ISCAS 中国科学院软件研究所学术年会’2024
暨重点实验室科技活动周

学术论文

Perceiver-Prompt: Flexible Speaker Adaptation in Whisper for Chinese Disordered Speech Recognition

江怡聪, 王天资, 谢旭荣, 刘娟, 孙伟, 燕楠, 陈辉, 王岚, 刘循英, 田丰

Conference of the International Speech Communication Association (Interspeech) 2024

联系方式: 江怡聪 jiangyicong231@mails.ucas.ac.cn

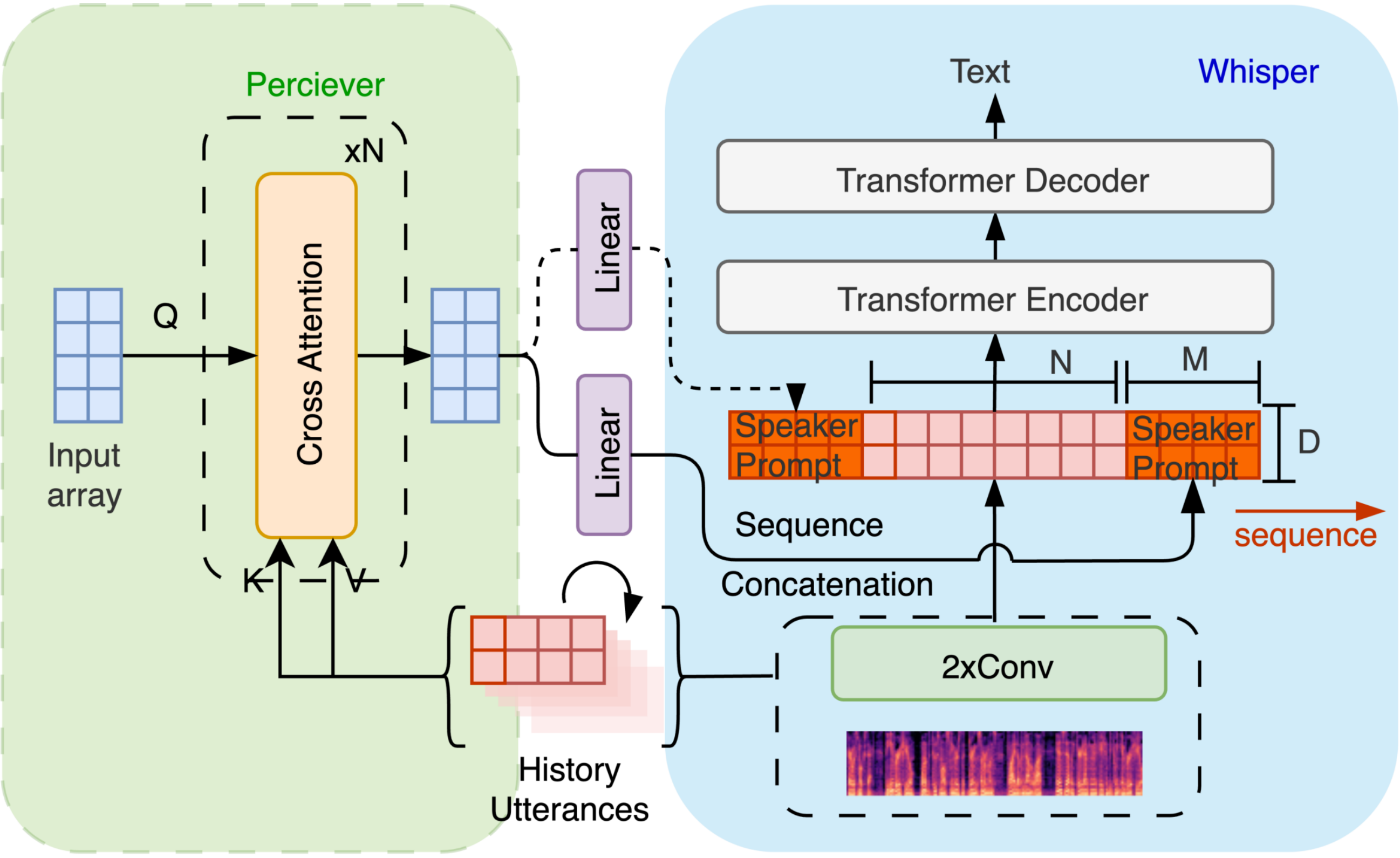
研究背景

障碍语音识别对于改善诸如构音障碍患者的生活质量具有深远影响，但与此同时也面临着包括数据有限、构音障碍与非构音障碍说话者之间的显著差异等挑战。因此，如何利用有限的数据尽可能改善障碍语音识别性能意义重大。

主要贡献

在新一代人工智能国家科技重大专项“面向神经系统疾病预警的智能人机交互关键技术”项目支持下，本文提出了**Perceiver-Prompt**，一种将提示学习应用于语音大模型上的说话人自适应方法。

- 该研究是首个将微调方法P-Tuning应用于语音大模型Whisper的研究；
- Perceiver-Prompt能够处理可变长度的输入并生成固定长度的说话人提示；
- 实验表明，我们的方法在字符错误率（CER）方面相对于微调后的Whisper减少了**13.04%**，特别是对于构音障碍严重程度最高的语音，CER相对减少了高达**51.38%**。



方法与结果

Perceiver-Prompt以其高度的灵活性在各种实验配置下均有着优秀的结果，并且优于现有的先进性能方法。我们在Perceiver的位置、与输入的连接位置、使用的历史语音数、说话人提示长度等不同的配置下进行了实验，同时采用了不同的训练方法。

$$\text{Prompt} = \text{Perceiver}(\mathbf{X_P}), \quad \text{Prompt} \in \mathbb{R}^{M,D}$$

$$\mathbf{X}_{\text{new}} = \text{Concat}(\mathbf{X}, (\mathbf{W}(\text{Prompt}) + \mathbf{b}))$$

$$\mathbf{X} \in \mathbb{R}^{N,D}, \mathbf{X}_{\text{new}} \in \mathbb{R}^{M+N,D}$$

Model	FDA-CER(%)				TSK-CER(%)			CER(%)
	F.1	F.2	F.3	F.4	T.1	T.2	T.3	
Conformer	33.1	43.8	6.3	1.8	9.6	14.9	14.5	12.9
TDNN	61.4	22.6	2.9	0.3	7.3	8.7	14.1	10.4
Whisper-medium	10.9	18.8	5.7	1.4	6.1	7.0	7.4	6.9
Whisper-iVector	13.5	20.7	6.2	1.4	7.9	8.9	6.2	7.6
Whisper-PP	**7.0**	**16.1**	**5.4**	**1.4**	**5.3**	**6.5**	**6.2**	**6.0**

Position		Input Utterances		Prompt length	FDA-CER(%)				TSK-CER(%)			CER (%)
Concat	Layer	history	Sto.		F.1	F.2	F.3	F.4	T.1	T.2	T.3	
End	Before encoder blocks	self	-	64	5.3	19.5	5.6	1.4	6.0	6.6	6.7	6.4
		self	-	32	7.0	16.1	5.4	1.4	5.3	6.5	6.2	**6.0**
		self	-	16	13.1	16.8	5.6	1.4	5.8	6.4	7.8	6.7
Beginning		self	-	32	9.4	18.6	5.6	1.6	6.1	7.1	6.9	6.7
Both sides		self	-		5.3	18.8	5.7	1.4	6.0	6.7	6.4	6.4
End		self+1	✗		5.3	19.2	5.7	1.4	5.9	6.6	6.8	6.4
		self+3			7.0	18.3	5.7	1.5	6.1	7.0	6.4	6.5
		self+5			5.3	18.1	5.7	1.5	6.3	6.3	6.3	6.3
		self+1	✓		12.3	19.2	5.4	1.5	5.8	7.0	8.0	6.9
		self+3			9.9	19.2	5.7	1.5	6.3	7.0	7.4	6.9
		self+5			8.2	18.5	5.7	1.4	6.2	6.5	7.1	6.6
End	log-mel	self	-	32	6.5	19.4	5.7	1.5	6.1	6.7	7.0	6.6

结果分析

- 将通过Perceiver-Prompt获得的说话人提示展开并进行t-SNE分析。
- 可以观察到通过我们的方法得到的说话人提示可以区分不同的说话人和不同程度的构音障碍。
- 通过添加更多的说话人历史信息可以增强提示对说话人的区分能力。
- 但更多的历史信息不一定能够保证在构音障碍语音识别任务中的性能提升。