语音识别核心引擎2012年12月实验性能汇总

一、要点

1、10月底至12月初进行了声学模型数据更新训练，词表的扩充和更新（8w8至11w8），语言模型过滤训练算法更新，叠加效果性能相比当前微信语音输入性能提升相对15%以上，详细实验结果以及分析见具体说明部分。

2、目前各个测试集和讯飞公开sdk对比性能见下表，红色代表我们优于讯飞性能，黄色代表讯飞优于我们。

	讯飞字错误率	2012.12月效果字错误率
常用语1900	6.7	6.5
短信2044	15.8	18.8
实网1	38.3	34.5
实网2	28.1	28.9
地图测试集	14.9	20.8
记事本测试集	11	10
领域相关测试集	31.2	30.3
快速测试集	19.6	21.4

从测试集上的结果反映出以下几点：

a) 在 qq 输入实网测试集上，我们和讯飞的模型都表现出较差的性能，说明讯飞的模型和我们过去的实网数据也具备一定的不匹配性

b) 在常用语，领域相关以及记事测试集上，我们和讯飞表现出类似的性能。在短信和地图测试集上我们有一定差距，其原因有主要是缺乏实际的地图和短信的数据。

c) 在快语速的测试集上，讯飞比我们有一定的优势

3、年前工作计划要点如下：

a) 目前这个版本已经更新至微信输入，语音提醒等其他业务，作为保底版本，接下来会陆续替换 QQ 输入法 QQ 语音服务还有微信各个服务

b) 1 月 15 日前进行算法部分实验，再出一个版本，预估时间是 1 月 15 号前

声学模型：pitch，silence 加载到词典中，决策树和数据的关系，纯中文音素的实验

语言模型：高阶模型实验，分类语言模型实验，过滤算法中算法的进一步更新

解码器：分类语言模型解码器构造，二遍 rescore 部分实验

c) 春节前需要根据标注的 speex 数据以及录制的记事本数据，进一步更新微信输入和记事本的声学模型部分，语言模型利用微信的文本语料，2-6 号再出一次新版本。

二、具体说明：

1、 测试集说明

a) 常用语 1900：自行录制，内容为日常用语

b) 短信 2044：自行录制，内容为念聊天记录或者短信

c) 地图测试集：自行录制，内容为还有地名的语句

d) 记事本测试集：自行录制，内容包含时间，提醒词以及事件

e) 领域相关测试集：自行录制，内容包含汽车，房产，风景，美食，医药，娱乐等领域

f) 快语速测试集：自行录制，内容为短信 2044 中一部分，但是语速偏快

g) 安卓输入法实网 1：2012 年 3 月左右于安卓 qq 输入法实网采集数据

h) 安卓输入法实网 2：2012 年 5 月左右于安卓 qq 输入法实网采集数据

以上测试集总计 26870 条目，约 19 小时，均为 16k，16bit 采样。

2、 基线系统性能

基线系统描述：

词典条目 8w8；

声学模型使用实网数据 700 小时以及 863 购买 pc 录音 700 小时，ML+HLDA+MPE 训练算法

语言模型使用 3 阶回退语言模型

	基线字错误率
常用语 1900	8.4
短信 2044	22.4
实网 1	35.6
实网 2	29.6
地图测试集	24.5
记事本测试集	16
领域相关测试集	36
快速测试集	26.8

3、 语言模型更新效果

更新描述：

词典条目扩充至 11w8，添加了更多领域相关的词根；

声学模型同基线系统

语言模型过滤算法中加入智能去重算法，屏蔽某一种语言现象在语料中频繁过多出现的问题，以及利用词典过滤语料中大量垃圾的更新算法。

	基线字错误率	语言模型更新效果	相对提高
常用语 1900	8.4	8.6	-2.38%
短信 2044	22.4	21	6.25%
实网 1	35.6	35	1.69%
实网 2	29.6	28.9	2.36%
地图测试集	24.5	24.5	0.00%
记事本测试集	16	12.1	24.38%
领域相关测试集	36	33.9	5.83%
快速测试集	26.8	25.5	4.85%

本次语言模型的更新，由于词表的扩大，语料去重等过滤算法的更新，测试集整体上是有一定的提高，对领域相关测试，短信测试集尤为明显，对常用语测试集有负作用，原因主要是这次语言模型插值里面压低了一些聊天内容的比重。

4、 声学模型更新效果

更新描述：

词典语言模型同基线系统

声学模型: 基于基线系统数据基础，添加了从海天数据公司购置的 2200 小时手机端录制数据，使用训练算法亦为，ML+HLDA+MPE

	基线字错误率	声学模型更新效果	相对提高
常用语 1900	8.4	6.7	20.24%
短信 2044	22.4	19.9	11.16%
实网 1	35.6	35.4	0.56%
实网 2	29.6	29.6	0.00%
地图测试集	24.5	21.2	13.47%
记事本测试集	16	12	25.00%
领域相关测试集	36	32.1	10.83%
快速测试集	26.8	21.5	19.78%

本次声学模型更新，由于 2200 小时手机录制信道的加入，对自录测试集均有明显的相对提高 10%-25%。同时，对 3，4 月份采集的两个 qq 输入法线上数据测试集提高很微弱。其原因主要是由于实网测试集中存在环境音，用户说话并不规则（可能有情绪伴随，一字一顿等现象），同所加数据并不匹配。由此回顾之前加入 700 小时实网数据进行训练，对实网测试集性能的提升也有相当 10%以上的提高。

5、 叠加更新效果

更新描述：

结合上述 3 和 4 的模型更新，具体性能如下：

	基线字错率	叠加更新效果	相对提高
常用语 1900	8.4	6.5	22.62%

短信 2044	22.4	18.8	16.07%
实网 1	35.6	34.5	3.09%
实网 2	29.6	28.9	2.36%
地图测试集	24.5	20.8	15.10%
记事本测试集	16	10	37.50%
领域相关测试集	36	30.3	15.83%
快速测试集	26.8	21.4	20.15%