SLT 儿童口语识别比赛系统介绍

王永庆 闫志勇 张俊博

参赛队员列表

DATA AUGMENTATION FOR CHILDREN'S SPEECH RECOGNITION THE "ETHIOPIAN" SYSTEM FOR THE SLT 2021 CHILDREN SPEECH RECOGNITION CHALLENGE

Guoguo Chen[*1], *Xingyu Na*[*2], *Yongqing Wang*[*3], *Zhiyong Yan*[*3], *Junbo Zhang*[*3], *Sifan Ma*[3], *Yujun Wang*[3]

[1]Seasalt AI LLC, USA
[2]Xiaoice, Microsoft Corporation, China
[3]Xiaomi Corporation, China

guoguo@seasalt.ai, asr.naxingyu@gmail.com
{wangyongqing3, yanzhiyong, zhangjunbo1, masifan, wangyujun}@xiaomi.com

前五位（标星号）作者按姓氏字母排序，其中有三位是来自小米的工程师

数据集

- 训练集 (8月发布)

数据集代号	时长（h）	说话人个数	年龄	风格
034	342	1997	18 ~ 60	朗读
011	28	907	7 ~ 11	朗读
018	28	158	4 ~ 11	口语
Opensrl(track2)	1374	SLR18-Thchs30、SLR33-Aishell、SLR38-Free-ST-Chinese-Mandarin-Corpus、SLR47-Primewords-Chinese-Corpus-Set-1、SLR62-Aidatatang_200zh、SLR68-Magicdata		

- 开发集

数据集代号	时长（h）	说话人个数	年龄	风格
Dev-011	-	20	7 ~ 11	朗读
Dev-018	-	5	4 ~ 11	口语

- 测试集(9月底发布)

时长（h）	年龄	风格
10	4 ~ 11	朗读口语各占一半 很可能和011、018同源

比赛策略

- 数据特点：训练集以成人为主，测试集全部是儿童
- 比赛的关键：解决训练集同测试集声学特征不匹配的问题
- 思路
  - 重点对儿童数据做更多的data augmentation，儿童数据可以超过成人数据

数据扩增

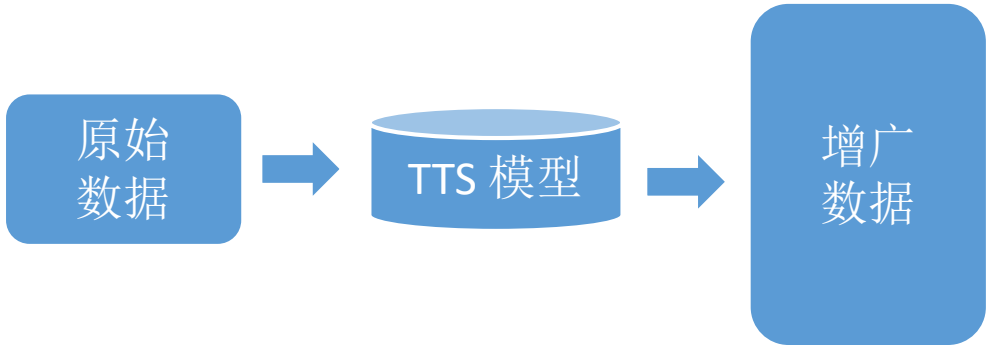
- 成人转儿童
  - 调Pitch，把成人数据改成听起来像儿童
- 语速 & 音量
  - 使用Kaldi的工具
  - 原语速的 85%、88%、90%、110%、112%、115%
  - 音量扩增倍数从0.125至2之间
- 混响
  - 用sox的reverb功能简单造了一些
- Spectral Augmentation
  - 19年google论文中的方法，Kaldi中有现成工具

Data Augmentation	Hours
Set A *	341.8
Set C1, C2 *	55.1
Set A, C1, C2 + rp + vp	396.9
Set A, C1, C2 + pp + vp	396.9
Set A + sp@{0.9,1.1} + vp	690.6
Set A + tp@{0.9,1.1} + vp	690.6
Set C1, C2 + sp@{0.85,0.88,0.9,1.1,1.12,1.15} + vp	342.7
Set C1, C2 + tp@{0.85,0.88,0.9,1.1,1.12,1.15} + vp	342.7
Total	3257.3

pp : Pitch perturbation
rp : Reverberation perturbation
sp : Speed perturbation, values inside the curly braces are different perturbation parameters
tp : Tempo perturbation, values inside the curly braces are different perturbation parameters
vp : Volume perturbation
* Our number is a little bit less than the official number

数据扩增 —TTS + NLG

- TTS 可以看作一种数据增广形式
  - 模型从训练样本中，学到如何增广（生成）数据
  - 相比于简单地改Pitch等方法，能生成更丰富的内容

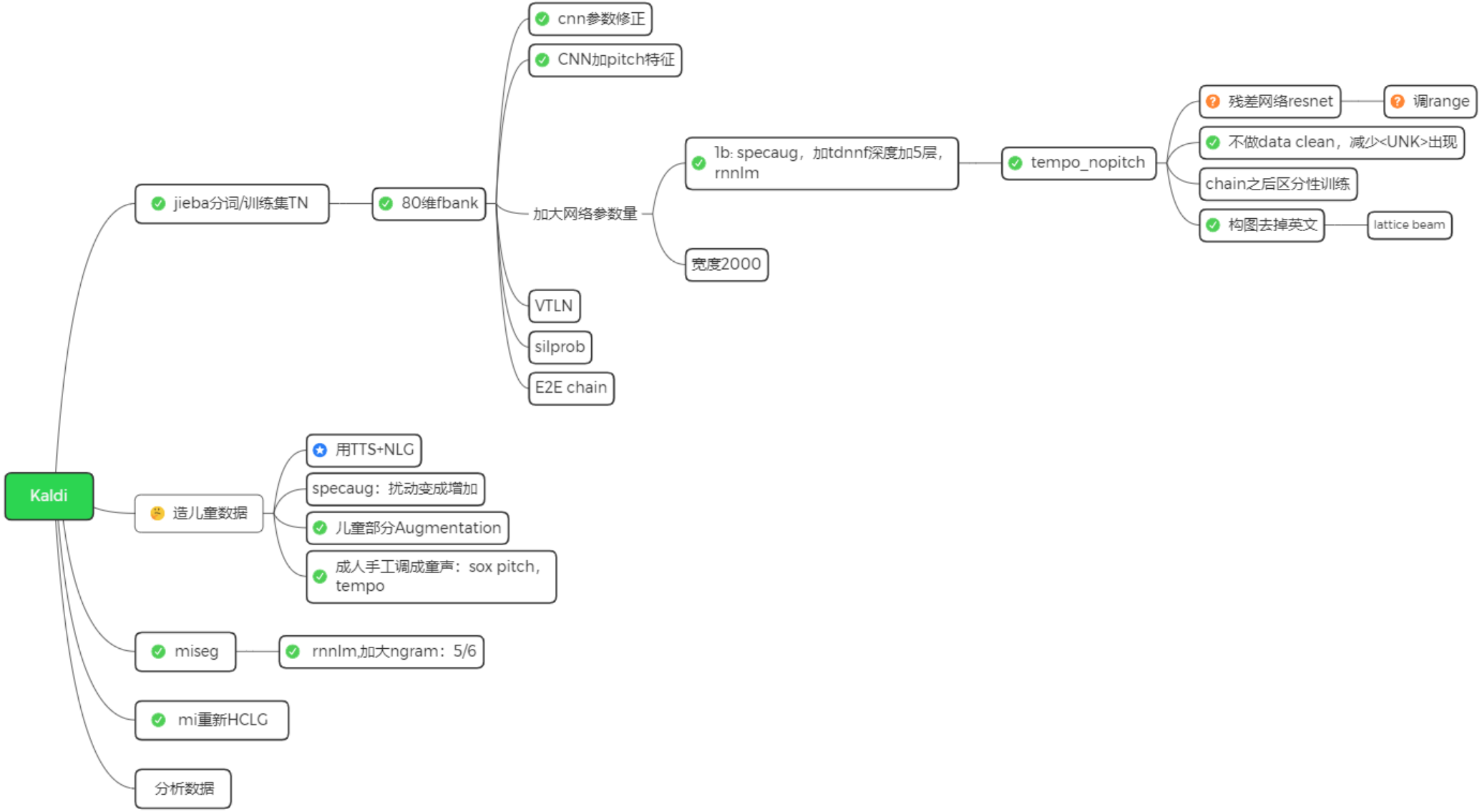


- 用NLG尝试解决口语文本不够的问题
  - 训练集以书面文本为主，但测试集有大量的口语文本
  - 尝试使用了GPT-2来生成口语文本

我也会唱歌曲我是
哦
你的话呀
你的是什么时候的
好不是不知道
你就是什么歌
我妈都是我不是一个小的那种
你们班的
嗯好的是什么

Kaldi框架

- kaldi工作细节



数据扩增测试结果

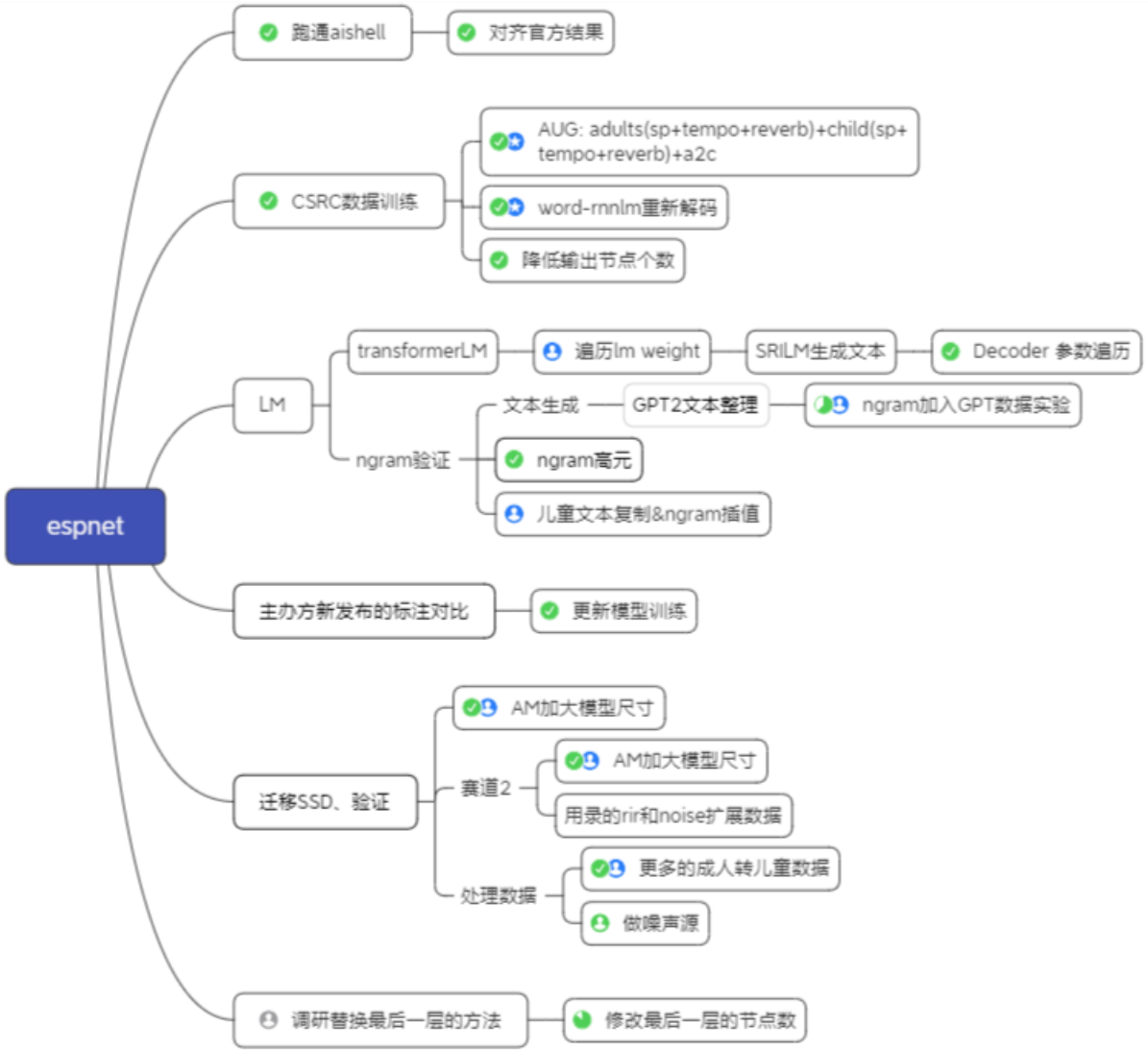
Kaldi-Exp	dev-011	dev-018	Average
Exp1: Baseline	10.2	22.36	16.28
Exp2: Baseline + 80-dim FBANK	10.22	21.95	16.09
Exp3: Baseline + 80-dim FBANK + SpecAug + depth tdnnf:17	9.95	21.66	15.81
Exp4: Exp3 + backward rnnlm: ngram order 4	9.19	21.66	15.43
Exp5: Exp3 + forward rnnlm: ngram order 4	9.12	20.98	15.05
Exp6: Exp3 + forward rnnlm: ngram order 5	9.09	20.45	14.77
Exp7: Exp3 + 3-dim pitch features	9.92	21.62	15.77
Exp8: Exp3 + 3-dim pitch features + forward rnnlm: ngram order 5	9.16	21.04	15.1
Exp9: Exp6 + *Set A, C1, C2*:rp,{sp,tp}@{0.85,0.9,1.1,1.15}	8.47	19.78	14.13
Exp10: Exp9 + remove clean + remove *Set A*:{sp,tp}@{0.85,1.15}	8.63	19.26	13.95
Exp11: Exp10 + *Set A, C1, C2*:pp + *Set C1, C2*:{sp,tp}@{0.88,1.12}	9.14	18.69	13.92
* Exp12: Exp9 + remove clean + *Set A, C1, C2*:pp + *Set C1, C2*:{sp,tp}@{0.88,1.12}	8.52	18.70	13.61

* Post-challenge experiment

Espnet模型结构与调参

- 端到端识别架构，模型结构采用 Conformer
  - [2005.08100] Conformer: Convolution-augmented Transformer for Speech Recognition
  - [2010.13956] Recent Developments on ESPnet Toolkit Boosted by Conformer
- 语言模型Shadow Fusion
  - 开发集上，RNN和Transformer都不如 4-gram 的效果好，CER 相差 0.2 个点
- 训练中的一些策略
  - 对开发集loss进行一些排序，筛选一些效果比较好的迭代模型做平均
  - Retrain的时候学习率和warmup的一些调优

- espnet工作细节


espnet
跑通aishell
对齐官方结果
CSRC数据训练
AUG: adults(sp+tempo+reverb)+child(sp+tempo+reverb)+a2c
word-rnnlm重新解码
降低输出节点个数
LM
transformerLM
遍历lm weight
SRILM生成文本
Decoder 参数遍历
ngram验证
文本生成
GPT2文本整理
ngram加入GPT数据实验
ngram高元
儿童文本复制&ngram插值
主办方新发布的标注对比
更新模型训练
迁移SSD、验证
AM加大模型尺寸
赛道2
AM加大模型尺寸
用录的rir和noise扩展数据
处理数据
更多的成人转儿童数据
做噪声源
调研替换最后一层的方法
修改最后一层的节点数

Q & A