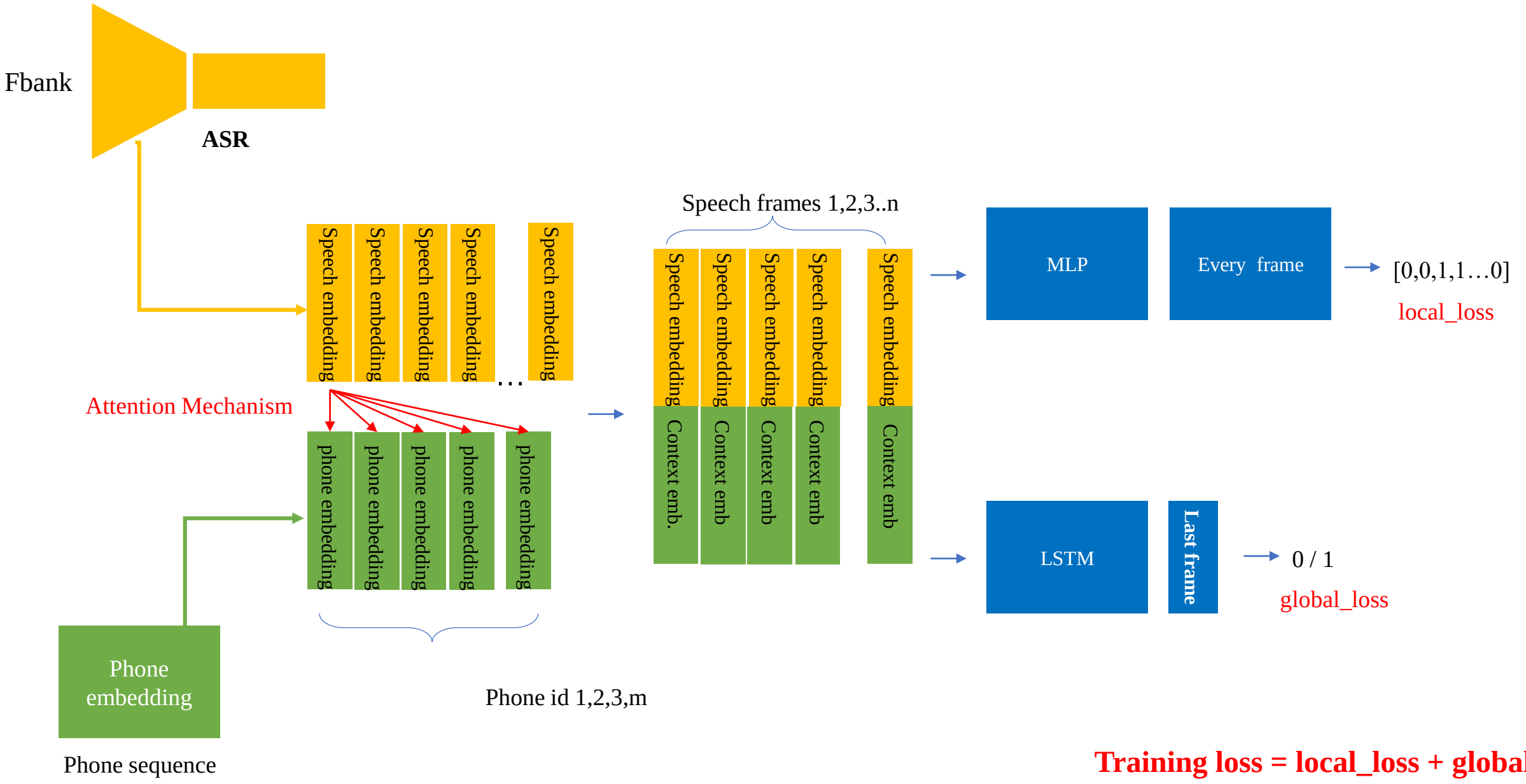
当前架构:



Training loss = local_loss + global_loss

当前主要结果Recall:

Data	th=0.1	th=0.2	th=0.3	th=0.4	th=0.5
Aishell	0.95	0.94	0.92	0.90	0.89
Fenyinta	0.67	0.60	0.54	0.49	0.45

当前存在问题：在 [sil,sil…sil, k1, k2, k3, k4,..kn sil, sil, sil] 这类的数据上表现不佳

已尝试过的方案:

1. 在text embedding 中加入garbage phone
2. 在training数据中加入 [sil,sil…sil, k1, k2, k3, k4,..kn sil, sil, sil] 这类训练数据
3. 改变attention的方向(text attention speech)

以上方案在分音塔的测试数据上都没有得到很好的效果。

分析初始方案，即：

不加garbage node

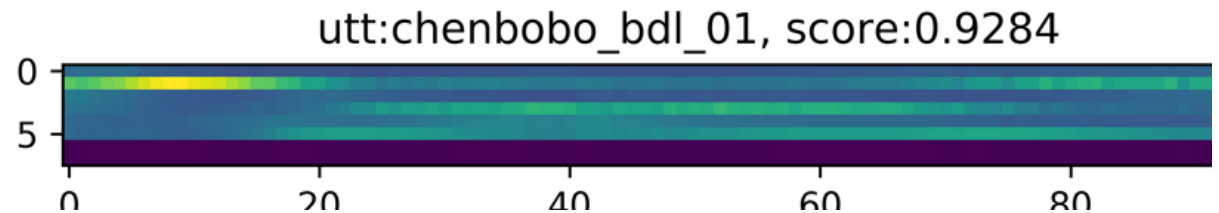
不人为引入[sil,sil…sil, k1, k2, k3, k4,..kn sil, sil, sil]数据

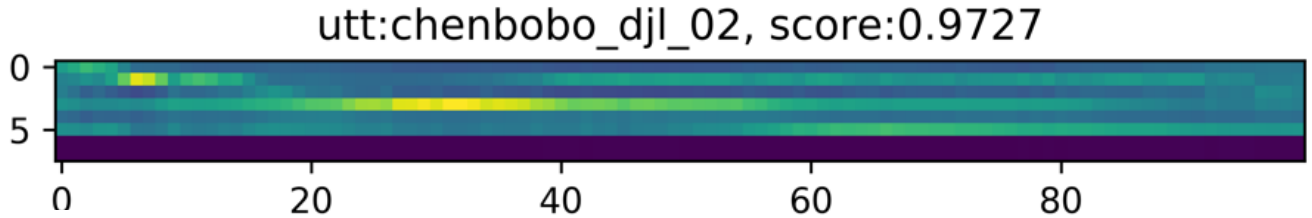
attention方向为speech attend text

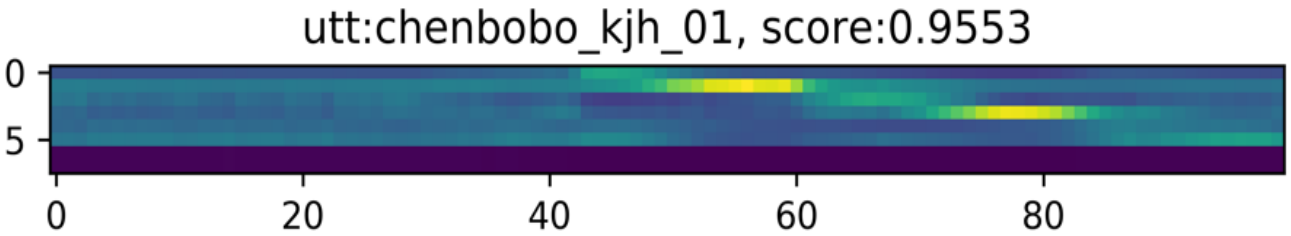
Attention weight 可视化：

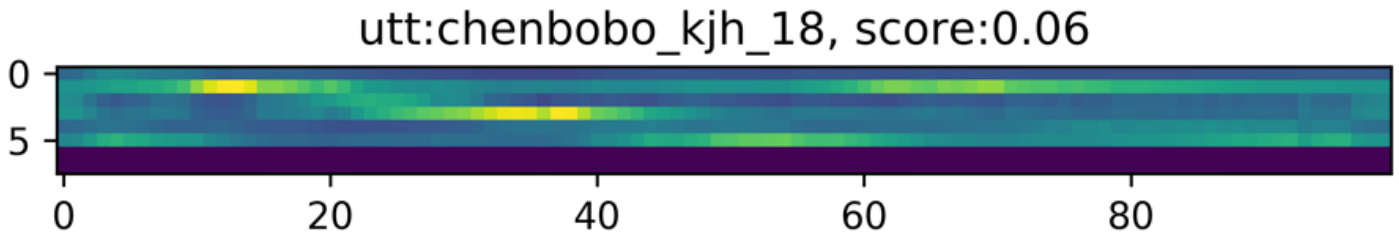
分音塔测试数据，横轴为时间，纵轴为命令词，以下五张图均为命令词段的语音

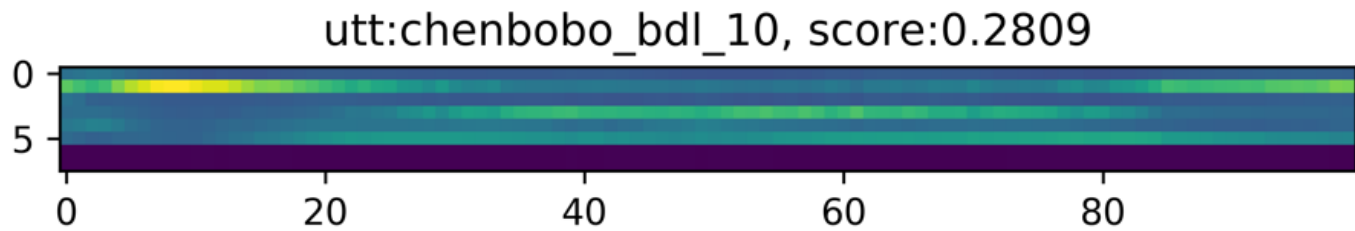
对应的attention weight 上三幅图score较高， 下三幅图score较低，但是图中大致均可看出隐约的对角线现象。





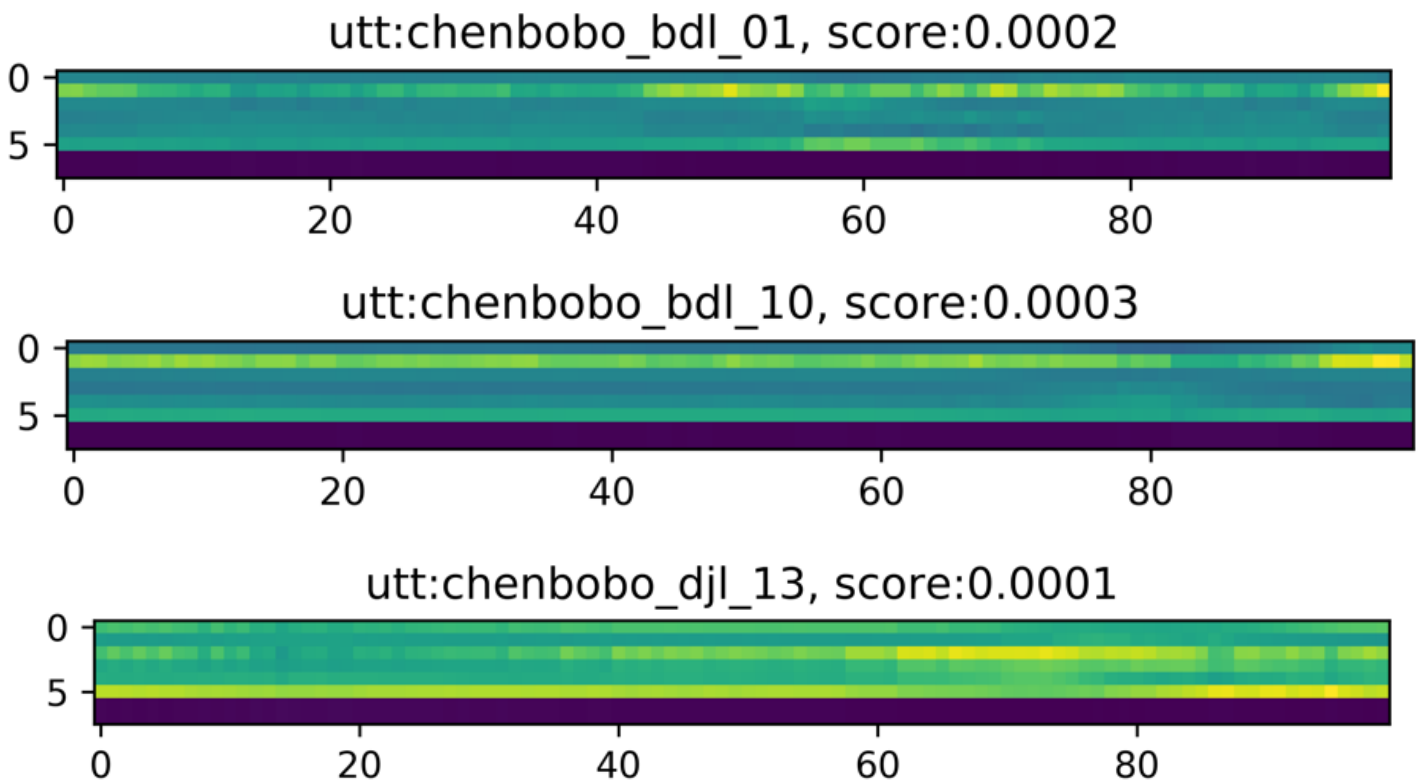






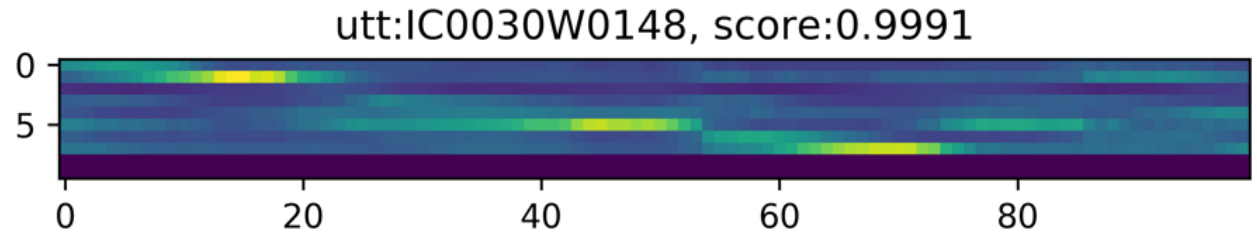
Attention weight 可视化:

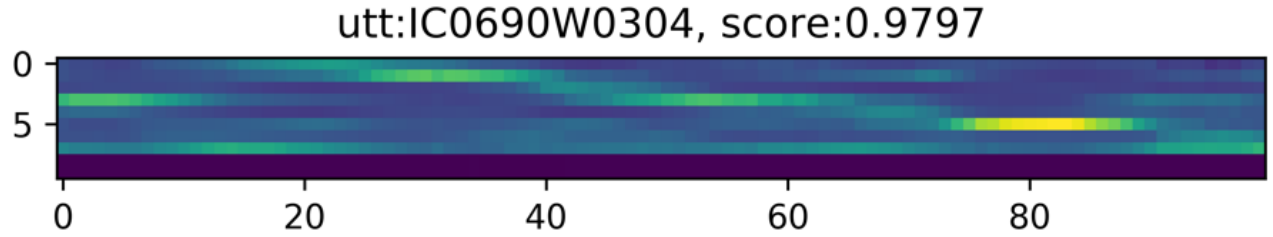
分音塔测试数据，非命令词段，可以看到attention很乱

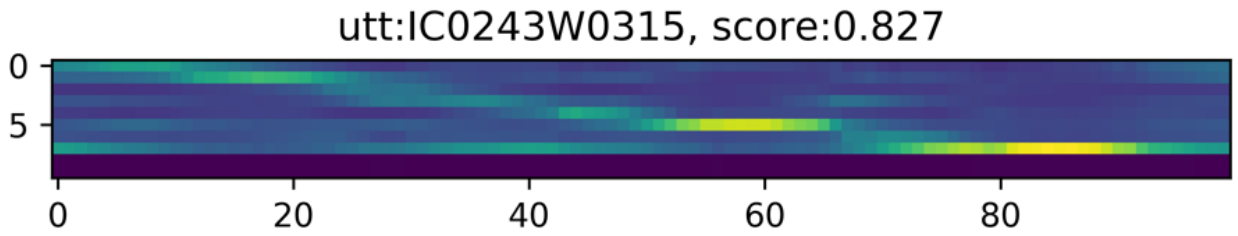


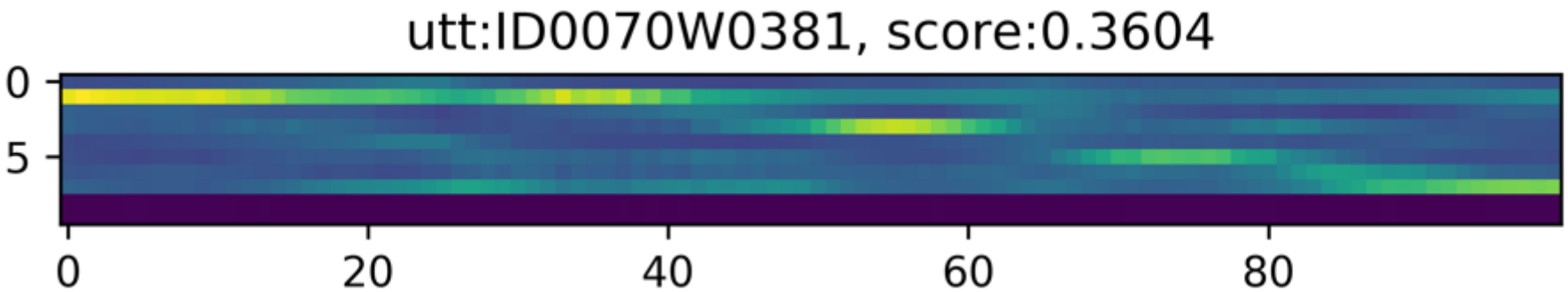
Attention weight 可视化：

Aishell测试数据，命令词段，也可以看到对角线现象，尤其是最后一幅图，分数较低的情况下对角线现象依然存在，从这个角度看，命令词段的语音，attention weight在两个数据集上的表现差异并不明显



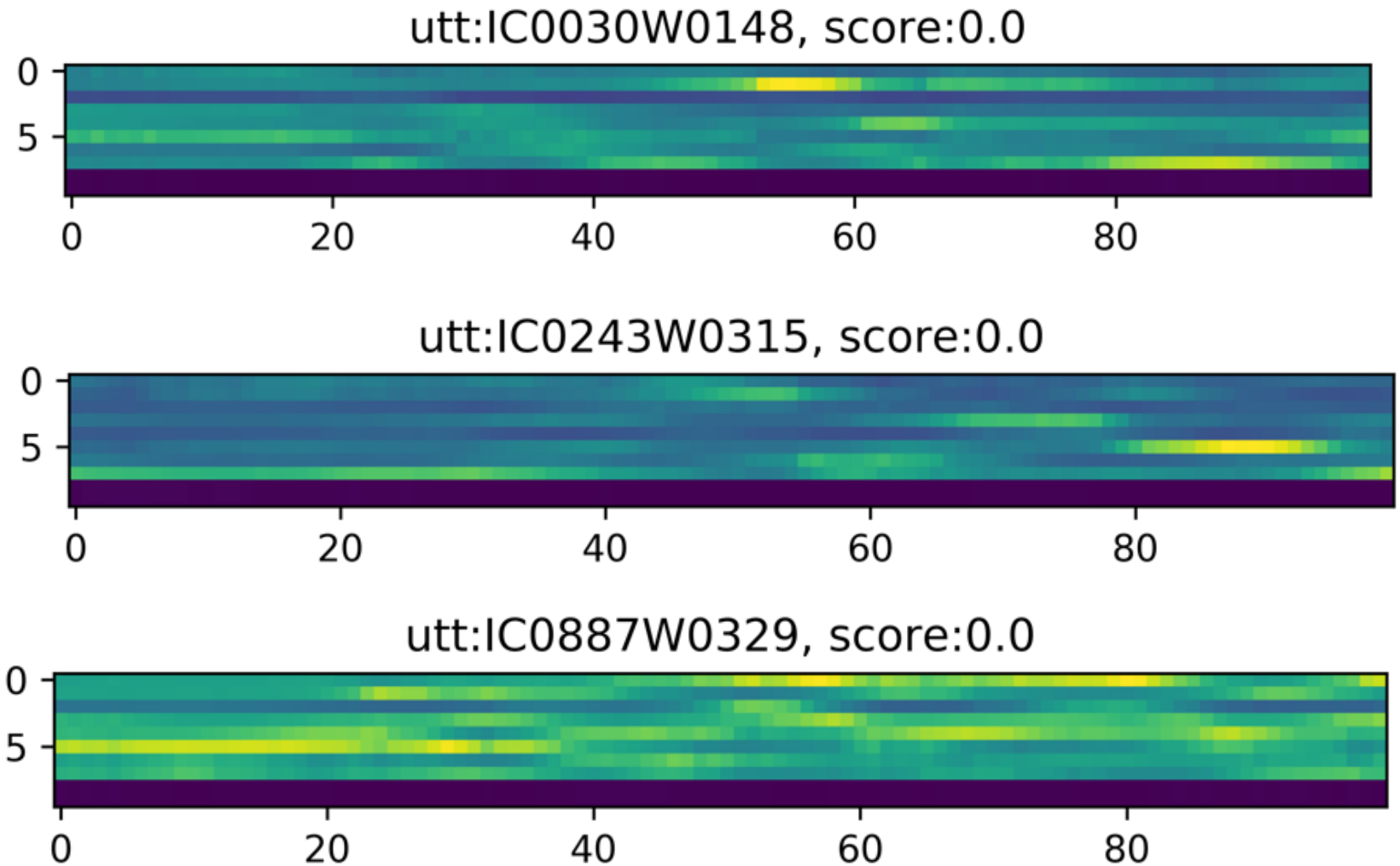






Attention weight 可视化:

Aishell测试数据，非命令词段，对角线现象弱化



从目前的现象来看，attention weight的差距在两个数据集上并不明显，所以两个数据集上性能的差异，可能是因为Speech embedding加入了单头self Attention 导致speech部分过拟合