**1 研究过程**

1.1 数据特点

1. 儿童语音普遍口齿不清晰
2. 部分音频本身存在清晰度问题，例如小朋友声音过轻等
3. 目前收录的语音存在不少无效音频，例如：空音频、极长极短音频等
4. 部分音频包含嘈杂的背景噪声
5. 部分音频包含较响的志愿者声音，例如：志愿者交谈、志愿者向小朋友提问题等
6. 部分声音清晰且表达正确的音频中，图片所对应的的答案不全，经常出现意思对了但是不符合标准答案的情况，应当扩充正确答案
7. 最初的音频由人工进行判定，内容包括：音频内容、拼音、是否包含志愿者声音、是否包含明显噪声、发音标准程度、认知正确性、音频有效性等方面

1.2 方法的调研与选择

1. DTW

DTW是最开始采用的方法，用于简单计算音频之间距离，工作原理是将两段音频按照一一对应的方式分割裁剪，计算每两段对应距离的总和，从而得到两段音频之间的总距离。

距离的计算方式可自定义，但对音频质量本身有很高的要求。这个方法的主要问题在于：

1. 需要儿童标准音频进行比对，但儿童标准音频本身难以定义。
2. 该方法对音频质量本身要求很高，对于标准普通话正确率尚可，但测试儿童语音得到的效果很不理想。
3. 每进入一段音频都要进行一次对比运算，耗时耗力。
4. ASRT

ASRT是GitHub上的一个开源的中文普通话语音识别项目，使用的训练集也是公开的普通话数据集，目前可以达到80%左右的识别正确率。

测试了ASRT的demo后，总结了以下问题：

1. 对标准普通话识别率尚可，儿童语音识别率很低。
2. 模型更偏向长句的识别，联系上下文可获得更好的识别率，而儿童语音基本都是词语词组，并不适用。
3. 对噪声的抗干扰性一般，这部分或许可以通过预处理进行规避
4. 如果要自己训练模型来投入使用，缺少大量的数据作为支撑（若之后全国数据足够，可以考虑训练针对儿童语音的识别模型），而市面上标准成人普通话公开数据集并不适用。

**2 目前采用的方法**

2.1 科大讯飞识别接口

现在市面上很难找到针对儿童语音识别的商业接口，而在普通话识别方面，科大讯飞是佼佼者，因此，我们采用讯飞的接口来作为识别方式。

在测试中我们发现，由于儿童语音本身口齿不清，加上本身都是短词短句，识别得到的结果并不十分理想。后来我们发现，对于许多错误的识别结果，尽管中文字本身识别错误，但是若将文字拆成拼音进行比较，却有一定的相似性，因此我们设计了基于拼音距离的滑动窗口算法来解决这个问题。

2.2 基于拼音距离的滑动窗口算法

对于一段音频，其对应的标准答案我们可以将其拆分成拼音，其对应的讯飞识别结果也可以将其拆分成拼音，对于这两段拼音我们利用滑窗思想可以计算出两段拼音之间的最小距离，作为标准距离。

例：标准答案【牙齿】，讯飞识别【这是牙此】，转拼音【ya-chi】 vs 【zhe-shi-ya-ci】，滑窗长度为2，依次为【zhe-shi】、【shi-ya】、【ya-ci】，分别与【ya-chi】求距离，最小距离作为最后判定的依据。

在此基础上，我们给这个距离设置了一个阈值（可以自行设置、更改），高于这个阈值就认为这个音频是错误的，反之正确。对于图片包括多组正确答案的情况，取距离的最小值作为标准距离。

后来，由于需要部分数据进行人工判断，因此，我们将单个阈值改为距离区间（可调节），区间内的部分最难判断，交给人工；把区间外最接近正确答案的部分判定为正确，最远离正确答案的部分判断为错误。

例：距离区间设定为[1, 2]，利用滑窗求得最小距离为0.9，则判定为正确；距离为1.5，则交给人工；距离为2.5，则判定为错误。

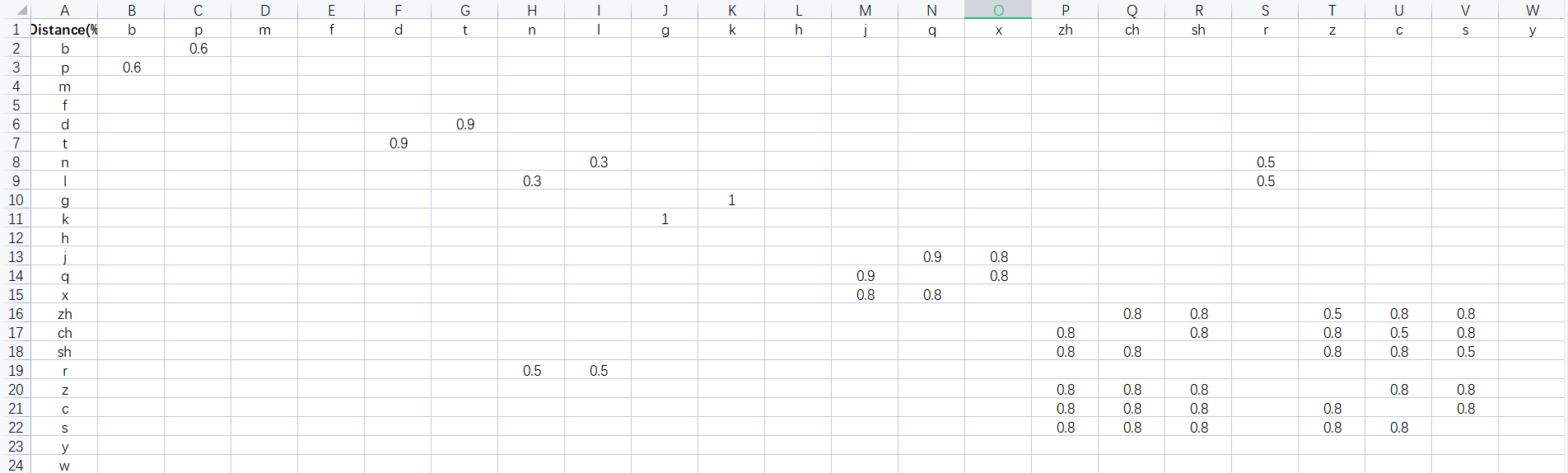
2.3 距离计算方法

最初的距离计算方式是最原始的编辑距离，即增删改的次数，我们在参考阅读了一些语言学的文献之后对其进行了更改：将单字拼音拆分成了声母、韵母两部分，并根据发音相似程度，设计了距离权值矩阵，分别求声母、韵母之间的编辑距离，并乘以权值，然后求和，得到最终距离。

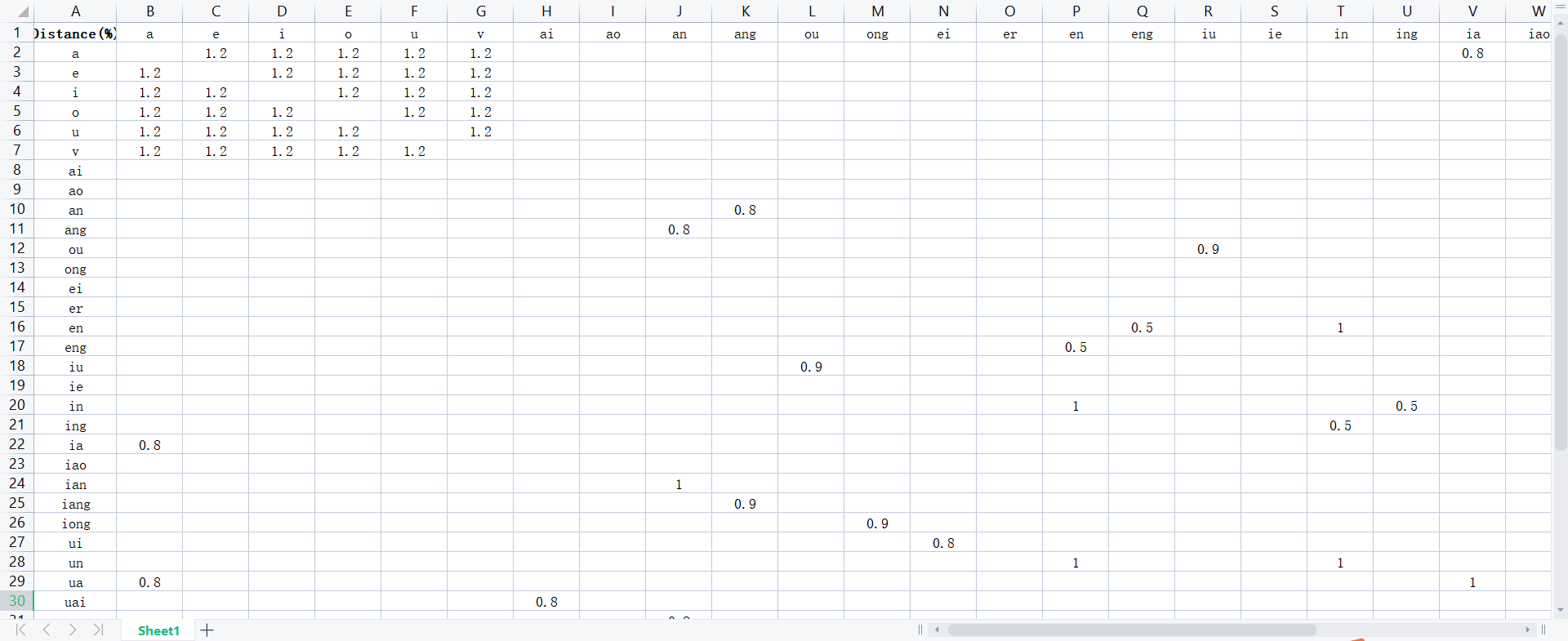
例：【beng】、【pen】的距离 = 编辑距离(b, p) \* 权值(b, p) + 编辑距离(eng, en) \* 权值(eng, en)。

该方法使得准确率进一步上升。但是由于这个权值表没有标准的定义和规定，未找到相关研究，因此该表是否合适，是否有进一步提升的空间以及如何提升还有待商榷。

距离权值矩阵示例：声母距离权值矩阵v2



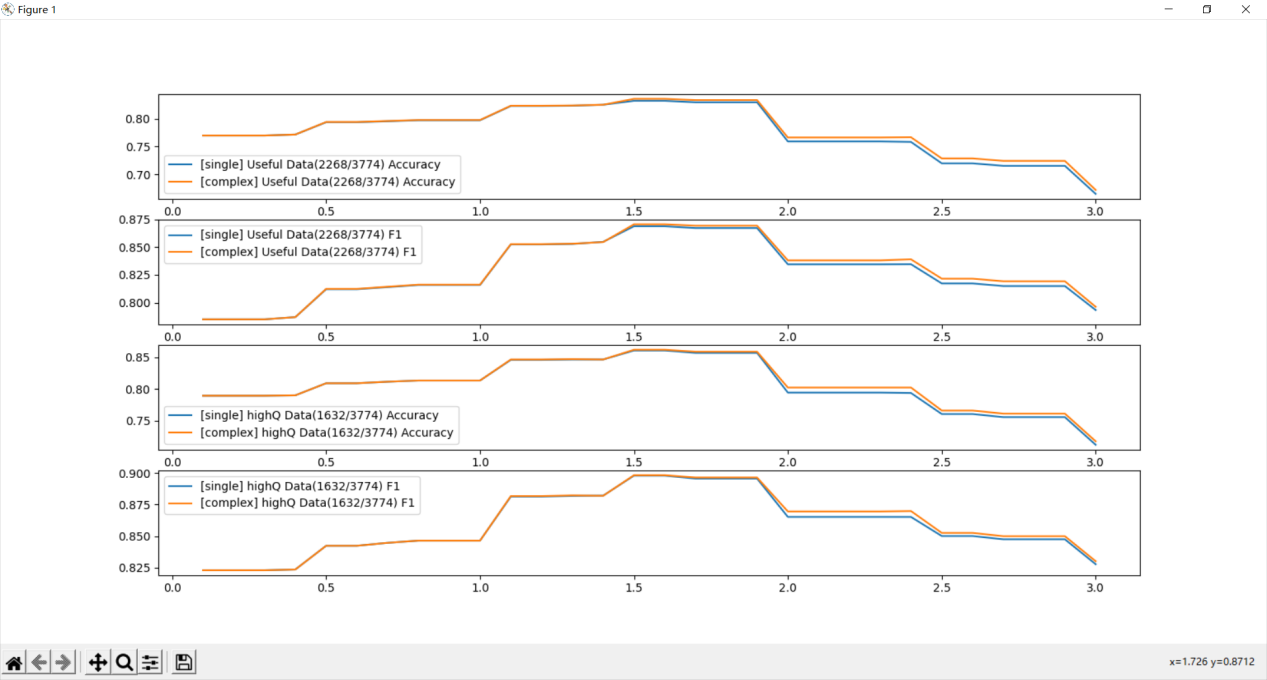
距离权值矩阵示例：韵母距离权值矩阵v2

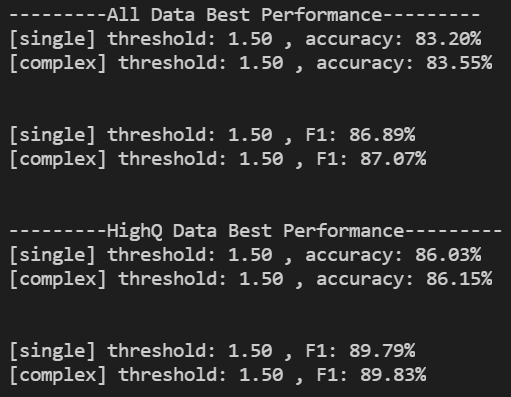


**3 实验结果**

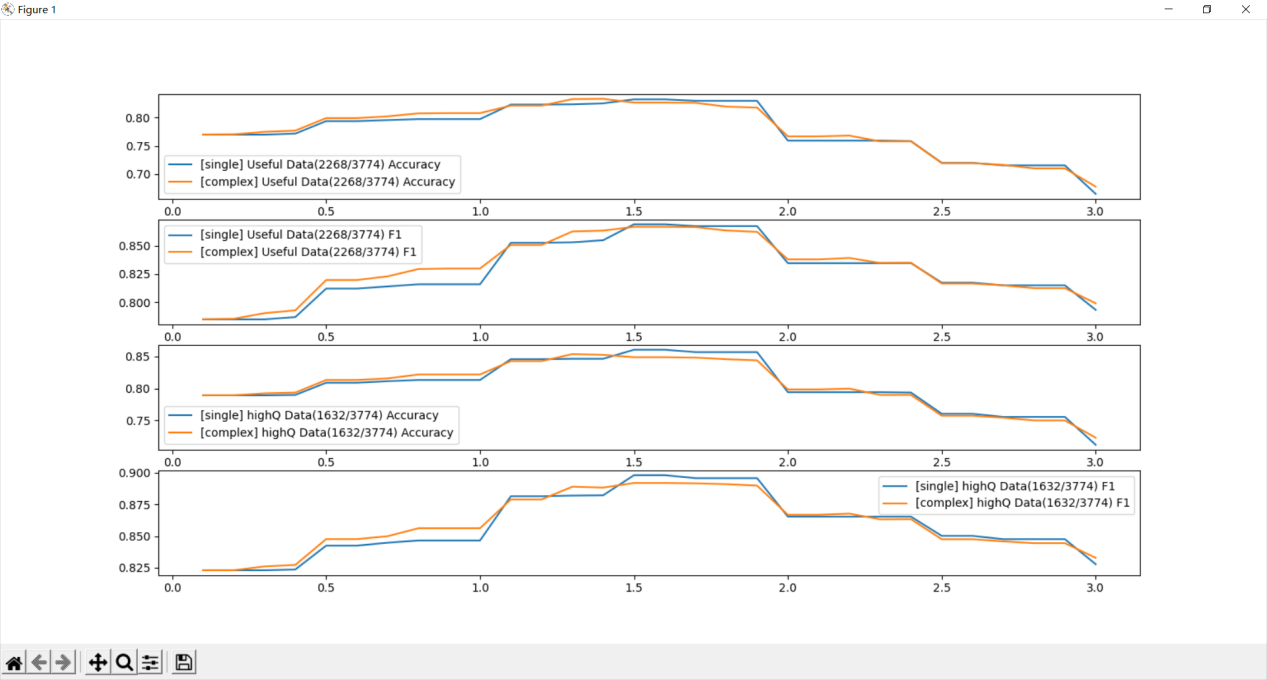
3.1 单阈值

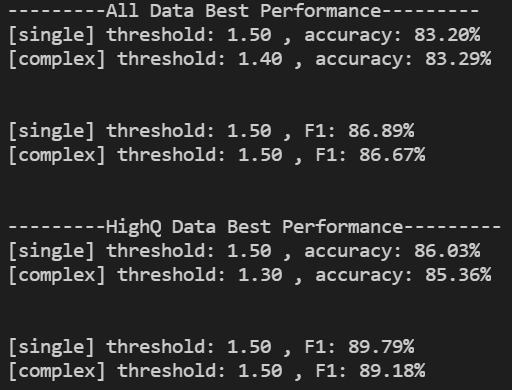
1. 原始拼音 vs 声母、韵母分离



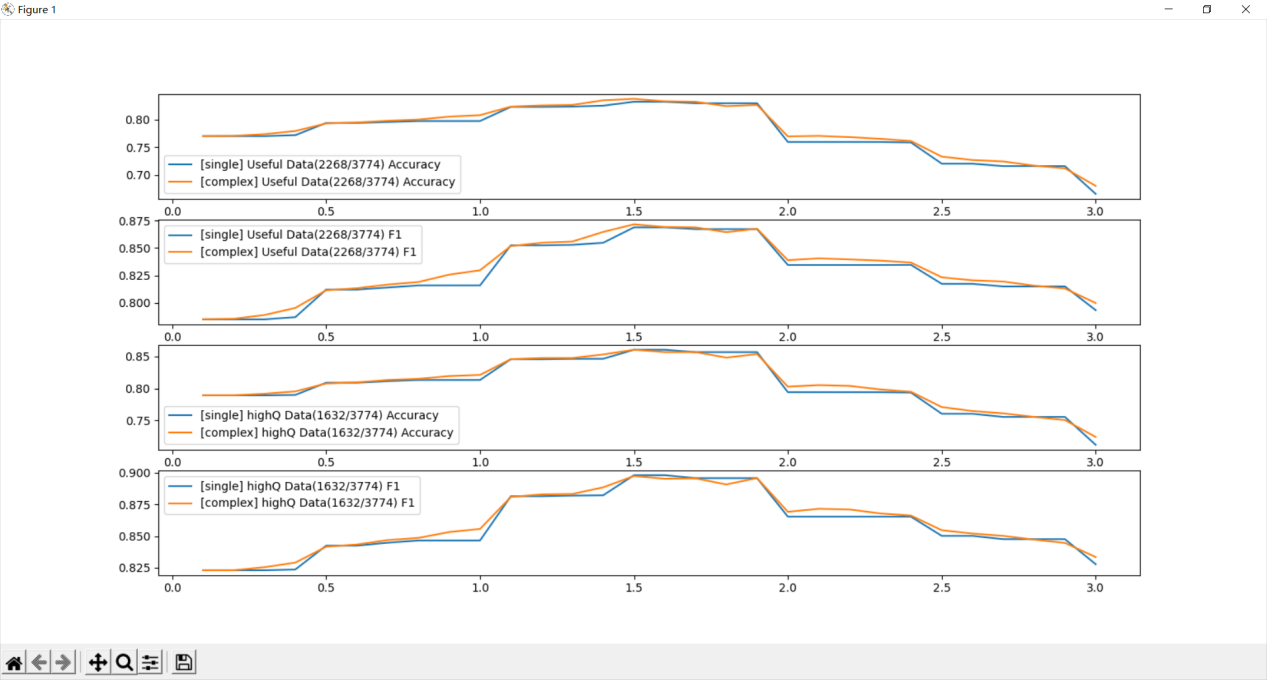


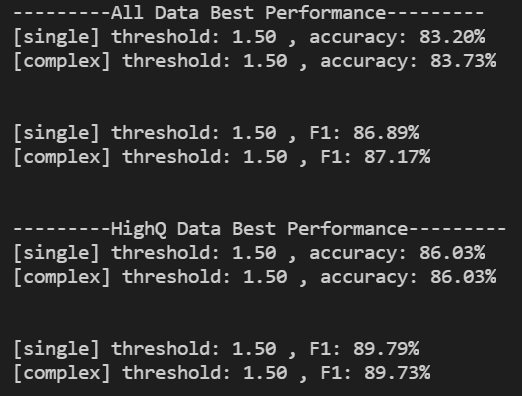
1. 原始拼音 vs 声母、韵母分离（相似拼音按0.5，1.5等比例加权）





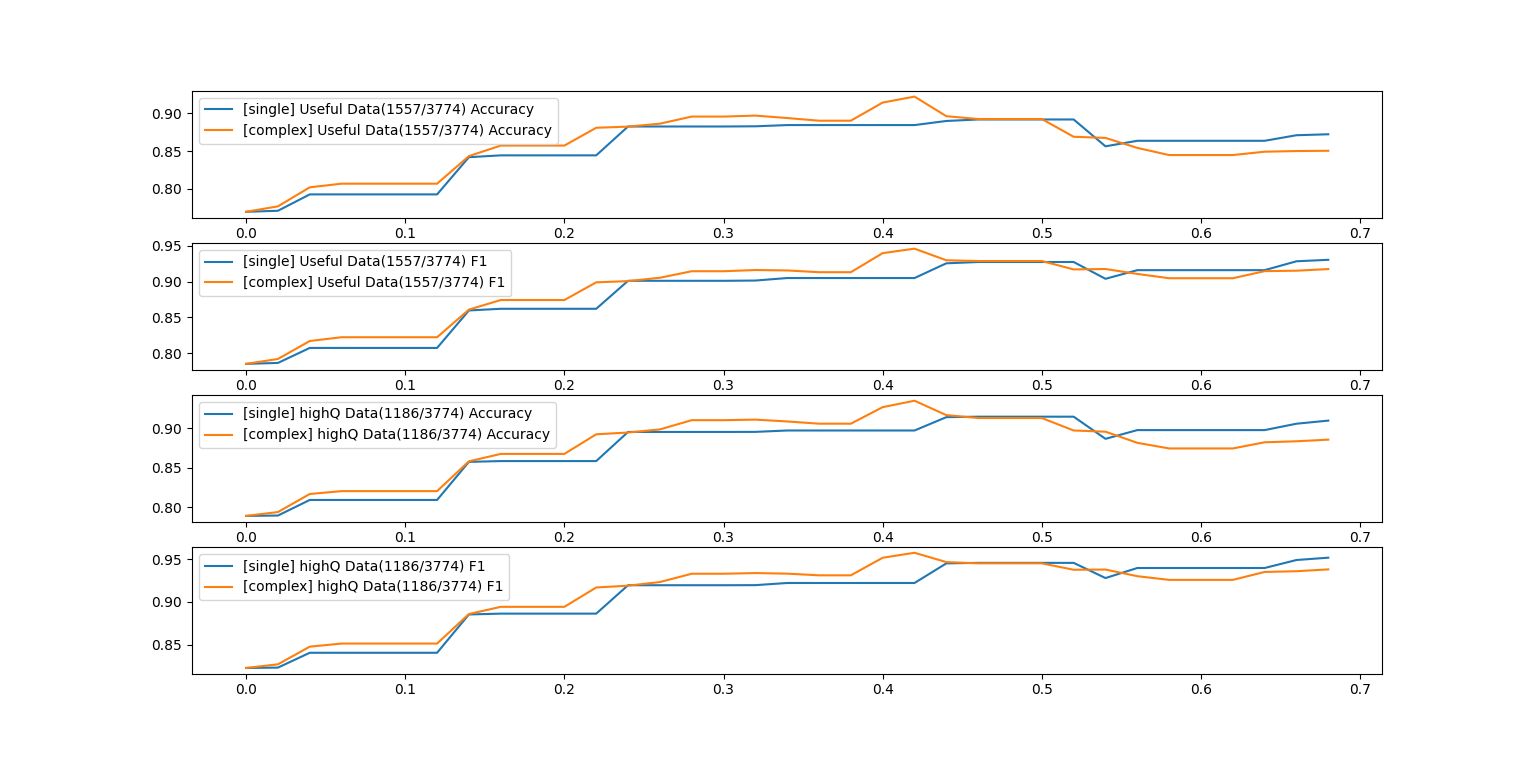
1. 原始拼音 vs 声母、韵母分离（相似拼音按各种比例加权）

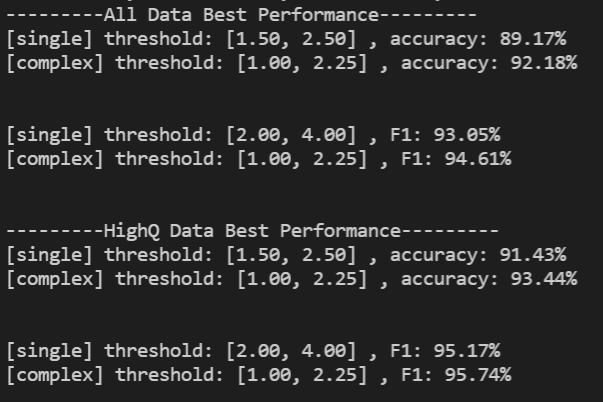




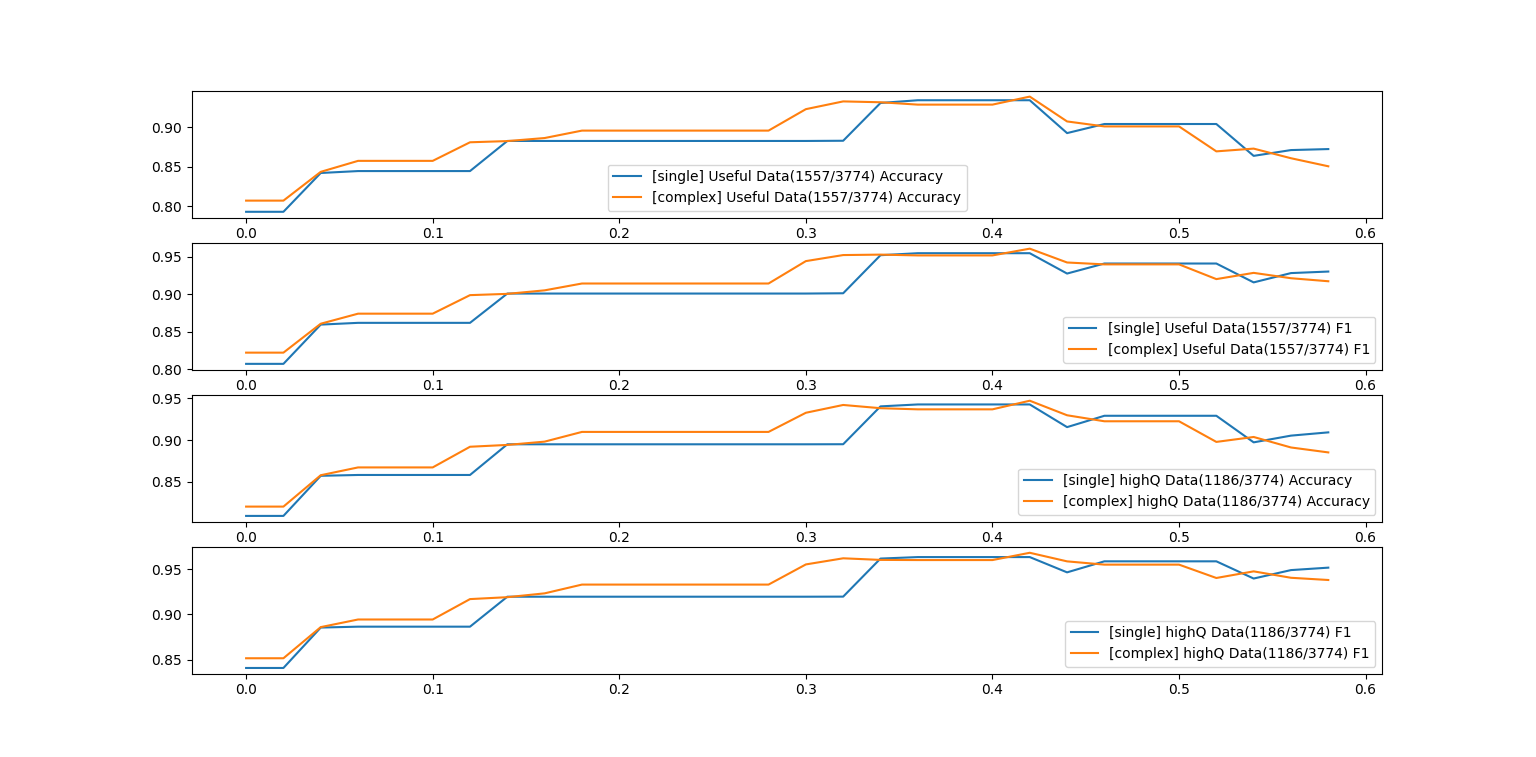
3.2 距离区间

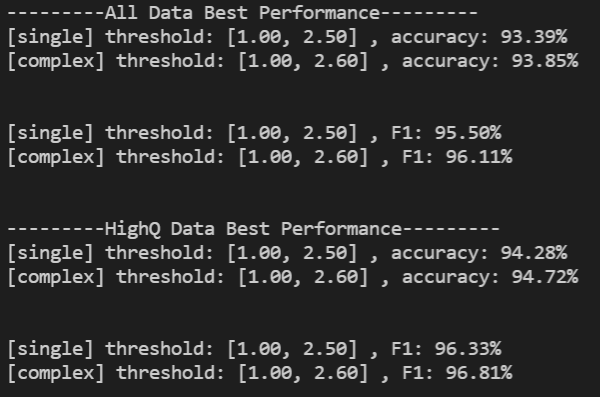
1. 人工判定30%





1. 人工判定40%





**4 模型功能（Demo）**

Page 1

1. 上传单个题目的zip文件，目录格式指定

2. 文件列表显示了服务端保存的各个题目的讯飞识别和认知判定情况

3. 点击某个题目ID，跳转到Page2

Page2

1. 显示图片答案，答案可编辑，但必须以“、”分隔，不包含多余空格

2. 显示音频数量

3. 提供讯飞一键识别按钮，该过程耗时可能较长（几分钟）

4. 用户可编辑人工判定的比例，默认0.3

5. 音频可播放，并显示判定结果列表，认知判定结果有4个状态（正确、错误、无、待定），待定是交给人工判定