L Decoder各由3层Transformer Layer堆叠而成
- **系统构建流程**
  - 训练阶段：使用训练集进行50轮训练，平均15轮验证集损失模型作为验证集解码模型
  - 微调阶段：使用训练阶段的平均模型进行初始化，使用全量CNVSRC-Single/Multi数据进行10轮微调，并平均全部10轮模型作为最终测试集的解码模型
  - 解码阶段：使用WeNet Attention-Rescore解码方式，beam size 32，ctc weight 0.3

实验结果与分析

- 不同编码器之间的性能比较
  - 从实验结果可以看出在相同视觉前端模块以及解码器的情况下，仅改变编码器，三种编码器之间的性能呈现出：

E-Branchformer > Branchformer > Conformer

与CNVSRC 2023中实验的结论相同

Table 1: The CER(%) results of our VSR systems on Dev and Eval sets in T1 and T2 tasks. Crop refers to the size of training lip video data. Suffix FT represents the decoding result after fine-tuning.

System	Visual Frontend	Encoder	Decoder	Crop	T1.Dev	T1.Eval	T1.Eval.FT	T2.Dev	T2.Eval	T2.Eval.FT
M1	ResNet3D	E-Branchformer	Transformer	80	41.76	-	-	48.50	-	-
M2	ResNet3D	Conformer	Transformer	96	39.27	-	-	46.21	-	-
M3	ResNet3D	Branchformer	Transformer	96	38.86	-	-	46.49	-	-
M4	ResNet3D	E-Branchformer	Transformer	96	38.39	-	-	45.97	-	-
M5	Enhanced ResNet3D	E-Branchformer	Transformer	96	38.30	-	-	45.79	45.43	41.75
M6	ResNet3D	E-Branchformer	Transformer	112	38.13	-	35.91	44.88	44.56	40.37
M7	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	112	36.60	36.79	34.56	43.29	43.13	38.73
M8	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	128	36.48	37.04	**34.27**	43.23	42.99	**37.93**
R1	ROVER **Eval** of M1∼M8					32.7199			38.3602	
R2	ROVER **R1** and **Eval.FT** of M1∼M8					**30.4679**			**34.2955**	

实验结果与分析

- 不同裁剪尺寸之间的比较
  - 对比M1、M4以及M6实验结果可以得出在相同模型结构下，数据尺寸由80 -> 112，越大越好
  - 对比M7、M8实验结果可以得出数据尺寸进一步由112增大至128依然有收益
  - 综上，唇动数据裁剪尺寸在80到128区间内越大，基于相同模型结构构建的VSR系统性能越好

Table 1: The CER(%) results of our VSR systems on Dev and Eval sets in T1 and T2 tasks. Crop refers to the size of training lip video data. Suffix FT represents the decoding result after fine-tuning.

System	Visual Frontend	Encoder	Decoder	Crop	T1.Dev	T1.Eval	T1.Eval.FT	T2.Dev	T2.Eval	T2.Eval.FT
M1	ResNet3D	E-Branchformer	Transformer	80	41.76	-	-	48.50	-	-
M2	ResNet3D	Conformer	Transformer	96	39.27	-	-	46.21	-	-
M3	ResNet3D	Branchformer	Transformer	96	38.86	-	-	46.49	-	-
M4	ResNet3D	E-Branchformer	Transformer	96	38.39	-	-	45.97	-	-
M5	Enhanced ResNet3D	E-Branchformer	Transformer	96	38.30	-	-	45.79	45.43	41.75
M6	ResNet3D	E-Branchformer	Transformer	112	38.13	-	35.91	44.88	44.56	40.37
M7	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	112	36.60	36.79	34.56	43.29	43.13	38.73
M8	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	128	36.48	37.04	**34.27**	43.23	42.99	**37.93**
R1	ROVER **Eval** of M1∼M8					32.7199		38.3602		
R2	ROVER **R1** and **Eval.FT** of M1∼M8					**30.4679**		**34.2955**		

实验结果与分析

- 使用CNVSRC数据集微调带来的收益
  - 以效果最好的单系统M8为例
    - 在单说话人测试集上经过CNVSRC-Single微调取得了**2.77%**的CER绝对降低
    - 在多说话人测试集上经过CNVSRC-Multi微调取得了**5.06%**的CER绝对降低
  - 单说话人微调较为困难，一定程度上会破坏初始模型的泛化能力

Table 1: The CER(%) results of our VSR systems on Dev and Eval sets in T1 and T2 tasks. Crop refers to the size of training lip video data. Suffix FT represents the decoding result after fine-tuning.

System	Visual Frontend	Encoder	Decoder	Crop	T1.Dev	T1.Eval	T1.Eval.FT	T2.Dev	T2.Eval	T2.Eval.FT
M1	ResNet3D	E-Branchformer	Transformer	80	41.76	-	-	48.50	-	-
M2	ResNet3D	Conformer	Transformer	96	39.27	-	-	46.21	-	-
M3	ResNet3D	Branchformer	Transformer	96	38.86	-	-	46.49	-	-
M4	ResNet3D	E-Branchformer	Transformer	96	38.39	-	-	45.97	-	-
M5	Enhanced ResNet3D	E-Branchformer	Transformer	96	38.30	-	-	45.79	45.43	41.75
M6	ResNet3D	E-Branchformer	Transformer	112	38.13	-	35.91	44.88	44.56	40.37
M7	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	112	36.60	36.79	34.56	43.29	43.13	38.73
M8	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	128	36.48	37.04	**34.27**	43.23	42.99	**37.93**
R1	ROVER **Eval** of M1∼M8					32.7199			38.3602	
R2	ROVER **R1** and **Eval.FT** of M1∼M8					**30.4679**			**34.2955**	

实验结果与分析

- **两阶段ROVER系统融合带来的收益**
  - 阶段一：融合所有没有经过微调的系统解码结果
    - 在单说话人测试集上取得了**4.07%**的CER绝对下降（M7 -> R1）
    - 在多说话人测试集上取得了**4.63%**的CER绝对下降 （M8 -> R1）

Table 1: The CER(%) results of our VSR systems on Dev and Eval sets in T1 and T2 tasks. Crop refers to the size of training lip video data. Suffix FT represents the decoding result after fine-tuning.

System	Visual Frontend	Encoder	Decoder	Crop	T1.Dev	T1.Eval	T1.Eval.FT	T2.Dev	T2.Eval	T2.Eval.FT
M1	ResNet3D	E-Branchformer	Transformer	80	41.76	-	-	48.50	-	-
M2	ResNet3D	Conformer	Transformer	96	39.27	-	-	46.21	-	-
M3	ResNet3D	Branchformer	Transformer	96	38.86	-	-	46.49	-	-
M4	ResNet3D	E-Branchformer	Transformer	96	38.39	-	-	45.97	-	-
M5	Enhanced ResNet3D	E-Branchformer	Transformer	96	38.30	-	-	45.79	45.43	41.75
M6	ResNet3D	E-Branchformer	Transformer	112	38.13	-	35.91	44.88	44.56	40.37
M7	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	112	36.60	36.79	34.56	43.29	43.13	38.73
M8	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	128	36.48	37.04	**34.27**	43.23	42.99	**37.93**
R1	ROVER **Eval** of M1∼M8					32.7199			38.3602	
R2	ROVER **R1** and **Eval.FT** of M1∼M8					**30.4679**			**34.2955**	

实验结果与分析

- 两阶段ROVER系统融合带来的收益
  - 阶段二：使用阶段一的结果与所有经过微调的系统的解码结果进行二次融合
    - 在单说话人测试集上进一步取得了**2.25%**的CER绝对下降（R1 -> R2）
    - 在多说话人测试集上进一步取得了**3.63%**的CER绝对下降 （M8 -> R2）

Table 1: The CER(%) results of our VSR systems on Dev and Eval sets in T1 and T2 tasks. Crop refers to the size of training lip video data. Suffix FT represents the decoding result after fine-tuning.

System	Visual Frontend	Encoder	Decoder	Crop	T1.Dev	T1.Eval	T1.Eval.FT	T2.Dev	T2.Eval	T2.Eval.FT
M1	ResNet3D	E-Branchformer	Transformer	80	41.76	-	-	48.50	-	-
M2	ResNet3D	Conformer	Transformer	96	39.27	-	-	46.21	-	-
M3	ResNet3D	Branchformer	Transformer	96	38.86	-	-	46.49	-	-
M4	ResNet3D	E-Branchformer	Transformer	96	38.39	-	-	45.97	-	-
M5	Enhanced ResNet3D	E-Branchformer	Transformer	96	38.30	-	-	45.79	45.43	41.75
M6	ResNet3D	E-Branchformer	Transformer	112	38.13	-	35.91	44.88	44.56	40.37
M7	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	112	36.60	36.79	34.56	43.29	43.13	38.73
M8	Enhanced ResNet3D	E-Branchformer	Bi-Transformer	128	36.48	37.04	**34.27**	43.23	42.99	**37.93**
R1	ROVER **Eval** of M1∼M8					32.7199			38.3602	
R2	ROVER **R1** and **Eval.FT** of M1∼M8					**30.4679**			**34.2955**	

目录

总结

- 本次报告◎ ◎ 了我◎ 在CNVSRC 2024中单说话人与多说话人唇语识别的工作
- 在模型结构上，相较于CNVSRC 2023中的方案，我们基于ResNet3D视觉前端进行升级，设计了Enhanced ResNet3D；解码器部分引入了双向Transformer解码器
- 在数据处理方面，我们沿用了CNVSRC 2023中的数据增广策略，包括静态和动态增广两部分，同时将最大的唇动视频裁剪尺寸由112进一步扩大至128
- 在训练流程上，我们采用了训练 + 微调的两阶段训练流程，取得了明显的收益
- 最后，在利用ROVER技术进行两阶段系统结果融合后，我们取得了单说话人测试集上CER **30.4679%**，多说话人测试集上CER **34.2955%**，取得了单说话人任务固定赛道**第一名**，开放赛道**第二名**，多说话人任务固定赛道**第一名**、开放赛道**第一名**的好成绩

Thank You!

He Wang

Audio, Speech & Language Processing Group (ASLP@NPU)

www.npu-aslp.org