2021 第一届 Speech home 语音技术研讨会

中国科学院自动化研究所 Institute of automatic, Chinese academy of sciences

低延迟非自回归语音识别方法

田正坤　中科院自动化所在读博士

zhengkun.tian@nlpr.ia.ac.cn

导师：陶建华 研究员

目录

- 背景介绍
- 非自回归语音识别方法介绍
- 我们团队的工作
  - 尖峰触发非自回归语音识别方法
  - 混合自回归与非自回归语音识别方法
- 总结与展望
- ICASSP2022 ADD 比赛简介

背景介绍

1. 会议字幕（线上or线下）/手机输入法
2. 语音交互（手机/音箱/智能助手）
3. 其他各种后台语音数据转录任务

背景介绍

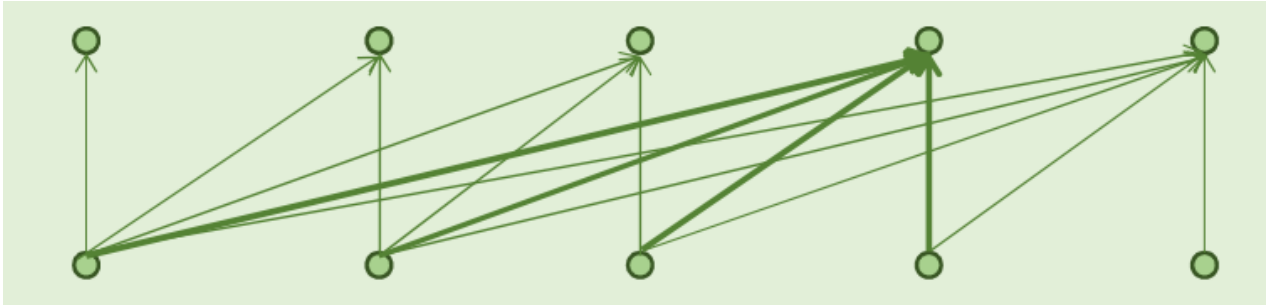
- 流式任务（边听边识别）
  - CTC
  - RNN-Transducer/SA-Transducer/Transformer-Transducer
  - MoChA/CIF/SMLTA/SCAMA/Triggered Attention
  - FastEmit/Fast-Skip Regularization

- 非流式任务（逐句识别）
  - 基于深度卷积构建的CTC模型
  - Speech Transformer
  - LAS各种Attention-based Encoder-Decoder模型

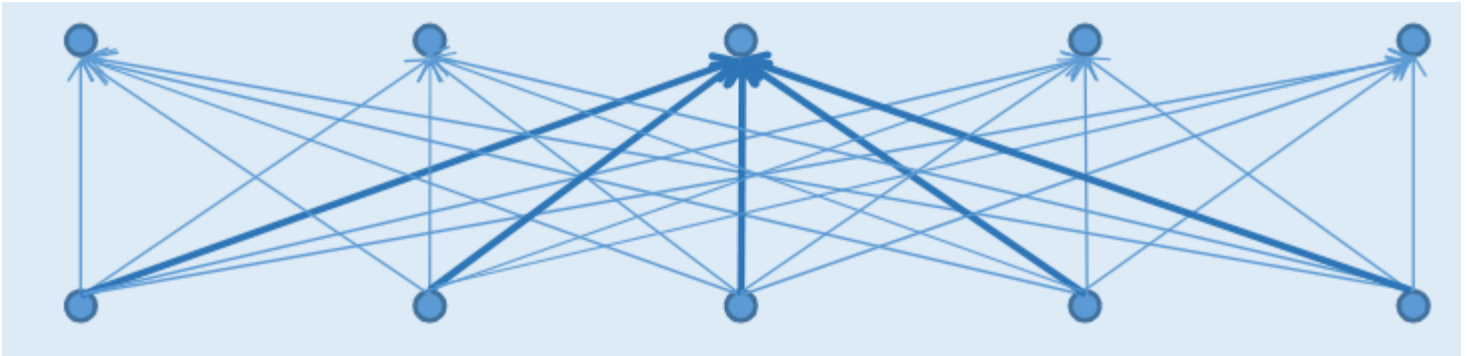
降低语音识别系统的延迟是不懈的追求！

非自回归语音识别方法

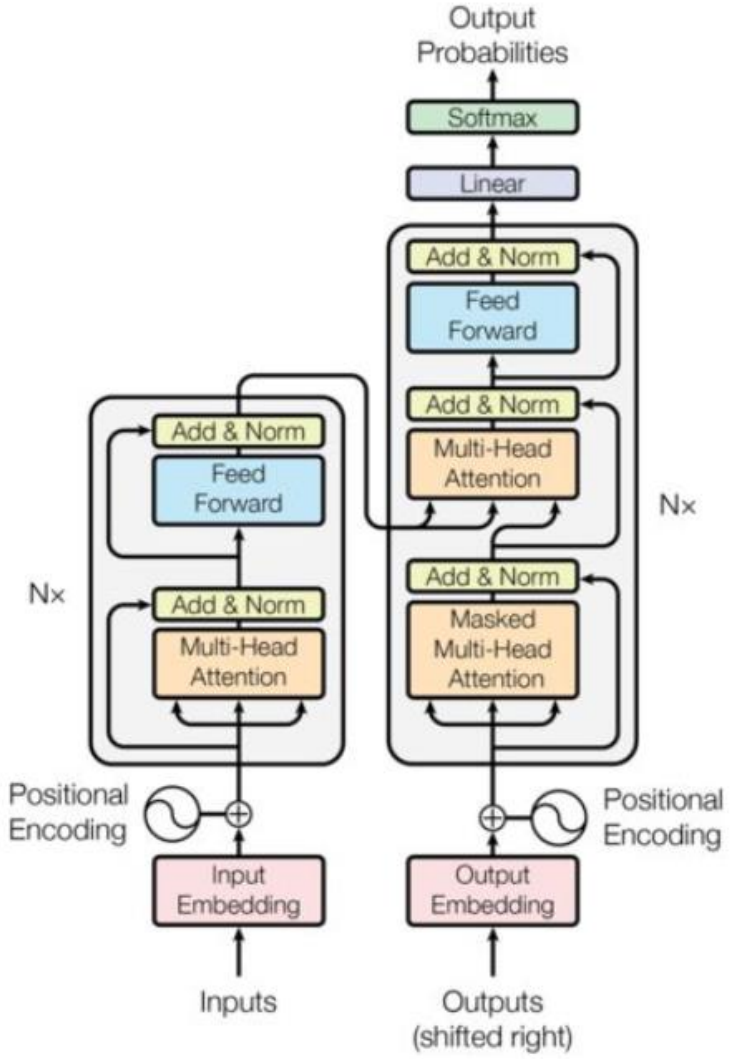
- 自回归模型 （时序依赖，不能并行）

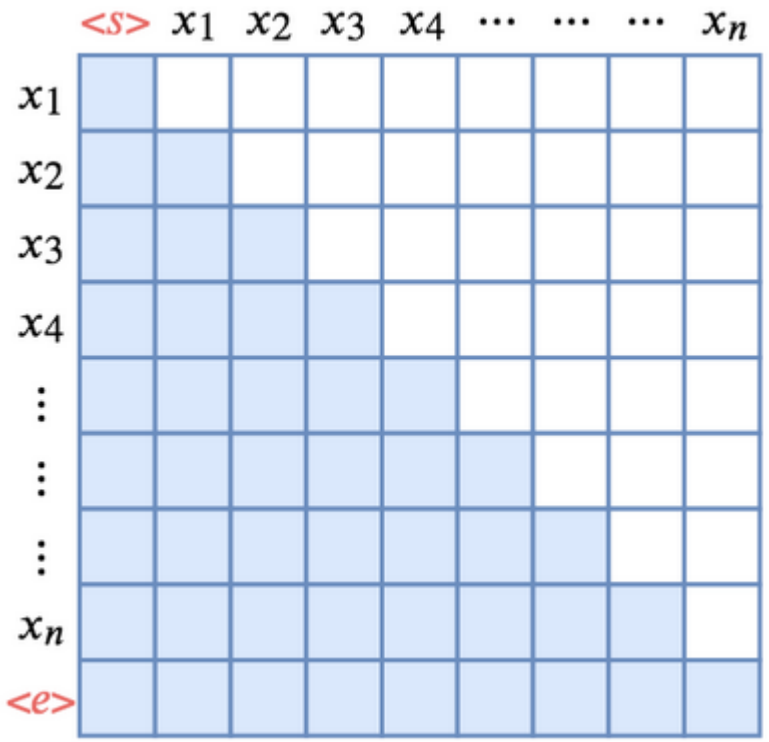


- 非自回归模型 （时序独立，完全并行）



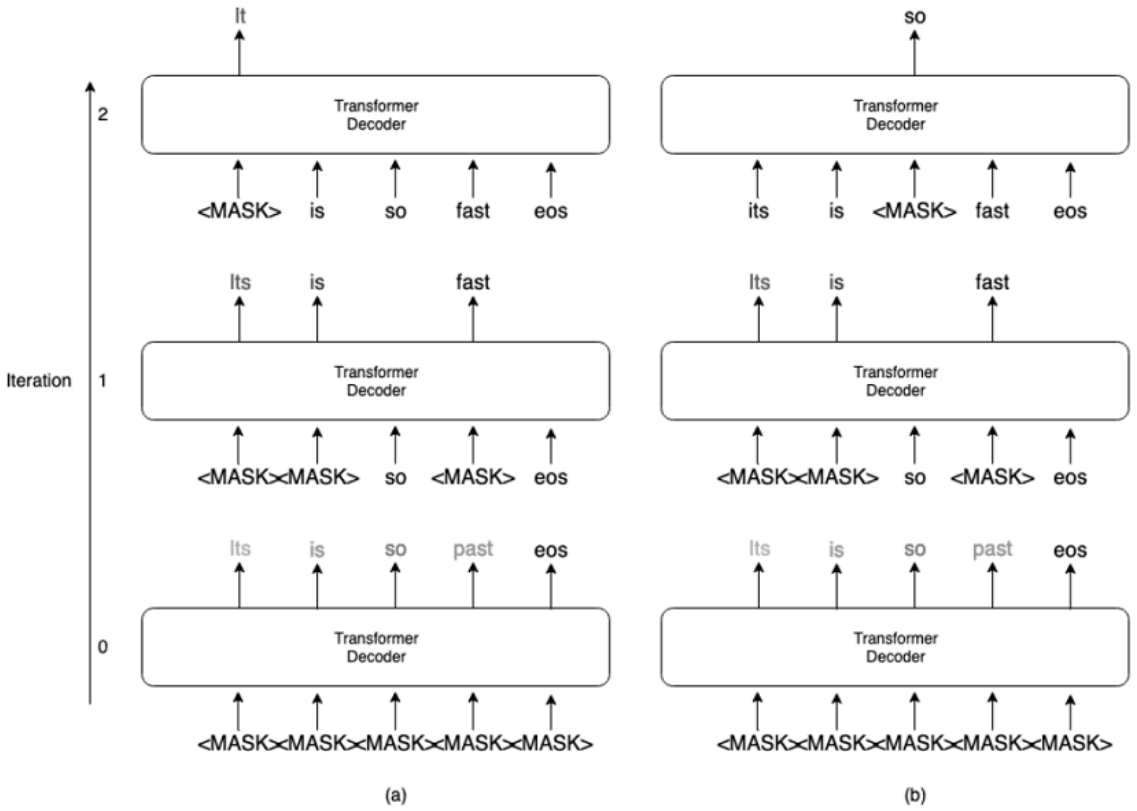
非自回归模型的构建

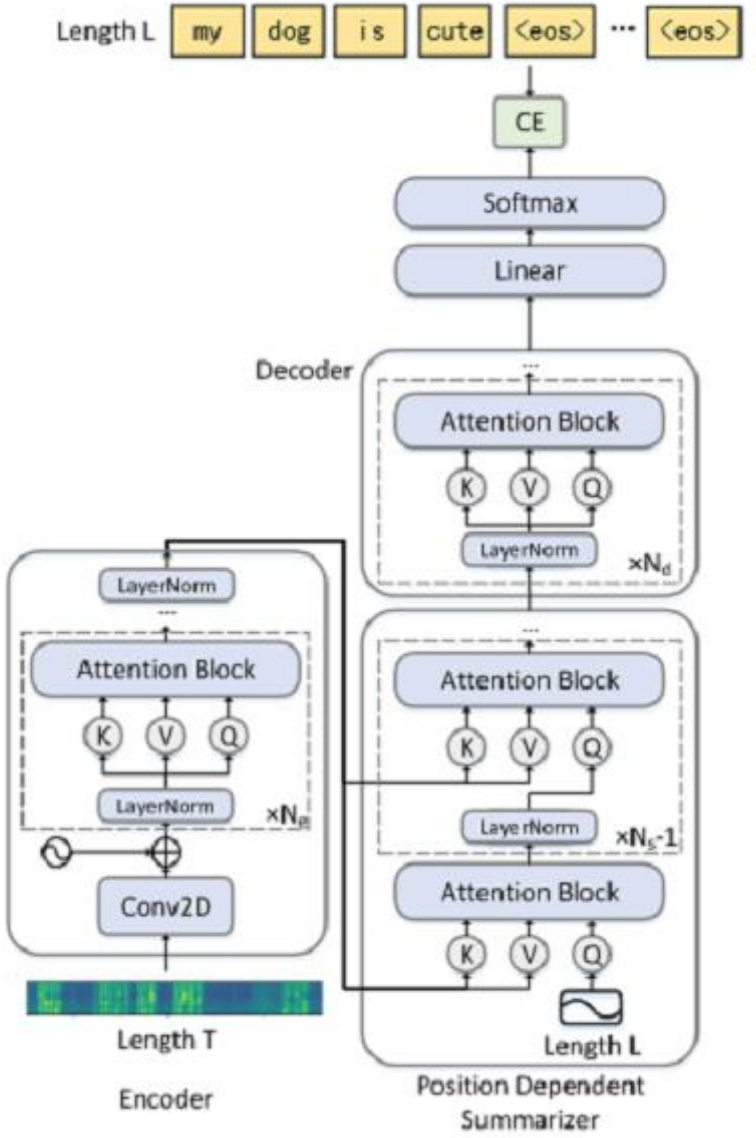

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)


<s> x1 x2 x3 x4 ... ... xn
x1
x2
x3
x4
xn
<e>

非自回归语音识别方法

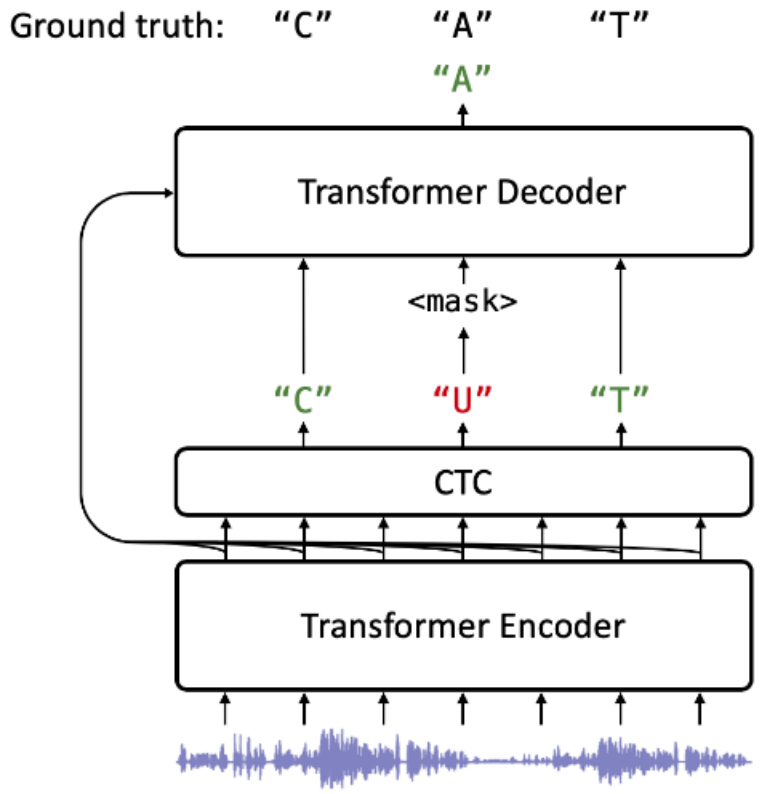
- 直接解码式非自回归

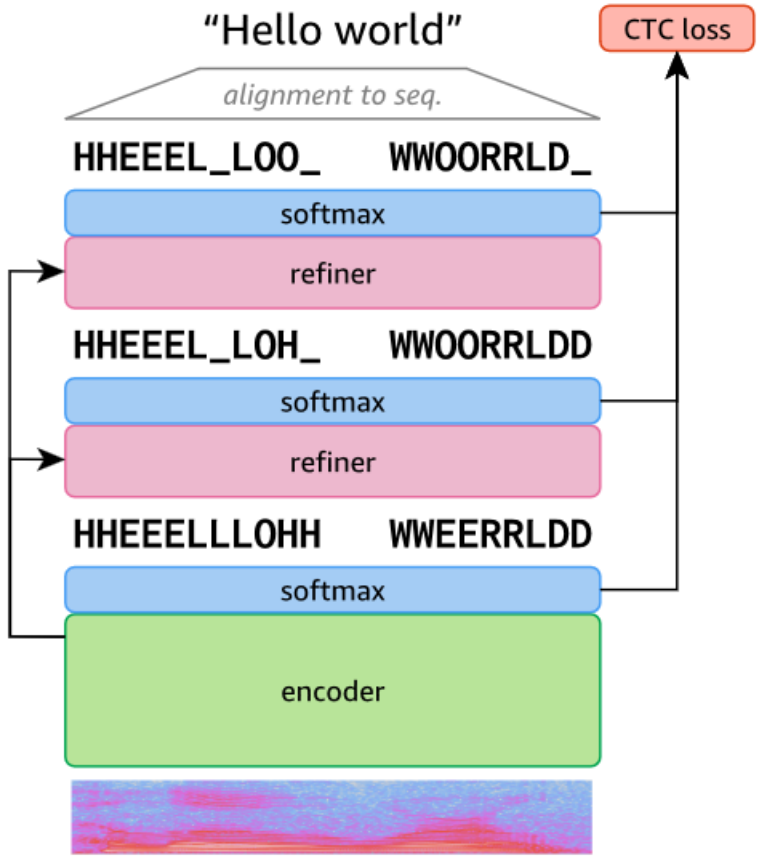




非自回归语音识别方法

- 纠错式非自回归





非自回归模型的训练方法

- 预训练模式
  - 将预训练好的自回归模型直接改成非自回归模型进行微调

- 随机遮蔽模式
  - 类似于BERT，通过随机掩蔽进行训练
  - 逐步提高掩蔽的概率

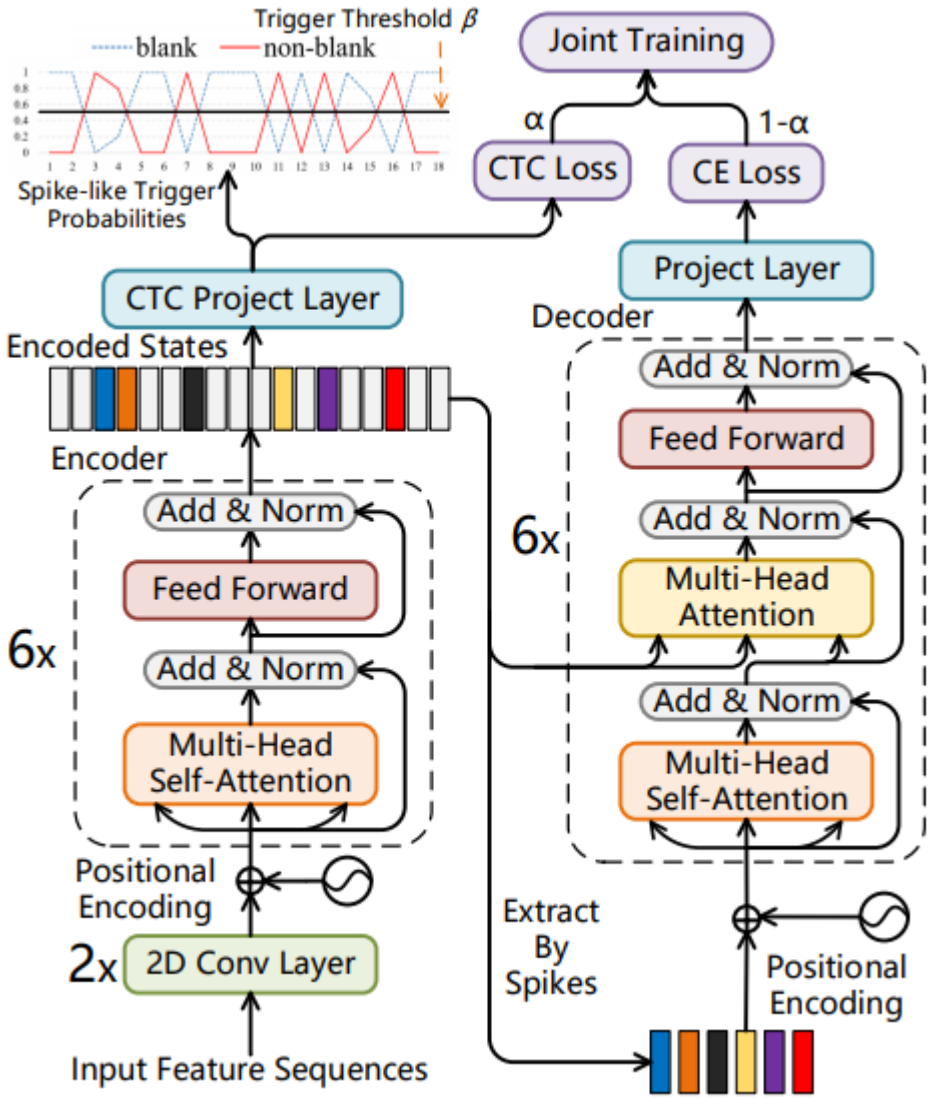
- 多任务联合模式
  - 联合CTC损失
  - 联合自回归损失或者其他正则化方法

工作1：尖峰触发非自回归语音识别方法

- 动机
  - 非自回归模型解码不能预测目标序列中的标记个数
  - 长度预测问题引发了大量的冗余计算

- 非自回归模型预测长度的方法
  - 长度预测网络 （类似 SCAMA）
  - 经验预测

工作1：尖峰触发非自回归语音识别方法



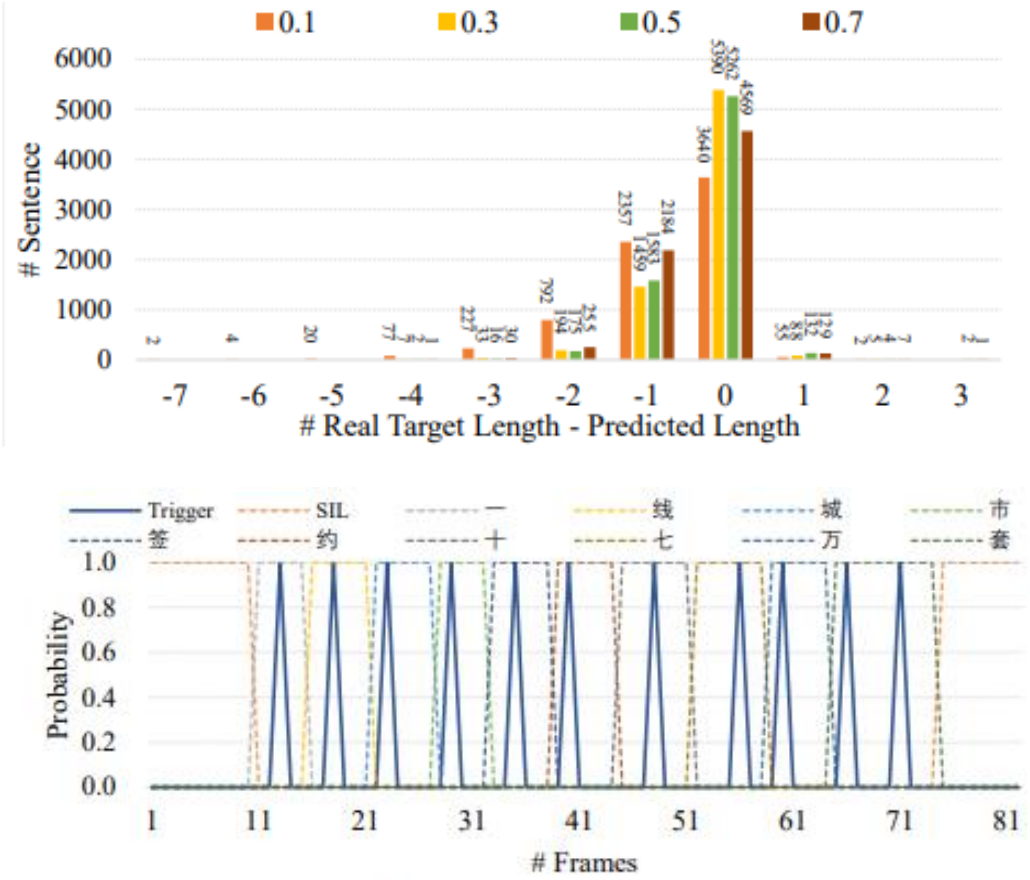
$$POS(i) = \begin{cases} triggerd, & 1 - p_b \geq \beta \\ ignored, & 1 - p_b < \beta \end{cases}$$

$$\mathcal{L} = \begin{cases} \alpha\mathcal{L}_{CTC} + (1-\alpha)\mathcal{L}_{CE}, & T' \geq T \\ \mathcal{L}_{CTC}, & T' < T \end{cases}$$

工作1：尖峰触发非自回归语音识别方法

Model	DEV	TEST	RTF
TDNN-Chain (Kaldi) [21]	-	7.45	-
LAS[22]	-	10.56	-
Speech-Transformer *	6.57	7.37	0.0504
SA-Transducer † [16]	8.30	9.30	0.1536
SAN-CTC * [23]	7.83	8.74	0.0168
Sync-Transformer † [24]	7.91	8.91	0.1183
NAT-MASKED * [11]	7.16	8.03	0.0058
ST-NAT(ours)	6.88	7.67	**0.0056**
ST-NAT+LM(ours)	**6.39**	**7.02**	0.0292

* These models are re-implemented by ourselves according to the papers.
† We supplement the RTF of our previous two models.

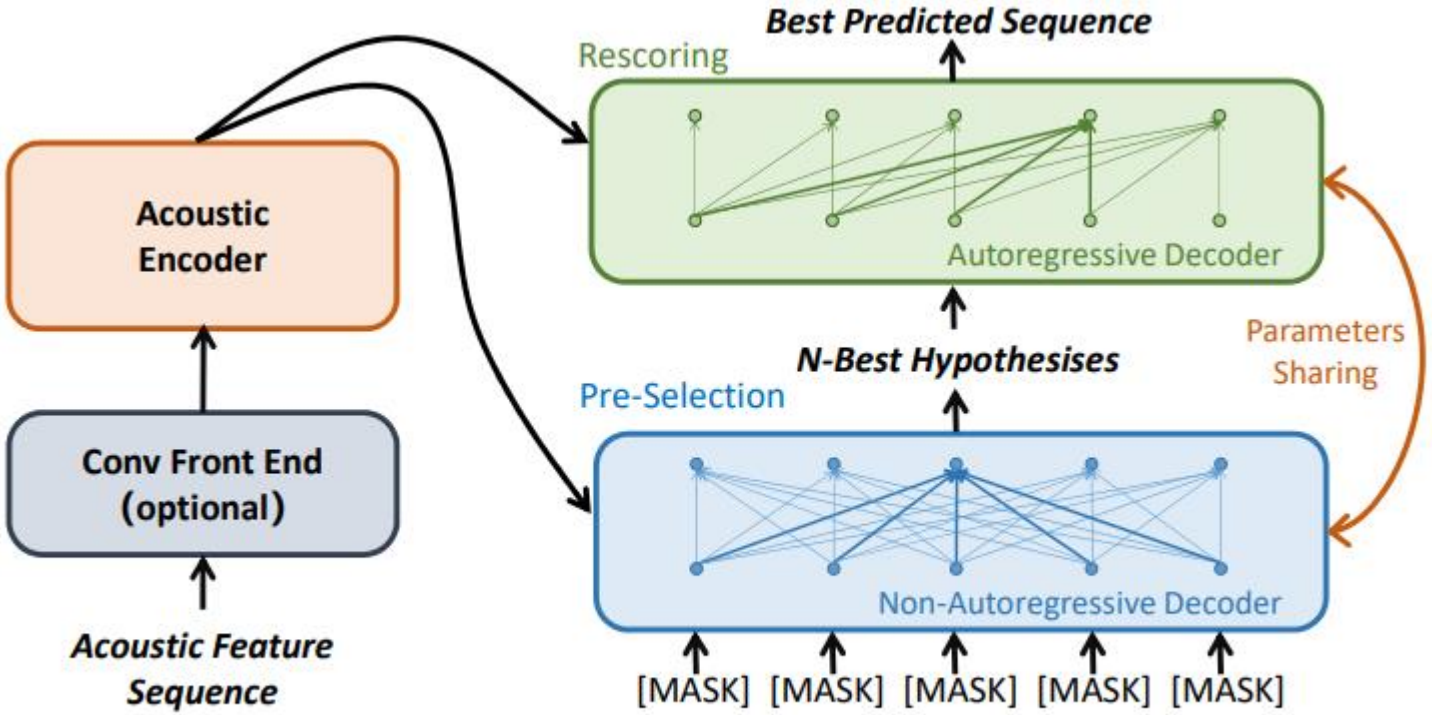


(a) The realtionship bettween trigger and word boundaries

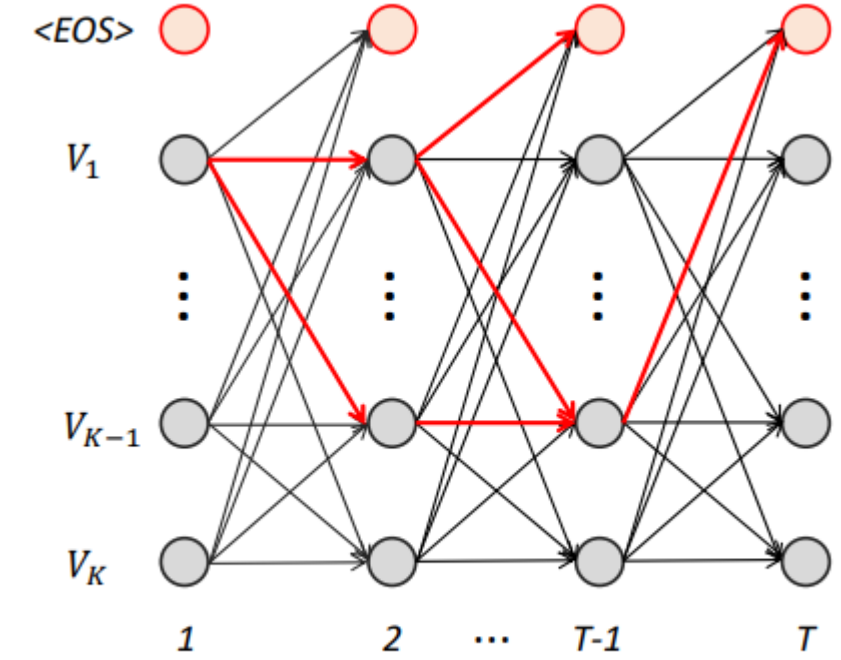
工作2：混合自回归与非自回归语音识别方法

- 动机
  - 非自回归模型难以训练
    - a. 更多的迭代次数
    - b. 辅助训练 （联合CTC）
  - 非自回归模型和自回归模型之间的性能差距
    - a. 辅助训练
    - b. 为解码器提供初始信息
- 思路
  - 共享 AR 与 NAR 部分参数 （加速收敛）
  - 两步解码 （性能提升）

工作2：混合自回归与非自回归语音识别方法



(a) The Structure of Two-Step Non-Autoregressive Transformer

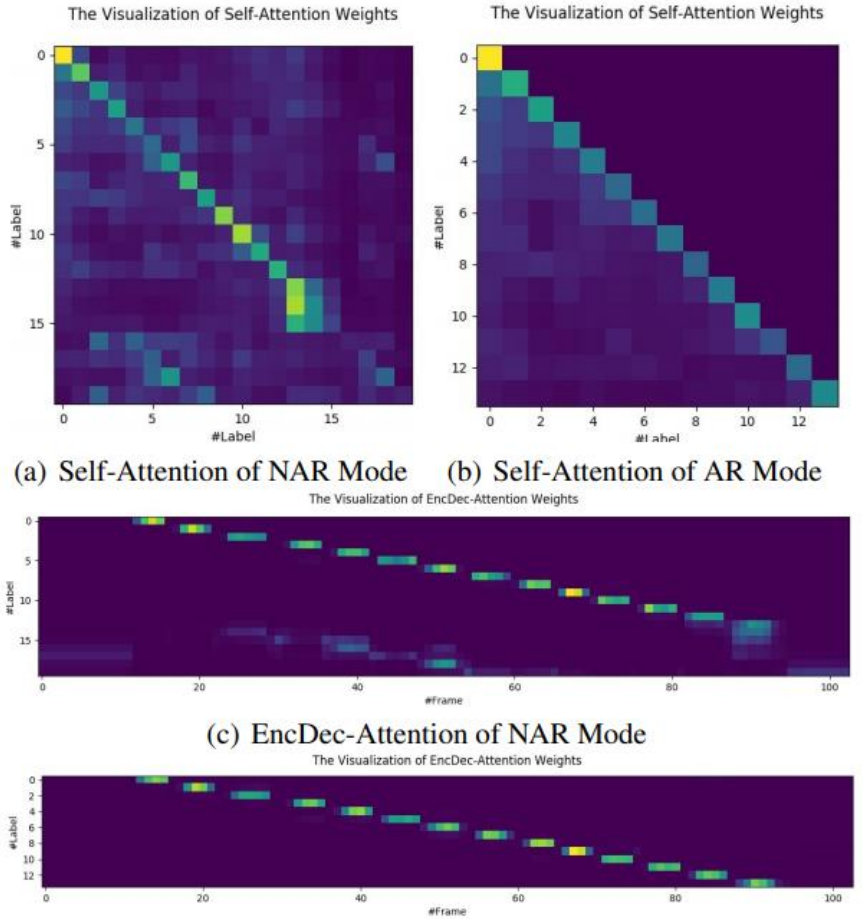


(b) The First-Step Inference Graph

工作2：混合自回归与非自回归语音识别方法

Model	LM	Dev	Test	RTF
A-FMLM(K=1) [10]	w/o	6.2	6.7	-
Insertion-NAT [25]	w/o	6.1	6.7	-
LASO-big [15] ◊	w/o	5.8	6.4	-
CASS-NAT [26] ◊	w	5.3	5.8	-
CTC-enhanced NAR [13] ◊	w/o	5.3	5.9	-
ST-NAT [14]	w/o	6.9	7.7	0.0056
ST-NAT [14]	w	6.4	7.0	0.0292
AR-Transformer-Small (34M)	w/o	5.3	5.9	0.0557
AR-Transformer-Middle (59M)	w/o	5.2	5.7	0.0613
AR-Transformer-Big (87M)	w/o	**5.0**	5.6	0.0721
HANAT-Small (34M) †	w/o	5.8	6.4	0.0054
Two Step Hybrid Inference	w/o	5.4	5.9	0.0173
HANAT-Middle (59M) †	w/o	5.4	6.0	0.0063
Two Step Hybrid Inference	w/o	5.2	5.7	0.0176
HANAT-Big (87M) †	w/o	5.3	6.0	0.0077
AR Inference	w/o	5.2	5.7	0.0735
Two Step Hybrid Inference	w/o	5.1	**5.6**	0.0185

◊ These models additionally use speed-perturb to augment the speech data.



(a) Self-Attention of NAR Mode (b) Self-Attention of AR Mode

(c) EncDec-Attention of NAR Mode

(d) EncDec-Attention of AR Mode

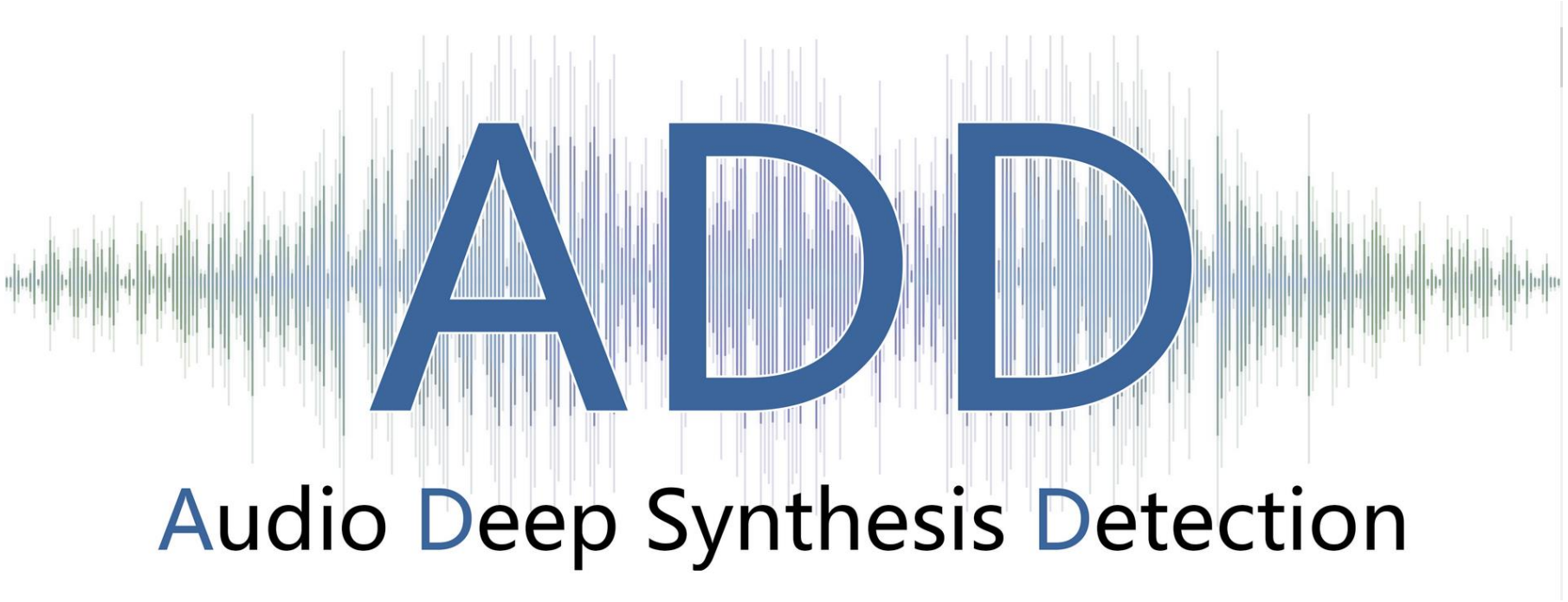
总结与展望

- 非流式解码方法有可能被非自回归模型颠覆（速度极快）

- 待解决的问题：
  - 消除自回归模型与非自回归模型之间的性能差距
  - 非自回归模型与流式任务的结合
  - 构建系统级的非自回归语音识别解决方案
    - 输出序列长度预测
    - 快速的准确解码方法
    - 热词定制等工业化问题

ICASSP2022 ADD Challenge

- 网址：http://addchallenge.cn
- 内容：合成音频检测与生成



谢谢大家的聆听！