更好的中英文混合语音识别系统-商迎新、王伟戌、王强强

1 研究背景

语音作为人与人交流的直接媒介，承载着人们日常生活中的大部分信息来源。基于近年来通信技术与物联网的发展，各式各样的语音助手、智能家具等软硬件层出不穷，人机交互技术的发展及人们对其需求日益攀升。语音识别技术在人机交互上扮演着重要的角色，任何因其导致的识别错误都可能在人机交互系统中的各个模块上传播，并最终导致交互失败。因此针对语音识别的研究具有重要的学术价值和应用价值。

混合语言现象常常出现在能够流利使用多种语言的群体中。英文作为全球的通用语言，就时常以各种形式与其他语言混合在一起。然而现有的大多数最先进的语音识别系统都专注于单语种语音识别，即它们一次只能处理一种语言，这样的系统无法识别中英混合语言的语音。随着语音技术开始渗透到人类生活的方方面面，混合语言的现象受到越来越多的关注。因此，开发用于中英文混合语言的自动语音识别(CSSR)系统尤为重要。

2 行业发展现状

中英文混合语音识别算法属于多语言语音识别领域。但与常规多语言语音识别不同，常规多语言语音识别仅针对一句话中出现一种语言，而混合语言语音识别则是指同一句话中说话人会在两种语言间切换使用。尽管语言学家对混合语言现象已经研究了长达半个多世纪，随着近年来语音技术的不断突破，对混合语言语音识别的研究近二十年才被人们重视。针对中英文混合语音识别也是近十多年来才开始研究。其技术难点主要表现为：嵌入语受主体语影响形成的非母语口音现象严重、不同语言音素构成之间的差异给混合声学建模带来巨大困难、带标注的混合语音训练数据极其稀缺。传统语音框架基于单一语种基础建模单元，如汉

语是基于拼音的声母韵母、英语则是英文的音素，这种技术架构对指定语种的语言学知识依赖较大，难以扩展到多语种识别 。由于不同语言之间的声学单元相互独立，且声学属性不同，常规基于声学单元建模的 DNN-HMM 语音识别模型无法很好的建模不同语言之间声学属性的联系。而端到端模型无需对于声学单元建模，转而采用字符建模，模糊了建模单元与声学属性之间的关联。并且由于端到端模型能够考虑帧的上下文信息，可以有效建模语言转换点的声学属性。因此最近几年的研究[1][2]偏向于采用端到端方式搭建混合语言语音识别系统。

基于深度学习的端到端模型灵活且复杂，相较于传统语音识别，融合多任务学习也能够提升模型性能。考虑到混合语言语音识别系统的特有属性，有学者提出可以鉴于 LID 模型[3][4]能够判别语言之间的差异性，以进行分类。中英文混杂识别联合语种识别受到越来越多的关注，在识别文本内容的同时进行语言分类，以增强对不同语言的分辨能力。本文从[5][6]获取灵感，在端到端网络模型基础之上添加语种信息进行联合训练，期望增强模型对不同语言的识别以及判别能力。

3 作业帮实践

3.1 实验细节介绍

为了便于后续讨论，首先介绍我们实验所采用的数据集、模型训练建模单元以及最终评价指标。

(1) 我们的训练语料为作业帮标注的约 1000 小时的老师上课的英文授课混合数据集，并在其中随机取 30 小时作为训练开发集，6.7 小时作为测试集，其余数据作为训练集。

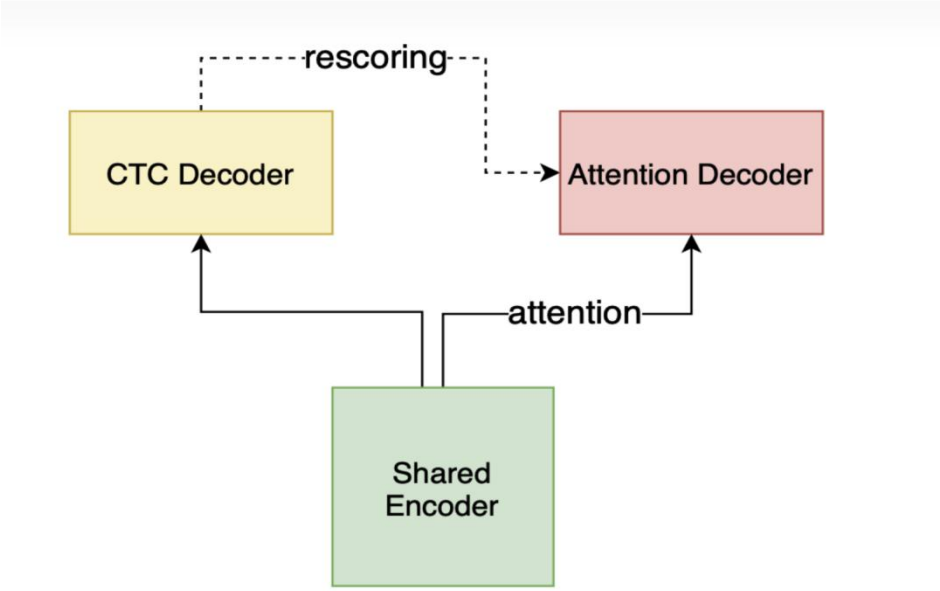
(2) 中文训练最小单元以汉字建模，但英文字母只有 26 个，因此以子词[7]替代字母作为英文的建模单元。从声学层面上，中文字符对应的音频长度远大于英文字母对应的长度，因此采用子词能够增大英文建模单元所对应的声学时长，从而减小中英文建模之间的差异，增强模型训练鲁棒性。

(2) 在语音识别系统的评估中，若单元为字符，则该错误率称为字符错误率 (Character Error Rate, CER)。若单元为单词，则称为词错误率 (Word Error Rate,

WER)。中英文混合语音识别系统由于包含两种语言，则中文部分计算 CER，英文部分计算 WER，称为混合错误率（Mix Error Rate, MER）。

3.2 基线识别模型

混合语言语音识别本质上还是语音识别任务，为了方便说明结果，我们的基线识别系统采用语音识别框架 Wenet[8]，Wenet 网络结构设计借鉴 Espnet 的 joint loss 框架，这一框架采取 Conformer Encoder + CTC/attention loss，利用帧级别的 CTC loss 和 label 级别 attention-based auto-regression loss 联合训练整个网络。这一框架是目前语音领域比较流行的框架之一。我们的实验也在其网络模型基础之上进行改进并对比实验效果。



3.3 语种信息联合训练

将语种信息加入中英文混合语言语音识别的网络模型训练，设 X 为声学模型的特征序列，Y 为模型的预测输出的中文字符或者英文字词，Z 为语种预测的输出类别，则最终的模型损失函数由之前的公式(1)增加为公式(2)。

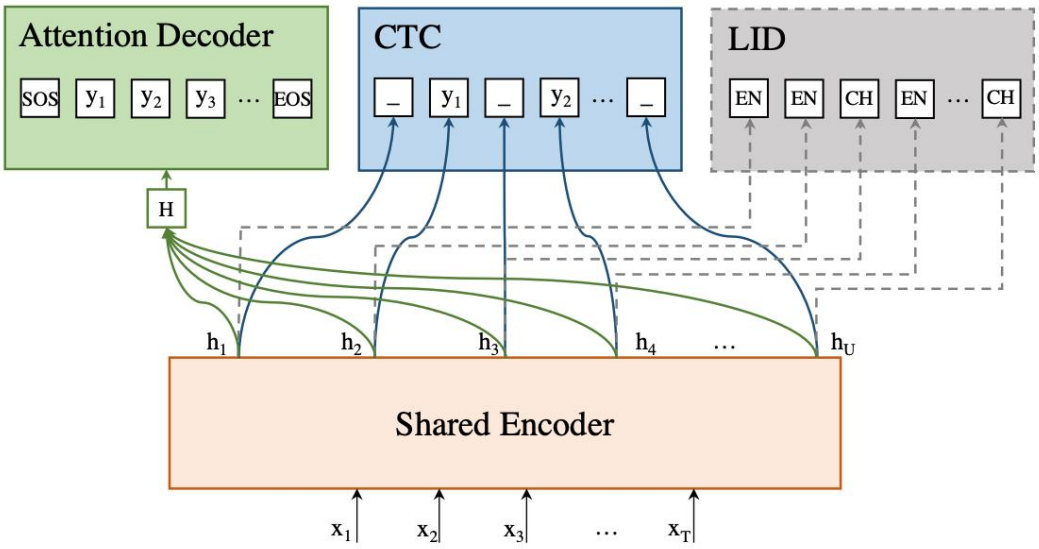
$$L_{MTL}(Y|X) = \lambda_1(Y|X) + (1-\lambda_1)L_{ctc}(Y|X) \quad (1)$$

$$L_{MTL}(Y|X) = \lambda_1(Y|X) + (1-\lambda_1)L_{ctc}(Y|X) + \lambda_2 L_{lid}(Z|X) \quad (2)$$

其中λ_1、λ_2分别模型训练的超参数 ，基线模型固定参数，在后续实验中，λ_1设置为 0.7。

3.3.1 帧级别语种 loss 联合训练

在混合语言语音识别中，不同语言的具有相似发音的单元很有可能被识别错误，因此我们考虑对识别系统加入语言识别。整体框架如下图所示，Wenet 编码器由基于 Attention 的解码器、CTC 和 LID 模块共享。它将输入序列 X 转换为高维特征 H。基于 Attention 的解码器和 CTC 生成输出序列 Y，而 LID 模块则输出每一帧的语种 ID。

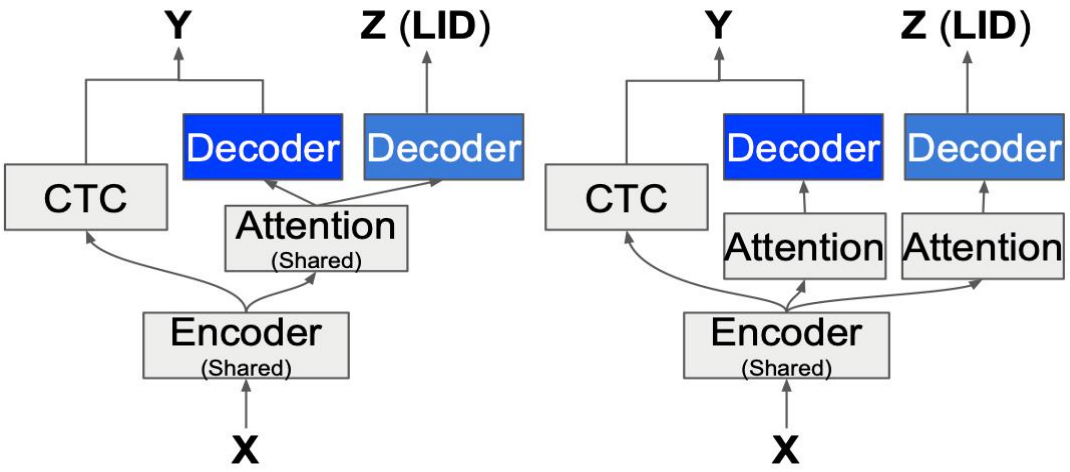


我们预先将数据集进行帧级别语种对齐，在实验中，我们利用的是之前训练过的中英文混合数据得到的 chain 模型，得到中英文混合数据的帧级别 LID-label，在训练过程中，将语种信息的预测和 label 的交叉墒作为语种的损失函数进行联合训练，最终目标函数为公式（2）所示。

3.3.2 token 级别语种 loss 联合训练

考虑到帧级别模型语种信息获取需要提前对中英文混合数据进行对齐，如果对齐不准，最终错误会累积。虽然模型在解码时，可以推断出语种信息，但是在

模型训练中未加入此信息，因此在这直接进行 token 级别的语种信息联合训练，网络框架如下图所示。



在这里，我们探讨了两种 LID 的 loss 添加方式，一种是与语音识别任务共享相同的注意力模型（LID-shared），另一种是单独学习一个独立的注意力模型（LID-indep）。两种方法都使用与等式 (2) 中相同的目标函数。

3.3.3 语种信息联合训练实验总结

1、3.3.2 介绍过，token 级别的语种信息添加有两种方式，即是否与识别系统共用 Attention，为了更有效率的对比两种方式的优劣性，我们先在小数据集上进行对比，在 1000 小时训练集中随机抽取 70 小时作为训练集，进行实验对比，效果如下表：

Model	λ_2	Test(%)
CTC+Attention	-	MER: 10.881
LID-shared	0.2	MER: 10.986
LID-indep	0.2	MER: 10.703

由以上表格可以看出，LID-indep 效果较好。因此后续添加 token 级别的语

种信息时，都采用 LID-indep 方式 ，并将系数设置为 0.2。

2、利用全部数据进行训练,为了方便进行记录，将帧级别添加语种信息训练模型称为 LID-frame。为了验证加入语种信息训练正确，我们还统计了 LID 的识别准确度 T-LID-ACC。以免中英文占比分布不均匀，这里统计的是 F1 score 。实验结果如下表:

Model	λ_2	Test(%)	T-LID-ACC(%)
CTC+Attention	-	MER: 5.843	-
LID-indep	0.2	MER: 5.740	98.8
LID-frame	0.2	MER: 5.901	97

从以上表格来看，LID-indep 在测试集上效果要优于基线模型，相对提升约 1.76%，并且语种判别准确度也达到了 98%，证明了添加 token 级别语种信息联合训练的有效性;同时 LID-frame 效果要比基线系统差, 分析原因可能有以下几个方面：(1) 帧级别的对齐 label 构建存在误差，传统混合语言语音 chain 模型的识别效果较差，会造成错误累积；(2) 在训练过程中，Wenet 会对数据进行语速增强，即随机对音频进行(0.9,1.0,1.1)倍变速，由于帧级别的 label 需要提前对齐，因此训练过程中只能关闭掉数据增强部分，模型鲁棒性变差，因此效果要低于基线模型。

3.4 混合语言错误类型分析

在验证添加语种信息训练的有效性基础之上，我们进行识别 case 分析，统计错误率分类占比，确定下一步优化策略。下表格是中英文错误单独统计的 MER。

Mix Error Rate, MER(%)	Character Error Rate, CER(%)	Word Error Rate, WER(%)
5.740	4.818 Sub：3.3 Ins:0.735 Del:0.775	12.512 Sub:9.015 Ins:1.215 Del:2.281

由上表格可以看出，混合语言语音识别中，无论是中文还是英文，均是替换错误占比较高，因此我们统计下替换错误的类型如下表（其中 E 代表英文，C 代表中文，$S_{E\to E}$ 则代表英文单词替换为英文单词）

Sub	Test(%)
$S_{E\to E}$	7.8
$S_{E\to c}$	1.1
$S_{c\to E}$	2.87
$S_{c\to c}$	0.15
SAME_SUB	3.729
DIFF_SUB	0.268

经过细致的错误分类，我们可以看出，不同语种之间的相互替换占比较低，替换错误主要发生在同类语种之间，尤其是英文单词替换为英文单词的占比较高为 7.8%，接下来我们通过添加语言模型辅助，期望降低英文同类语言的替换错

误。

3.5 语言模型增强

WeNet 中的 LM 支持方案，其中语言模型需要自己构建，依靠 ctc wfst search 生成 n-best， wfst search 为依靠传统解码图的传统解码方式。需要注意的是由于我们的建模单元是汉字加英文子词，因此构建的词典 L 是汉字映射为词语，英文子词映射为英文单词，英文词典的构建需要和训练时分词方式保持一致。T 为端到端训练时的建模单元 ，G 为语言模型即将 n-gram 的语言模型转为 WFST 形式的表示，最终构建 TLG 进行解码。以下是在 LID-indep 模型基础之上添加语言模型的效果。

Model	Test(%)
LID-indep	5.74
LID-indep + TLG	5.587

由上图表明，我们添加 TLG 实验的有效性，在测试集上错误率相对下降约 2.71%， 同时计算下测试集的替换错误如下表

Sub	Test(%)
SAME_SUB	3.388
DIFF_SUB	0.223

添加 TLG 同时也降低了中英文同类替换错误，由之前的 3.729 下降到 3.388，0.268 下降到 0.223，分别相对下降约 9.14% 和 2%，降低了混合语言测试集的替换错误。进一步验证了添加语言模型的有效性。

3.6 最终实验对比

在模型训练期间，词典是由训练文本产生，因此训练文本和开发集中几乎不含有<UNK>,在实际业务场景中是不合理的，因此在后续我们添加了约 50 小时的带有<UNK> 的数据，来增强模型鲁棒性，实验步骤和前面介绍的一致。最终实验结果如下表所示。

Model	Test(%)
CTC+Attention	5.843
CTC+Attention+LID-indep	5.743
CTC+Attention+LID-indep+<UNK>数据	5.56
CTC+Attention+LID-indep+UNK 数据 + TLG	5.42

另外，考虑到在实际的业务场景中，我们几乎不关注感叹词的识别情况，也不在意<他她它>的对错与否，因此将参考文本和识别结果中的感叹词过滤，统一<他她它>，重新将基线模型效果和最终实验效果进行测试，结果如下表所示

Model	Test(%)
CTC+Attention	MER：5.46 CH_WER 4.52 EN_WER 12.14
CTC+Attention+LID-indep+UNK 数据 +TLG	MER：5.08 CH_WER 4.23 EN_WER 11.14

最终将感叹词去掉后，和基线模型相比较，混合错误率相对降低约 6.96%，中文错误率相对降低约 6.41%，英文错误率相对降低约 8.24%。

4 总结与展望

在优化中英文混合语言识中，我们通过三个方面来提升中英文混合的识别效果。第一是模型训练层面，在 Wenet 的基础之上，我们对比了不同语种信息加入方式的优劣性，并从中选出最适合匹配基线模型的方式，测试集效果提升相对约 1.76%；第二是数据方面，为了更贴合实际业务场景，很多未在模型训练词典中的词可以识别出来，因此我们加入了部分<UNK>数据，进一步提升识别系统的可用性，相对提升约 3.1%;第三是考虑到中英文语言文本的连贯性，进一步通过语言模型来增强混合语言语音识别模型，构建 TLG，进一步相对提升约 2.5% 。最终相对基线模型提升约 7.8%。最终在实际应用方面，我们去除了感叹词和将<他她它>进行统一后，对比测试基线模型效果，从整体对比，混合错误率相对降低约 6.96%，中文错误率相对降低约 6.41%，英文错误率相对降低约 8.24%。

同时，实验也还有很多不足之处，后续会考虑从不同训练方式层面来提升中英文混合语言语音识别模型的识别效果。例如训练模型参数调优、预训练[9]以及无监督[10]的方式生成大量混合数据文本等。

5 参考文献

[1] Shan, Changhao, et al. "Investigating end-to-end speech recognition for mandarin-english code-switching." in Proc. of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019

[2] Emre Yılmaz, Samuel Cohen, Xianghu Yue, David van Leeuwen, and Haizhou Li, "Multi-graph decoding for code-switching ASR," in Twentieth Annual Conference of the International Speech Communication Association (INTERSPEECH). ISCA, 2019, pp. 3750 – 3754.

[3] Chan, Joyce YC, et al. "Detection of language boundary in code-switching utterances by bi-phone probabilities." in Proc. of the 2004 International Symposium on Chinese Spoken Language Processing. IEEE, 2004.

[4] Weiner, Jochen, et al. "Integration of language identification into a recognition system for spoken conversations containing code-switches." Spoken Language Technologies for Under-Resourced Languages. 2012.

[5] Zeng Z, Khassanov Y, Pham V T, et al. On the end-to-end solution to mandarin-english code-switching speech recognition[J]. arXiv preprint arXiv:1811.00241, 2018.

[6] Luo N, Jiang D, Zhao S, et al. Towards end-to-end code-switching speech recognition[J]. arXiv preprint arXiv:1810.13091, 2018.

[8] Yao Z, Wu D, Wang X, et al. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit[J]. arXiv preprint arXiv:2102.01547, 2021.

[9] Xinyuan Zhou, Emre Yilmaz, Yanhua Long, Yijie Li and Haizhou Li. "Multi-

Encoder-Decoder Transformer for Codeswitching Speech Recognition." in Proc. of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH). ISCA, 2020.

[10] Guo, Pengcheng, et al. "Study of semi-supervised approaches to improving english-mandarin code-switching speech recognition." arXiv preprint arXiv:1806.06200 (2018).