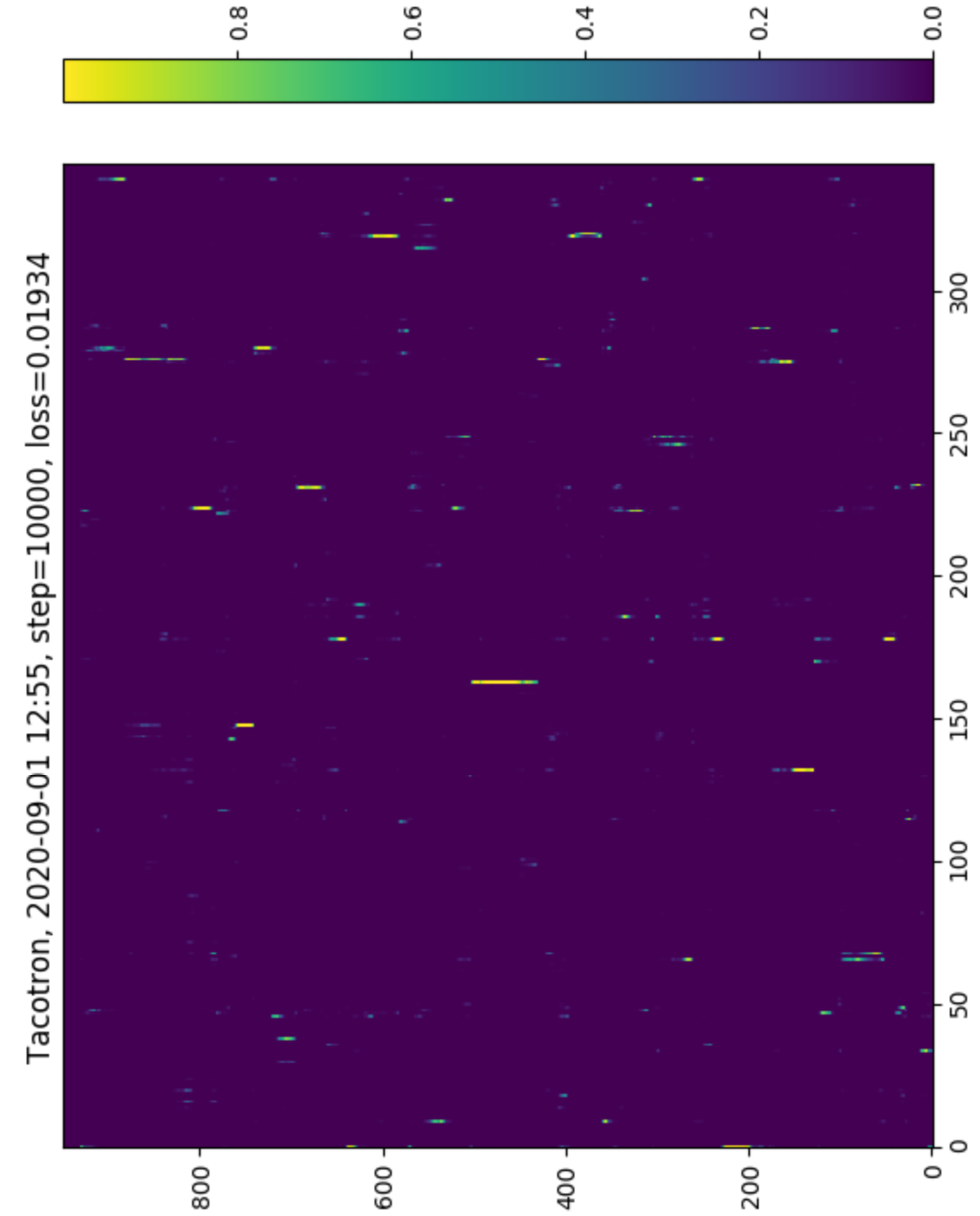
Tile：PPG Cross TTS

Abstract：略

1. INTRODUCTION
2. RELATEDWORK
3. BASELINE APPROACH
   1. PPG

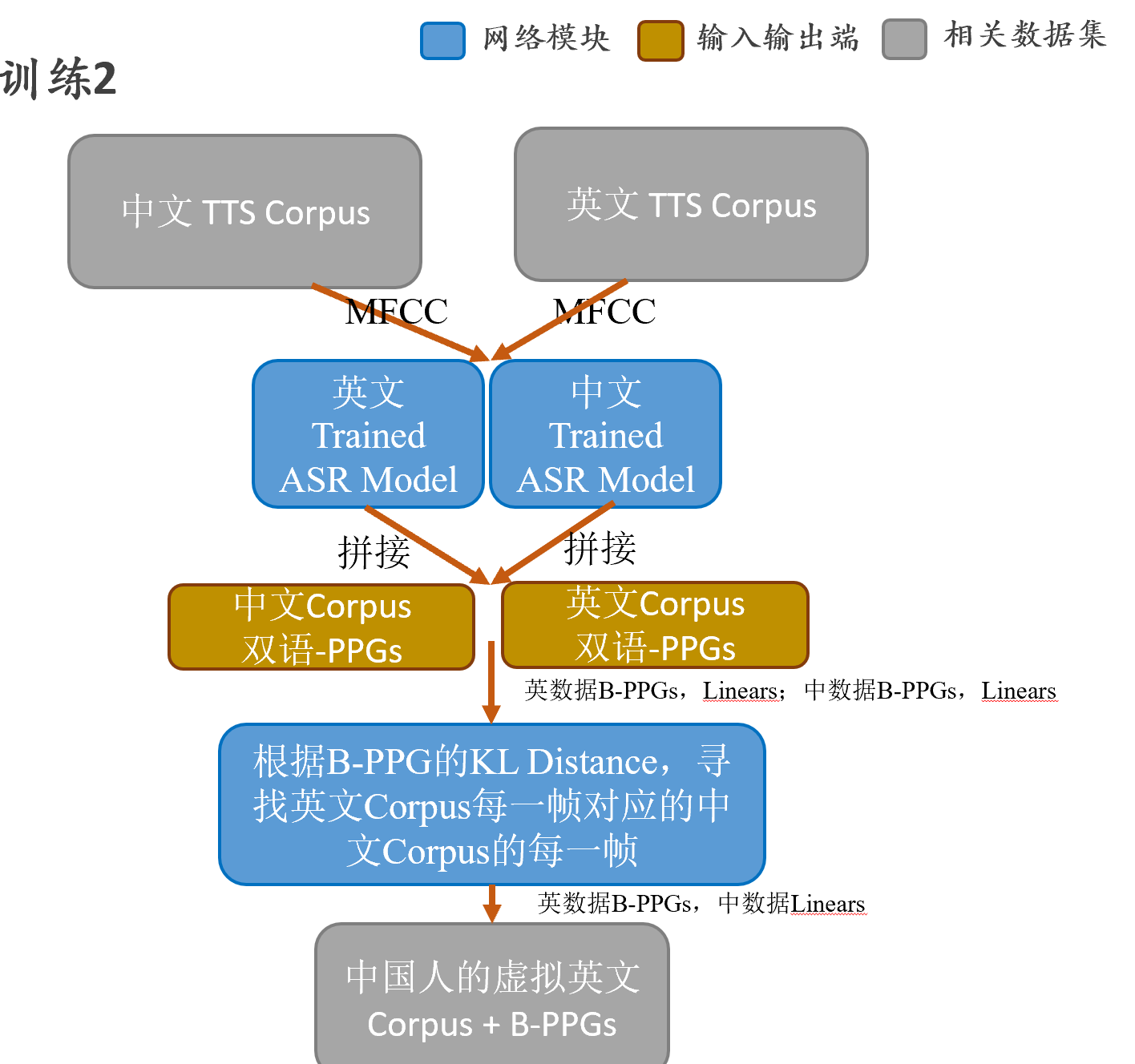


1. PROPOSED APPROACH
   1. Approach: FindA

算法的定义为：

我们采用了寻找维度空间里距离最为相近的中英文语料的PPG作为用来一一替换的映射对的寻找依据。

为了从一句英国人说英文的句子转变为中国人说英文的句子，为每一帧的英文语料PPG寻找对应最合适的一帧中文语料PPG，并用对应的中文的PPG（Spec）来替换原英文的PPG（Spec）



* + 1. 距离选取为欧氏距离

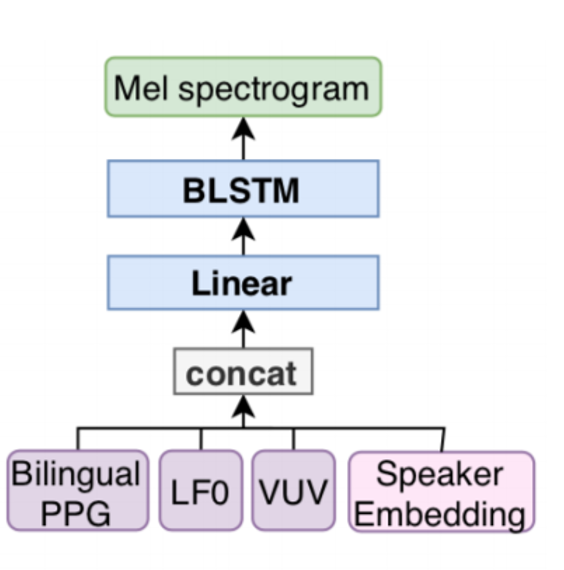
欧氏距离的定义：略

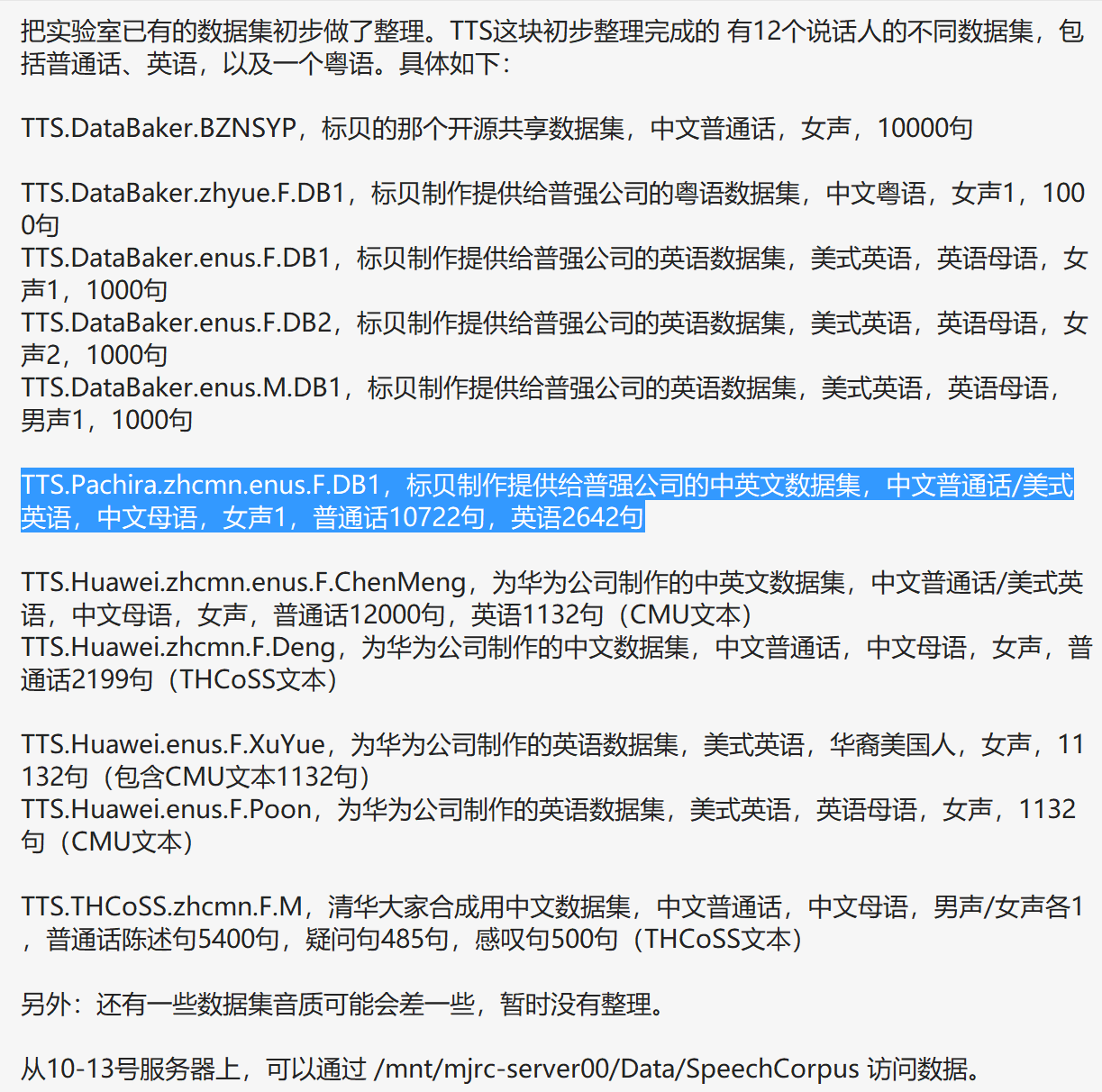
算法具体流程是：

1. 首先，分别读取训练数据集中的中英文语料的PPG，并将其拆分成相互独立的帧级别的PPG，将二者混合在同一个列表里面
2. 然后，为每一帧的英文PPG寻找欧式距离最近的中文PPG，由于训练数据集过大和时间复杂度过高，采用了KMeans的聚类方法，先对所有中英文帧的PPG进行聚类，然后再在每一个类内进行距离寻优，为每一帧的英文PPG寻找到类内最近的中文PPG，并将其认为是整个维度空间里距离最近的中文PPG
3. 类内寻找最近距离PPG时，类内建立基于欧氏距离的k-d tree数据结构，用于加快查询的时间复杂度
4. 经过3，4步骤，为每一帧英文PPG找到了全局中文PPG中最近的一帧PPG，记录为map矩阵
5. 最后，根据map矩阵，英文PPG为map矩阵查询索引，逐帧的替换此句话中的英文Spec为中文Spec
   * 1. 距离选取为对称KL散度

无法使用聚类和k-d tree优化，但是可以：

1. 写成矩阵形式并行计算
2. PPG只考虑维度中前三大的维度的数对对称KL散度距离的影响，剩下的近似为0
   1. Approach: FindB-Multi-Speaker-CBHG





1. EXPERIMENTS

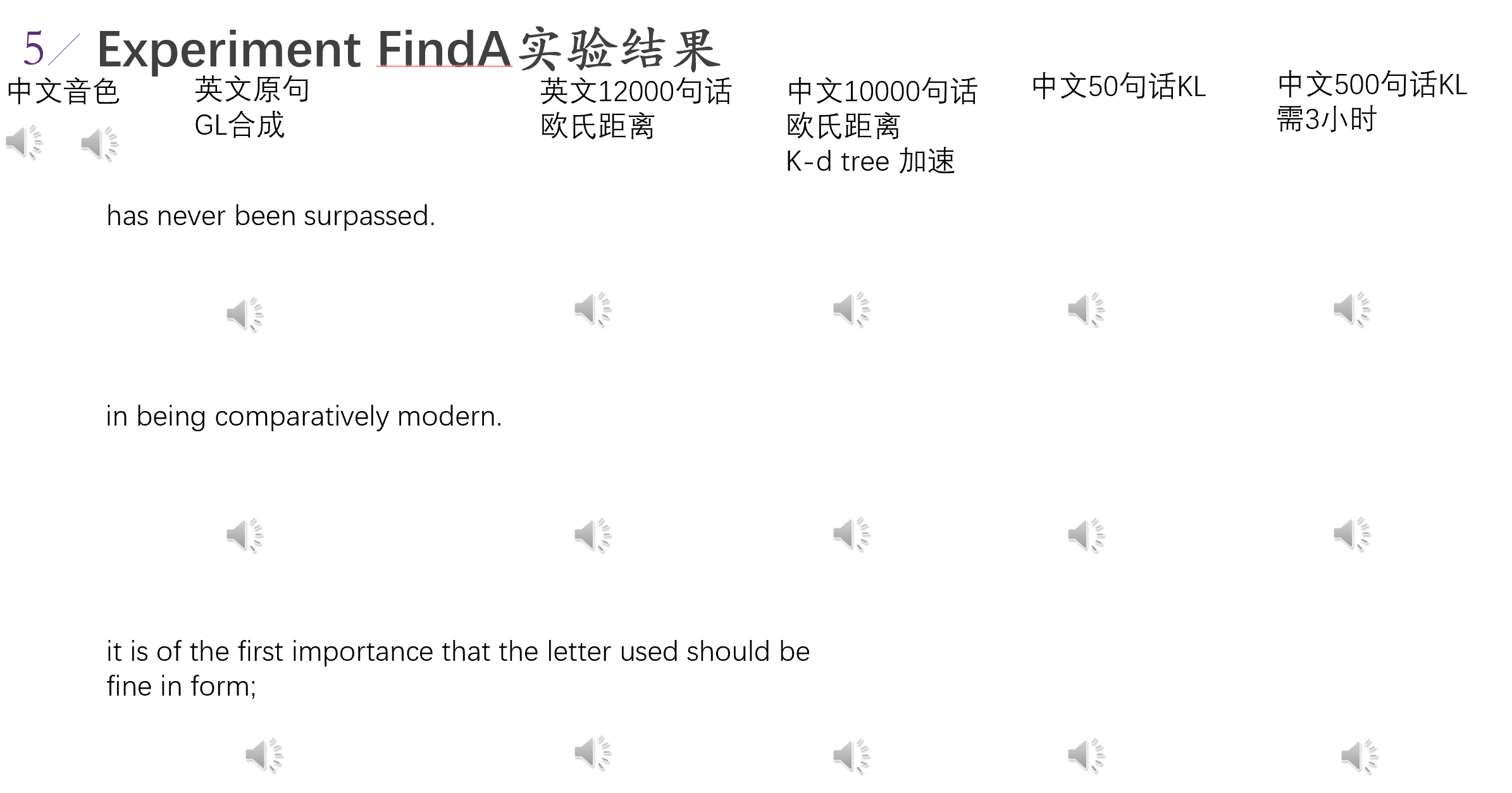
5.1. Experimental setup

ASR部分：英文ASR，使用5ms帧移，Librispeech训练出来，使用卢辉预训练正确率83%的ckpt

PPG提取：

* 英文语料：LJSpeech 13100句，用英文ASR提取PPG，并提取Spec
* DataBaker\_CN 10000句，用英文ASR提取PPG，并提取Spec

具体超参数见PPT：



1. 实验现象是不连续，呲呲啦啦，本质上是：每一帧挑选ppg是单独割裂开来挑的（拼出来的），拼起来的时候，spec一定会呲呲啦啦；所以目前重点是在从呲呲啦啦的声音到平滑的声音，这样做才能出demo，经典问题
2. 处理specs （不考虑）
3. 从呲呲啦啦的声音到平滑的声音方法一：**从时域上取一小段波形\***
   1. **呲呲啦啦信号波形，时域上加窗，信号角度平滑**
4. 拼帧，距离是|前后几帧，一个大的PPG-Triphone
5. 问题建模：中文的原始语料使得构造的PPG英文新句子，英文原始，关注输入：替换的PPG-Spc，目标是，“平滑，正确”的波形-TTS， 中文过度特性，英文正确发音

PPG -》 X feature represent

X， Spec

**验证：现在数据，实验结果：PPG与Spec发音相似的对应，是否是杂乱无章无序的，**需要重新排序/映射关系，保证映射关系一致。是否优缺点：不连续-》紧邻不代表Y近邻