1. 研究过程
   1. 数据的特点
      1. 儿童语音的声音普遍口齿较为不清晰
      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. 目前收录的语音有不少无效音频，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音
      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. 在声音清晰且表达正确的音频中，我们发现图片所对应的的答案并不周全，经常出现意思对了但是不符合标准答案的情况，我们认为可以适当的扩展正确答案的集合
      7. 最初的音频是由我们人工进行判定，判定内容包括了：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、音频有效性这7个方面。
   2. 方法的调研与选择
      1. DTW

DTW是最开始我们采用的方法，这是一个单纯的计算音频之间距离的方法，它的工作原理是将两段音频按照一定的、对应的方式分割裁剪，计算每两段对应距离的总和，从而来得到两段音频之间的总距离。

距离的计算方式是可以自己定义的，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. 如果采用这种方式，我们需要儿童标准音频进行比对，儿童标准语音本身不好定义。
2. 这个方法对音频质量本身要求很高，如果是标准普通话正确率尚可，但我们之后试验了儿童语音，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。
   * 1. ASRT

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

我们在使用了ASRT的试用demo后，总结了以下的一些问题：

1. 对标准普通话识别率尚可，儿童语音识别率很低。
2. 模型更加偏向长句的识别，联系上下文可获得更好的识别率，而我们的儿童语音基本都是词语词组，不适用。
3. 对噪音、杂言、环境音的抗干扰性一般，这部分或许可以通过预处理进行规避
4. 如果要用训练自己的模型来投入使用，我们缺少大量的数据作为支撑（若之后全国数据足够，可以考虑训练自己的针对儿童语音的识别模型，也是我们下一步考虑更进一步的方向），而市面上标准成人普通话公开数据集是不适用的。

2、下面是目前采用的方法：

* + 1. 科大讯飞（使用中）

现在市面上很难找到专门面向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，所以我们试用了讯飞的接口来识别测试。

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语，其实测试得到的结果并不十分理想。后来我们发现，许多错误的识别结果虽然中文字本身识别错误，但是若拆成拼音来看，确实相差无几，所以我们想到了一种滑动窗口的思路来解决这个问题。

* + 1. 拼音滑动窗口（使用中）

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，其对应的科大讯飞识别结果我们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端拼音之间的最小距离，作为标准距离。（也许这里可以做一个滑窗的演示小动画？就几页ppt顺序放然后滑窗那种）

（重点）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。

在这个基础上，我们给这个距离设置了一个阈值（可以自行设置、更改），高于这个阈值我们就认为这个音频是错误的，反之正确。若出现多组正确答案，则取其最小值作为标准距离。

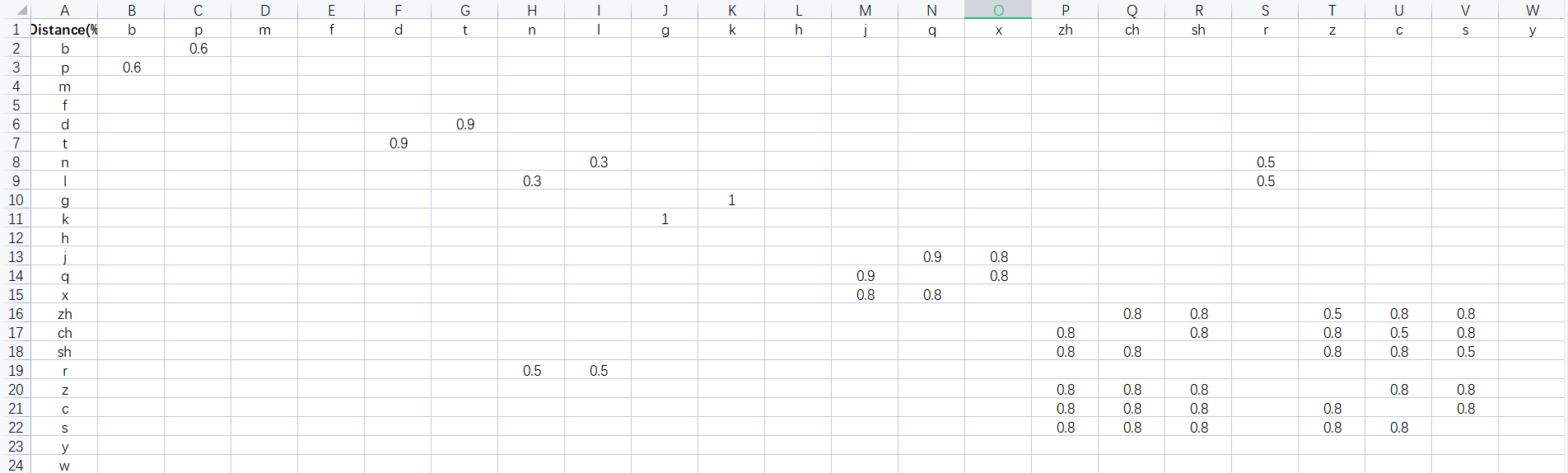
在得到距离后，如果需要部分数据进行人工判断，因此，我们把单个阈值改为距离区间，区间内的部分最难判断，交给人工。我们把区间外距离最近，即最靠近正确答案的部分判断为正确，距离最远，即最远离正确答案的部分判断为错误，剩下的部分（可调节）交给人工，这部分在后面的模型功能会详细介绍。

（重点）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。

* + 1. 距离计算方法（使用中）

最初的距离计算方式是最朴素的编辑距离，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。我们将单字拼音拆分成了声母、韵母两部分，并对两者都做了相应的距离对照表，从而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、规定，所以目前的是否合适，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

