作业

王成仲 202218001027011

中国科学院声学研究所

中国科学院大学

1. 摘要

本作业基于离散标签和MFCC，分别实现了48个单词的孤立词语音识别系统。对于离散标签输入，采用词向量优化方法；对于MFCC输入，采用同态滤波+Boosting技术；测试集精度分别达到91.95%和87.36%，两个系统在未知数据上达到了77.35%的一致率。这些表明，在现实中使用两者进行48词孤立词语音识别是基本可行的。

1. 基线

2.1 离散标签输入基线

离散标签输入基线系统如图1所示。输入每个标签对应的独热表示，进入CNN网络执行降采样，再通过LSTM层和线性层得到每个帧降采样后对应字母和空标签的log概率输出。

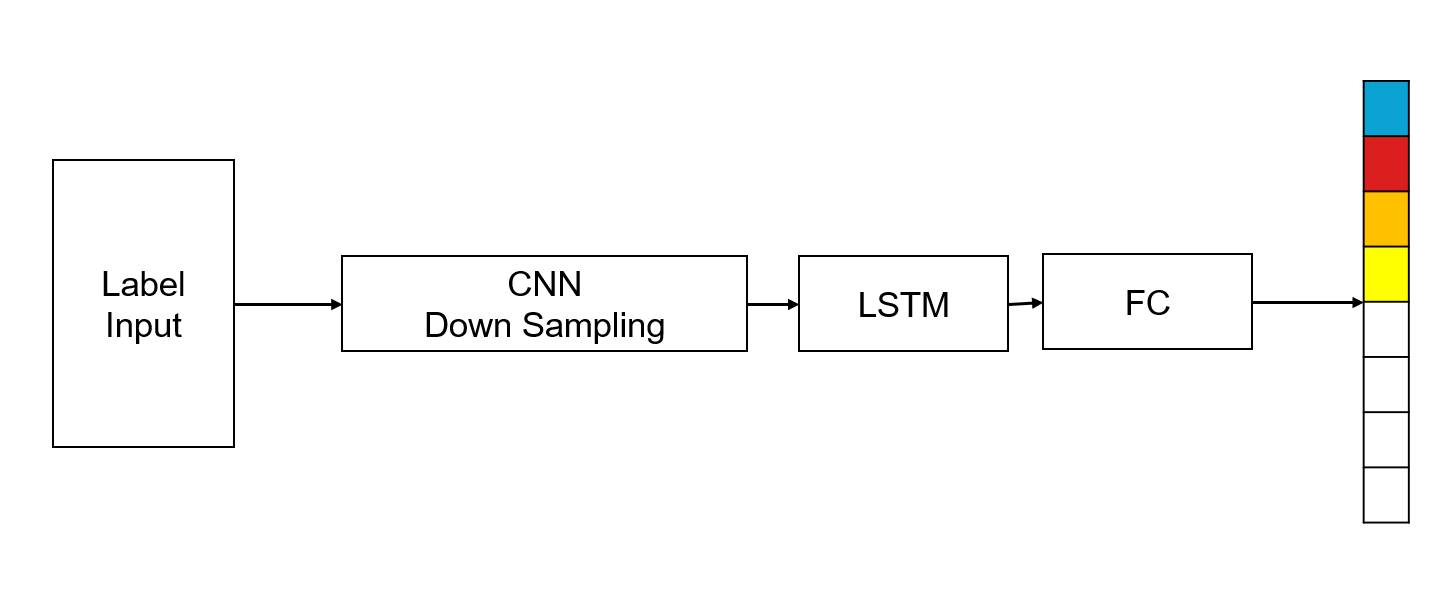


图1 离散标签输入基线

卷积降采样层使用二维卷积，以减少LSTM层的输入负担，如图2所示。

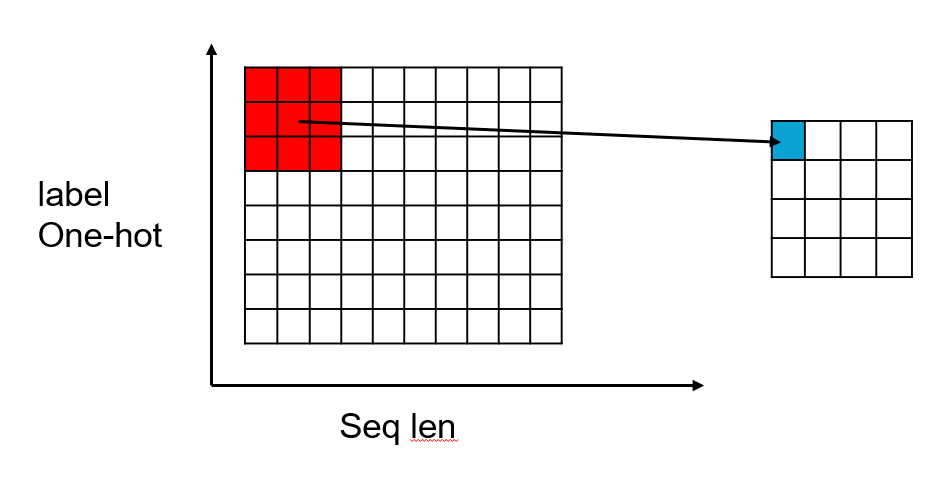


图2 卷积层

各Batch输入的序列长度不同。记录降采样后的维度，在输入LSTM之前解padding，在输出之后Padding，并沿着LSTM输出的特征维度作用一个线性层，提取最终的26维输出，每一维表示输出对应25个字母和空标签的log概率（注意到，48个单词没有出现z的，所以输出仅含25个字母），如图3所示。

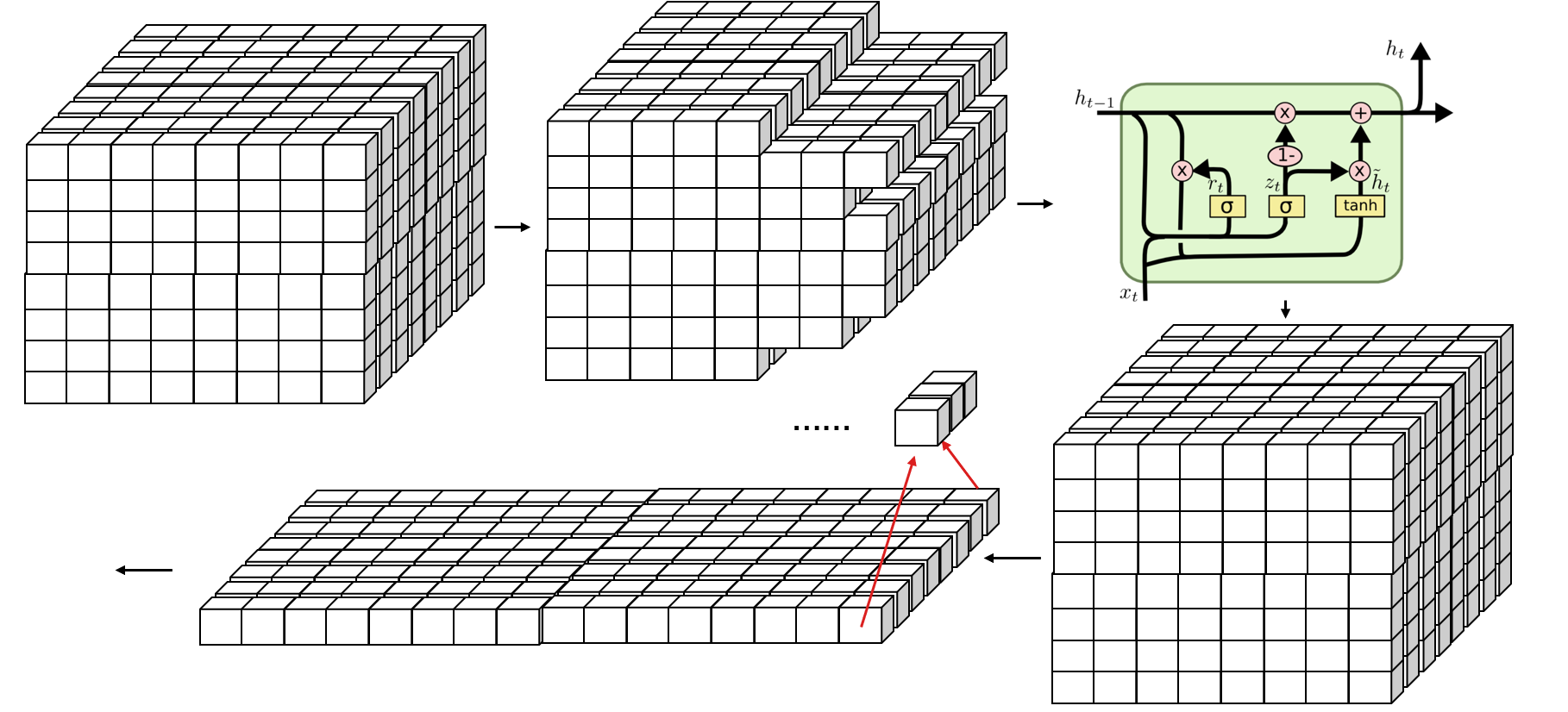


图3 LSTM padding & packing和线性层

使用CTC-Loss作为优化准则。

2.2 MFCC输入基线

基线模型结构与2.1相同。输入MFCC特征时，将一阶差分和二阶差分连接输入，如图4所示。

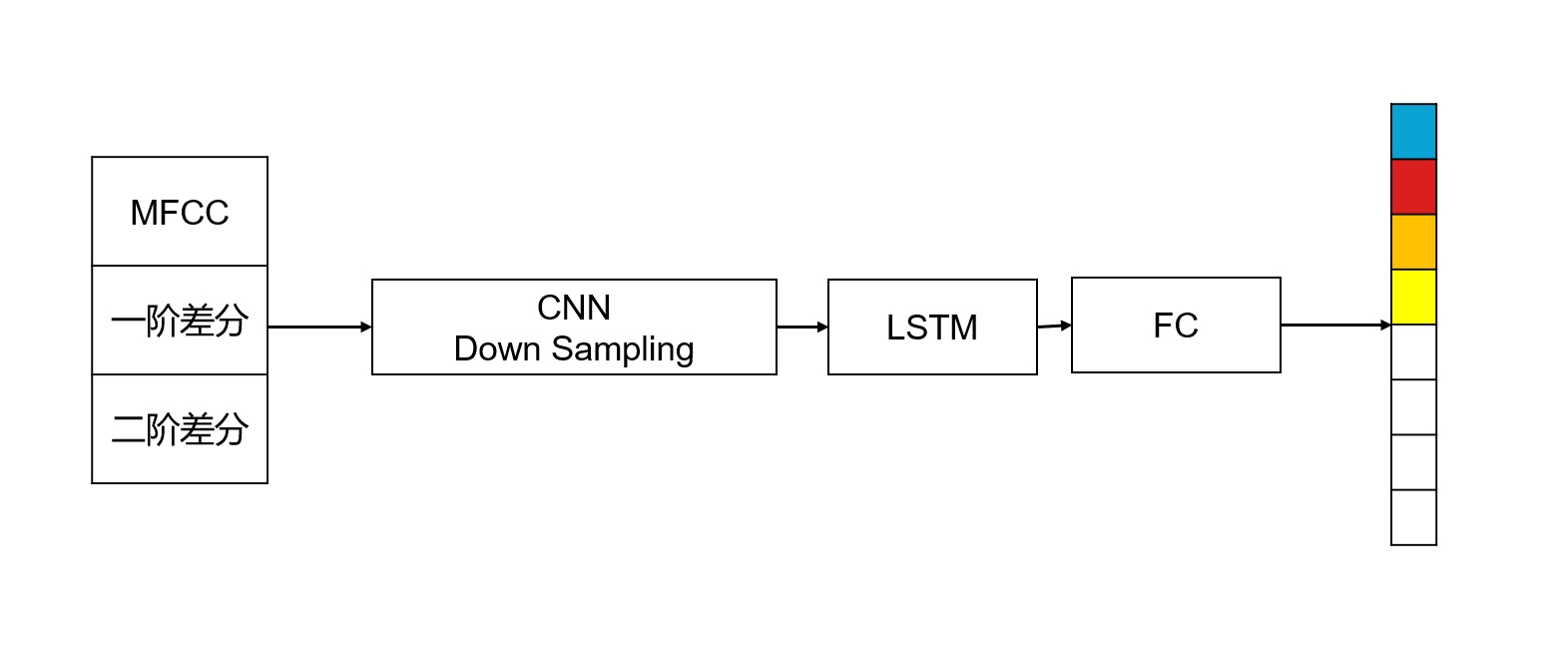


图4 MFCC输入基线

2.3 解码方法

系统输出降采样后每一帧的字母估计，故需要解码方能得到48个词的输出。

2.3.1 贪婪解码

贪婪解码是CTC-loss解码常用的方法。在多数情况下，这种方法能给出良好效果。贪婪解码的基本过程如图5所示。在每个时间步，选择具有最大概率的输出标签，而后对相邻相同的标签进行合并。这种解码方法不考虑输出序列的整体概率，只关注每个时间步的最大概率。但是，贪婪解码忽略了一个输出可能对应多个结果的问题。例如， [a, a, *ϵ*],  [a, a, *a*] 与 [b, b, b] 相比，前两者对应同一个解码结果，可能概率总和更大，但使用贪婪解码时只看当前时间步的解码结果，可能产生[b,b,b]的错误输出。



图5 贪婪解码

2.3.2 束解码

束解码可以解决贪婪解码一个输出对应多个对齐结果时解码错误的问题。它的基本流程如图6所示。确定一个束大小N，在每个时间步按照概率同时选定历史上前N条可能到达的路径，而后舍弃剩余的路径。这种解码方法在解码精度和计算复杂度之间取得了折衷。

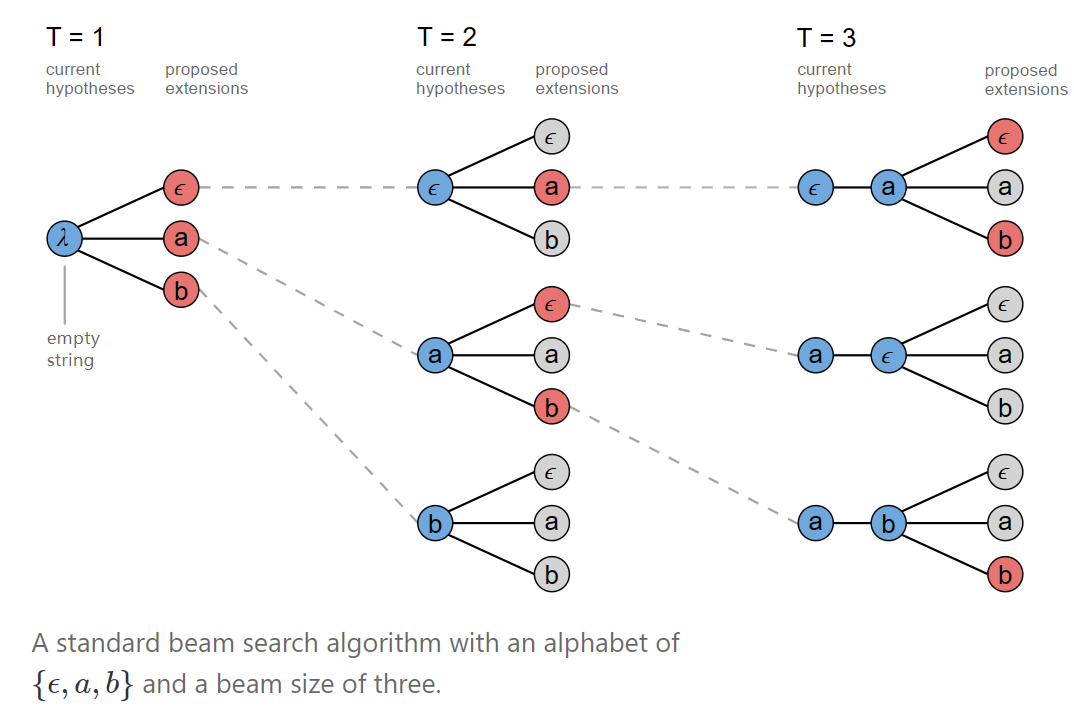


图6 束解码

2.3.3 针对此作业的解码方法

该作业已知目标只有48种可能。为此，可以直接对每个单词计算CTC-loss，然后取最小损失对应的单词作为最终输出结果，如图7所示。

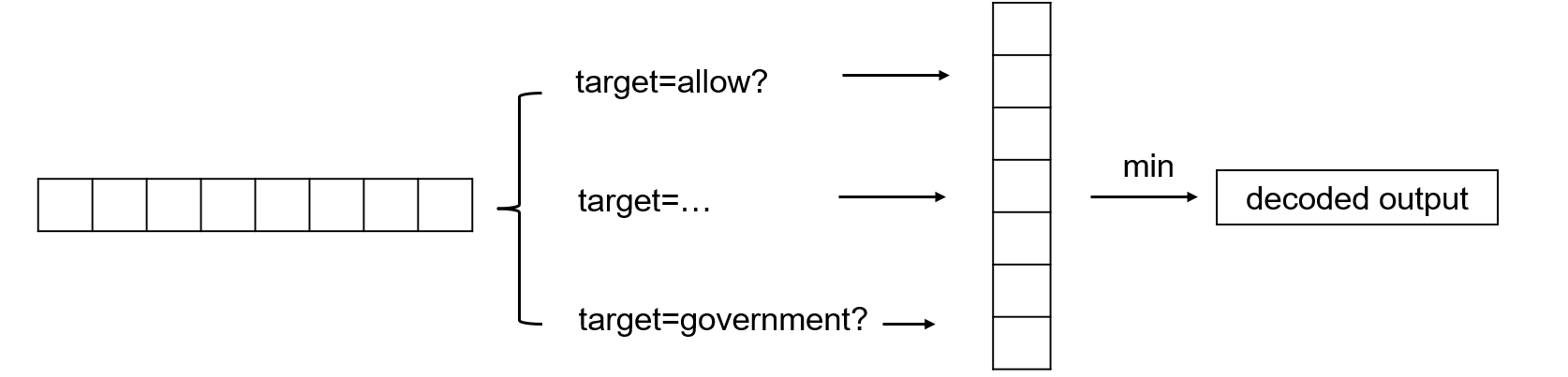


图7 针对此任务的解码方法

2.4 概率求解

为了求解解码出单词对应的置信概率，有以下三种思路：

* 使用贪婪解码的匹配过程给出置信度
* 使用CTC-Loss直接计算置信度
* 将48个单词依次代入模型输出计算CTC-loss

第一种方法较为简单，但CTC-Loss生成输出序列时的概率信息被忽略，因此是错误的求解办法；第二种方法十分复杂，但没有考虑仅仅有48个单词，因此也不合适；第三种方法简单且正确。它将48个单词计算得到的损失取指，反映当前输出对应某单词的概率大小，而后对这些概率进行归一化。

1. 提升

虽然基线方法可以达到作业要求，但是它存在以下三点问题：

* 训练不稳。不同次训练的测试集精度差别太大，可达10%以上；
* MFCC模型太大。在使用MFCC模型进行训练时，需采用300~500个LSTM Hidden Unit，使用多层LSTM层才能在测试集上达到目标效果，这造成15~30MB的模型大小。
* 泛化精度太差。使用测试集达到80%的模型拿到未知数据上对照，两基线系统的相合度远低于50%。我们认为是MFCC系统的严重过拟合导致差的泛化能力。

针对基线模型的问题，我们改变数据处理方式、改变模型结构和训练方法，从稳定性、精度和参数量等方面对基线系统进行提升。实验表明，这些提升手段带来了显著的效果。接下来的部分，将分别阐述针对离散标签输入和MFCC输入的提升技术。

3.1 针对离散标签基线的提升

3.1.1 词向量

反思离散标签基线（图1）中的问题。我们认为，独热输入对CNN来说是不合适的。这体现在以下两点：

* 输入向量如果是独热的（如图8所示），CNN对相邻的方块编码，但是相邻的方块并没有关联信息，使用CNN降采样没有物理意义。这与图片等场景下CNN处理的数据是不相同的。
* 256维的独热向量是按照文件的顺序随便读入的，标签序号是随机的排列，有可能给CNN神经网络带来困扰。

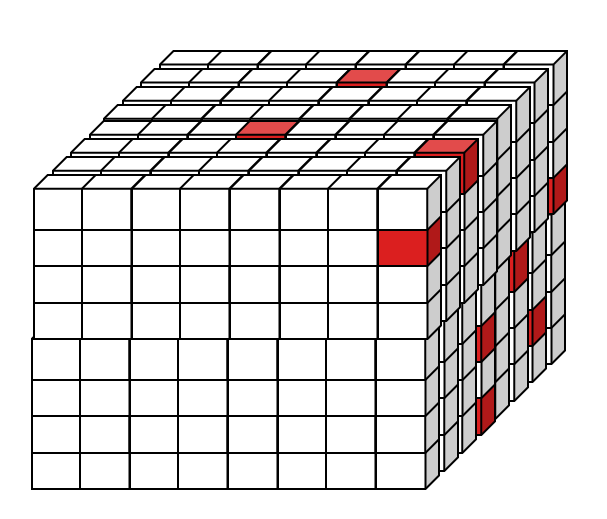


图8 独热的无意义编码向量给CNN理解数据带来困难

为此，我们采用词向量进行编码。词向量的原理如图9所示。训练目标是神经网络中的隐层表示参数，这些参数可以编码多个句子输入中的词相关信息，使得具有相似意义的特征在转换后的向量空间中的距离较近，这正是CNN想要的。

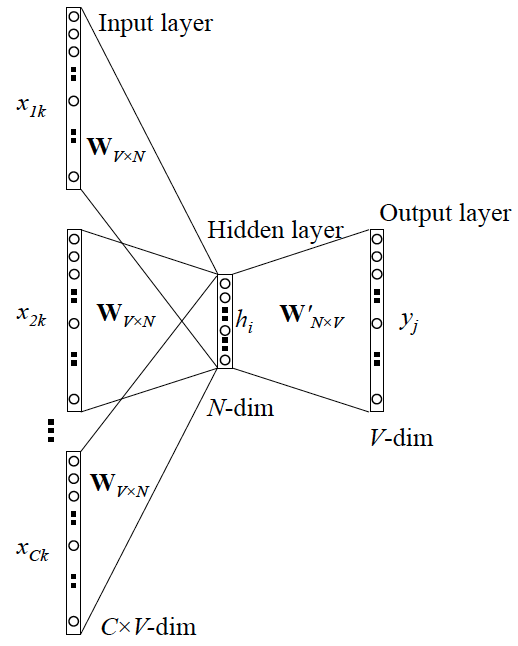


图9 词向量编码的原理。对作业来说，输入形如[[“GG”,“PP”,“BB”,“QQ”,…],[“AH”,”MM”,”SS”,”JJ”,…]…]，

而后输出每个标签的256维编码表示。

加入词向量后的结构如图10所示。

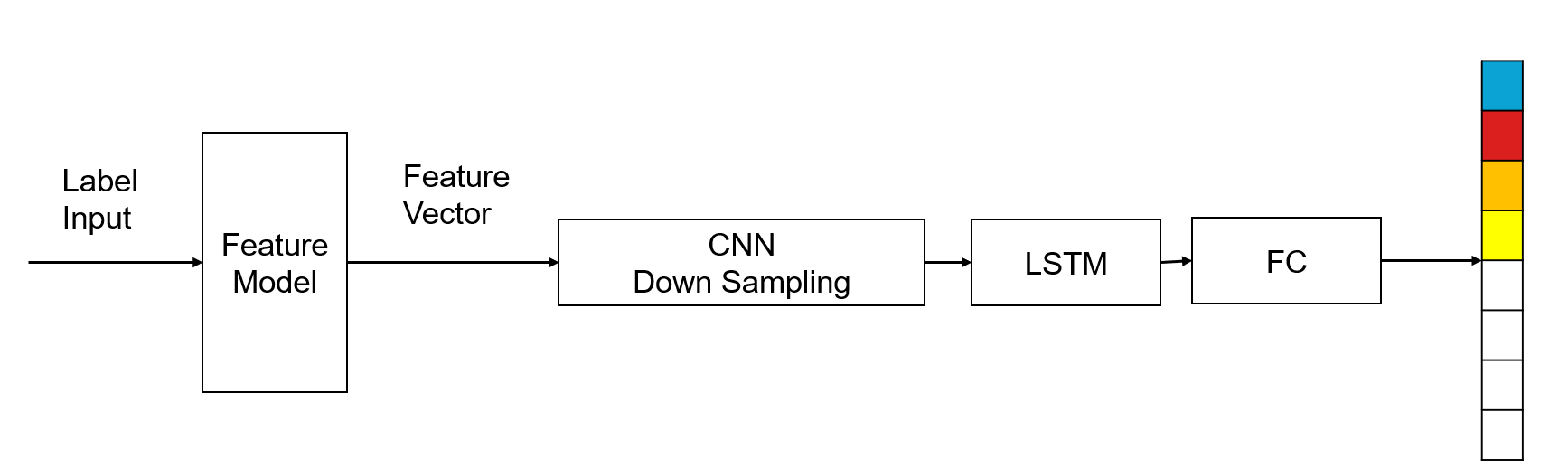


图10 加入词向量后的结构

3.2 对MFCC基线的提升

3.2.1 同态滤波

反思MFCC基线系统输出的问题。MFCC的输入对语音的特征信息并不敏感，对元音类似、爆破音有些许不同的语音估计往往错误，如图11所示。

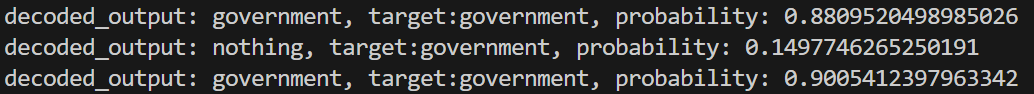


图11 MFCC语音估计错误的情况

为此，我们引入同态滤波。它主要用于提取语音信号的包络信息，将语音信号分为激励信号和声道响应两部分。激励信号主要包含语音的激励信息，而声道响应包含语音的共振峰信息。

我们对信号进行对数运算，将乘法运算转化为加法运算，然后在频域上进行滤波，将原信号MFCC、一二阶差分、e(n)和h(n)的MFCC分别作为五个通道输入卷积层，其流程如图12所示。

实际上，x(n)可以通过如下方式恢复。

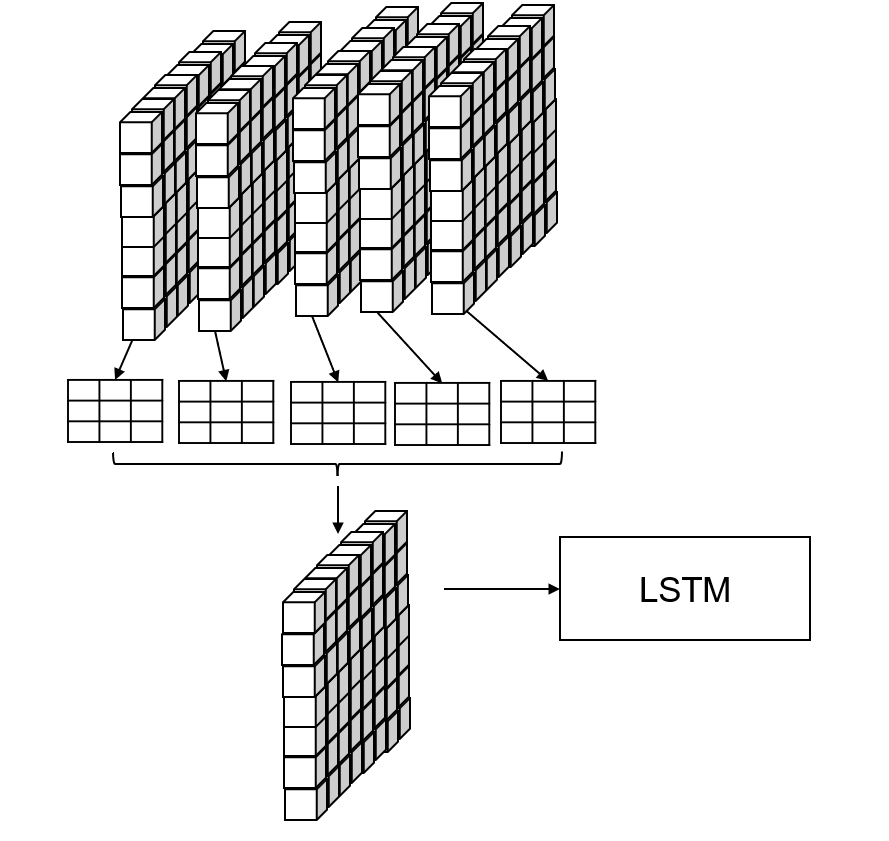


图12 将5个输入分量作为5个分量输入卷积层

3.2.2 Boosting

在3.2.1的提升后，我们发现，验证集的精度仍然不够高，且相邻epoch之间的精度差距可达10%以上，如图13所示。为此，我们将神经网络与传统的Boost技术结合，提出了一种并行输出分批训练的模型结构。实验证明，这种结构以约3倍的模型大小代价，实现了更加鲁棒的训练过程和明显更强的泛化能力。





图13 不同Epoch间的测试集精度差异可达10%以上。

上：Epoch74 下：Epoch75

Boosting的结构如图14所示。设置三个并行的网络，这三个网络依次变简单，以对更少的数据量进行分类。在训练时，当训练到一定epoch数时，检查对48个词的激活情况。如果某个词的激活概率全部大于0.75，则不加入这个词；如果某个词的激活概率小于0.75但与第二个激活词的概率之差大于0.4，也不加入这个词；仅当当前词的激活概率小于0.75并且与第二个激活词之差大于0.4时，加入这个词。这样分别对48个词依次运行一遍，如果加入的词数大于阈值（设为7），就开始对神经网络的下一部分进行训练。(控制逻辑有很多，需要精细控制)

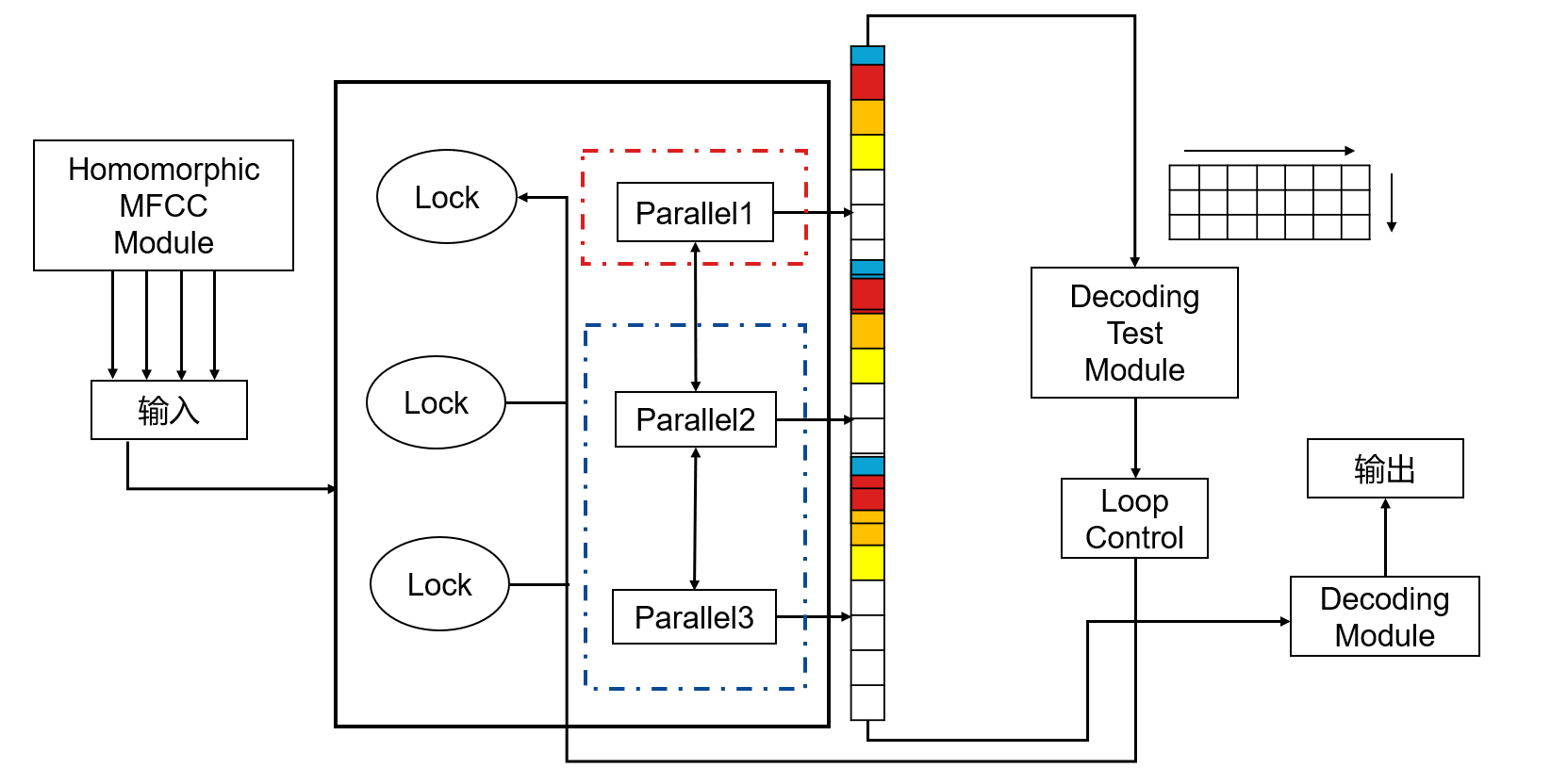


图14 Boosting的结构

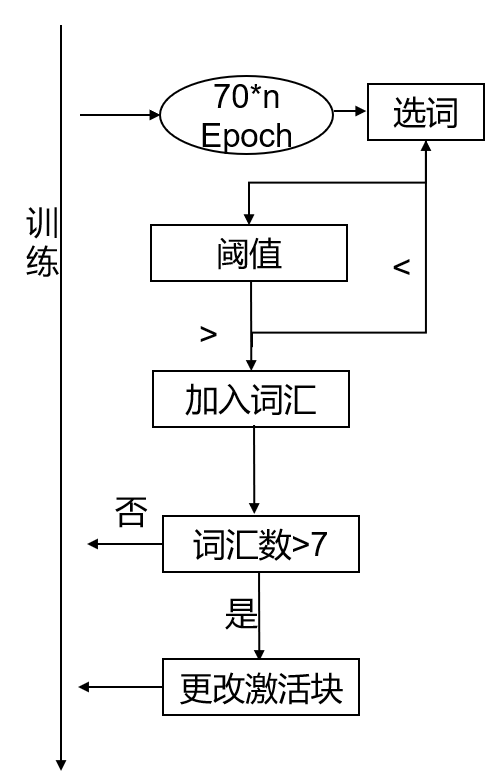


图15 Boosting的训练判断流程

Boosting后得到模型的解码过程如图16所示。

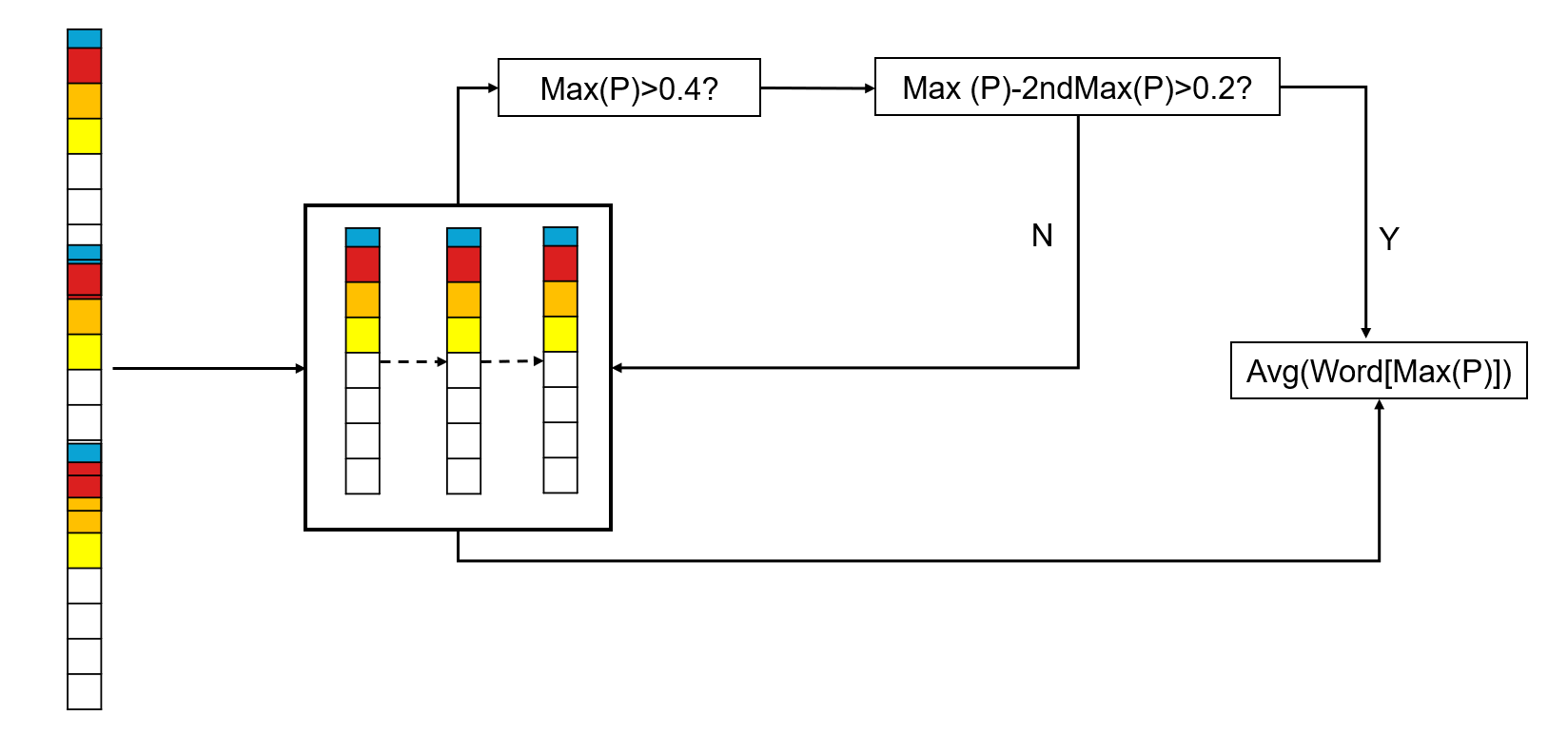


图16 Boosting模型的解码

1. 实验结果

4.1 概览

从图17中可见，基线系统和提升系统在测试集上都达到了80%以上的精度，但是提升系统的精度和稳定性都更好。

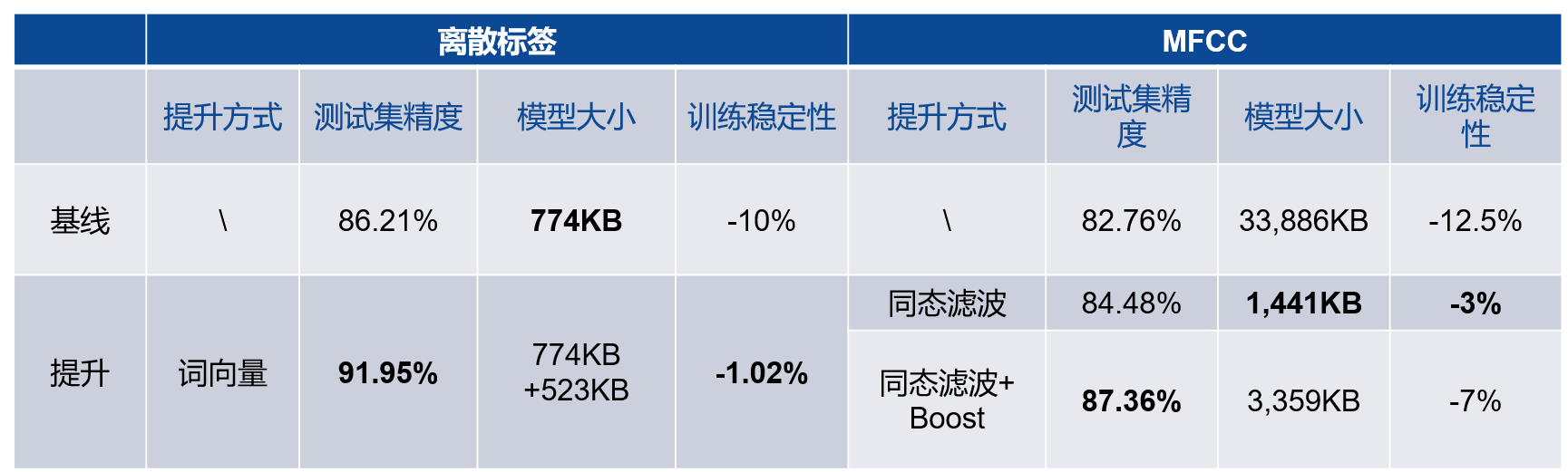


图17 概览

从对未知数据的泛化能力看，提升系统远强于基线系统，如图18.

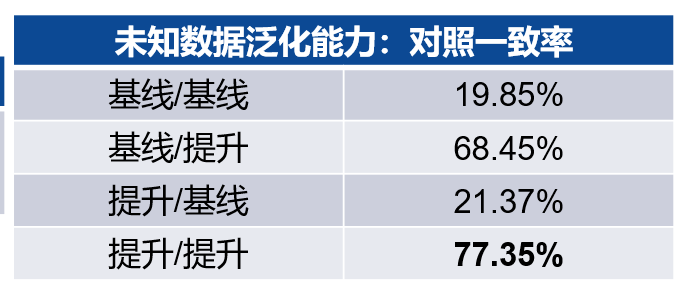
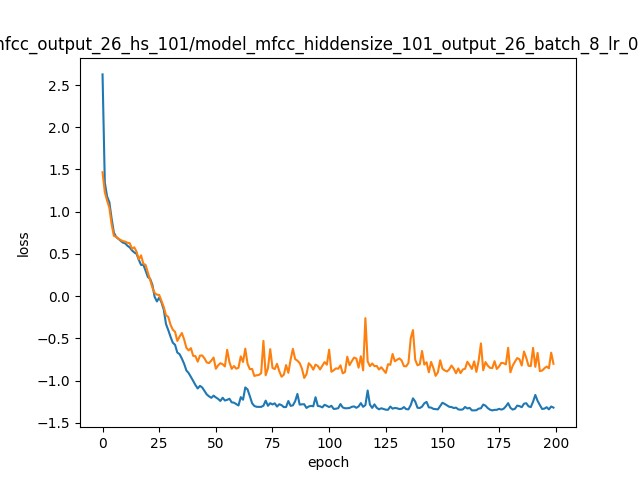
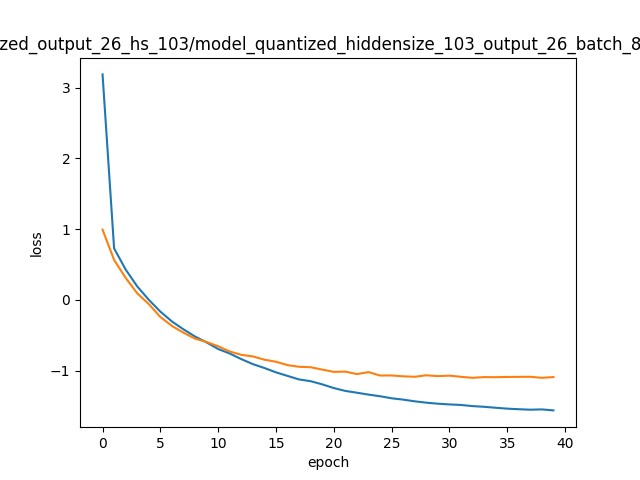


图18 未知数据的泛化能力对照

4.2 训练-测试loss曲线

如图19所示。



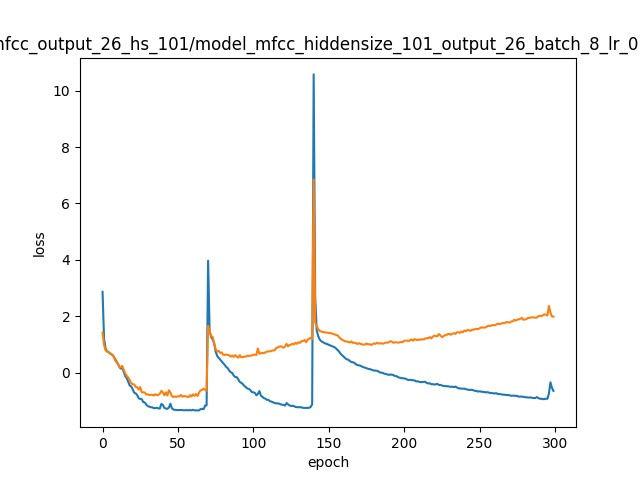


图19 系统的损失函数曲线对照。

左上：词向量提升的离散标签输入时的损失函数曲线

右上：同态滤波提升的MFCC输入时的损失函数曲线

下：使用Boosting和同态滤波提升的损失函数曲线

可以看出，使用词向量提升和同态滤波提升都没有问题；但是使用Boosting提升时，明显出现了第二、第三个块泛化能力差的问题。我们认为，虽然在本测试集中这种方法没有突出优势（仅提高4百分点的测试集精度且泛化能力没有明显改善），但在大规模数据集的训练和实际应用中，根据训练情况选用不同的系统以并行或串行的方式进行判断、并设置合适的控制逻辑进行级联，对提高模型对数据的针对性、提高测试精度、提高模型的领域适应性和泛化能力都有好处。

1. 总结

本文首先基于CNN-LSTM搭建了一个孤立词识别模型。对于离散标签和MFCC连续特征，测试集识别精度分别达到86.21%和81.76%；而后，对于离散标签采用词向量进行提升，测试集识别精度达到91.95%；对MFCC连续特征采用同态滤波和Boosting进行提升，测试集识别精度达到87.36%。在未知数据上，两种方法估计的相合程度达到77.35%，证明了两个提升后模型良好的泛化能力。我们指出，Boosting在该任务并没有表现出突出优势，但它具有推广应用的潜力。如何探索输入MFCC数据更合理的表征，使用音素建模（现在没有成功），使用更合适的模型和训练结构，是下一步的探索工作。