信也科技杯 FINVOLUTION BIG DATA COMPETITION | 信也 FINVOLUTION

数智创新 声至未来

DEEP IN DIALECTS, FOR FUTURE WAVE

第八届信也科技杯算法大赛

THE 8TH FINVOLUTION DATA SCIENCE COMPETITION

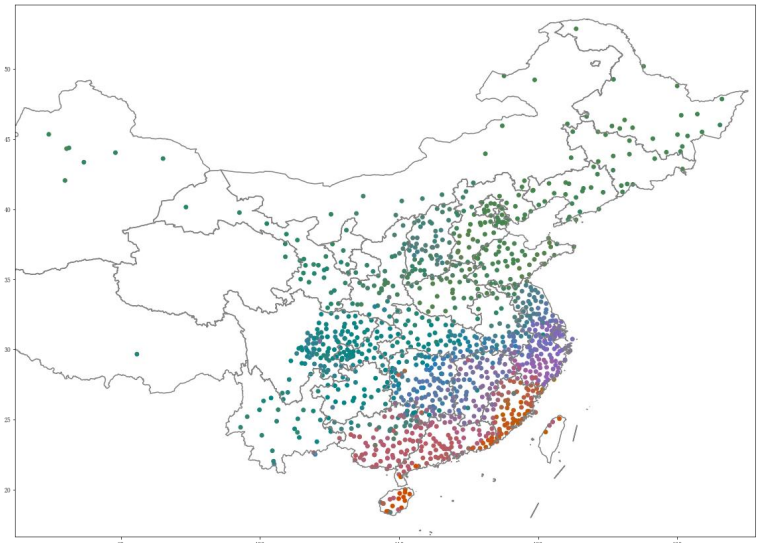
目录

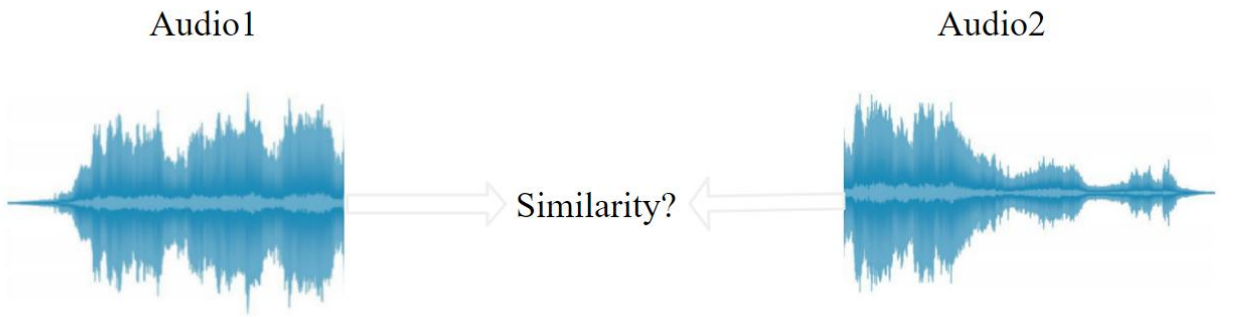
CONTENTS

一、问题分析

1.1 赛题目的

估计不同方言之间的距离特征，以便提供更好的语音服务。





- seen & known
- unseen & known
- unseen & unknown

初赛：

- unseen & known
- unseen & unknown

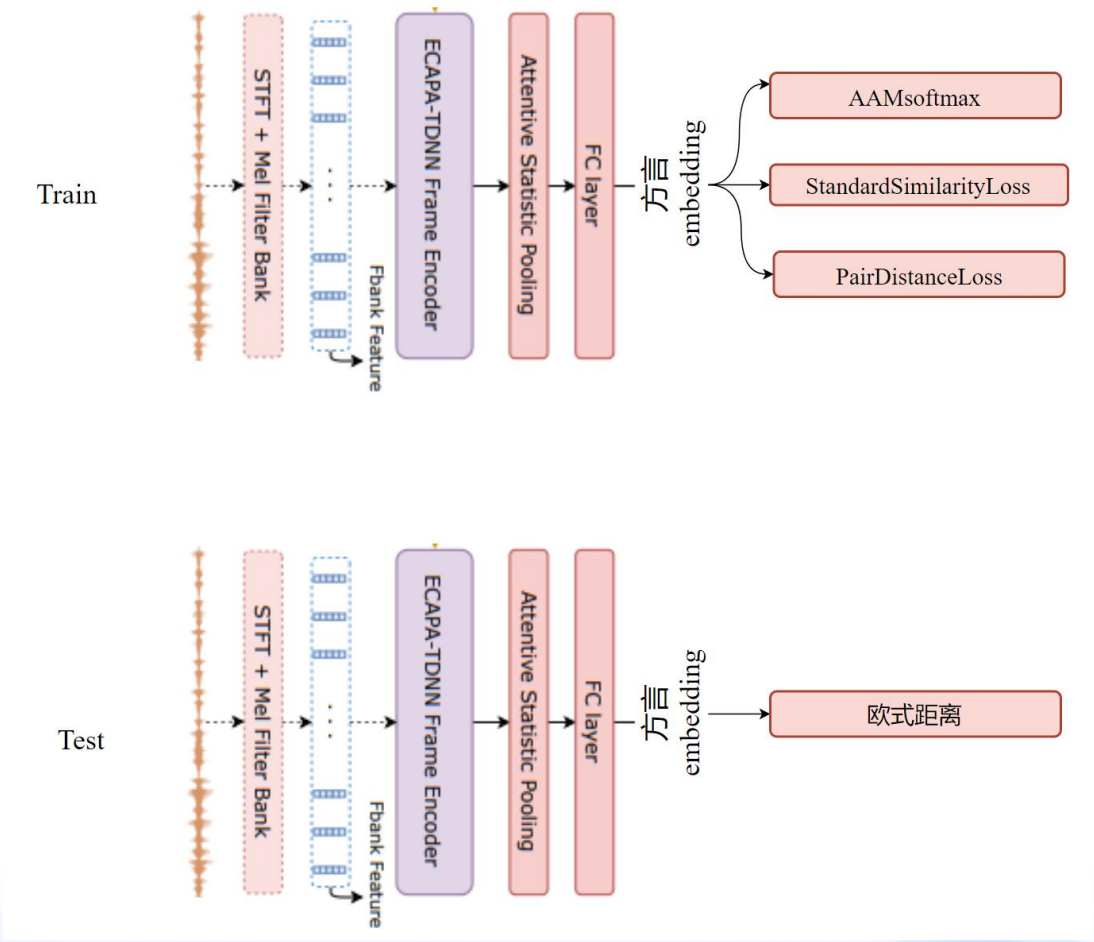
复赛：

- unseen & unknown

1.2 主要挑战

- 如何提升对unseen、unknown 集外数据的表征能力；
- 如何提升对短音频的表征能力；
- 最优化利用内存和运行时间限制；

二、Baseline



Model：ECAPA-TDNN

- Channel- and context-dependent statistics pooling
- Dimensional Squeeze-and-Excitation Res2Blocks
- Multi-layer feature aggregation and summation

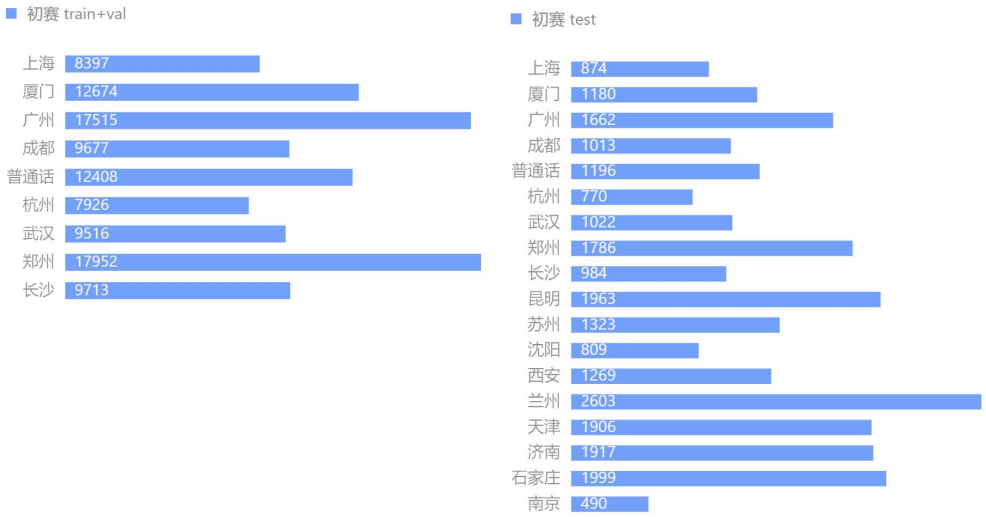
Loss:

- AAMSoftmax：分类问题
- StanddardSimilarityLoss：分类问题，但利用标定距离计算相似度
- PairDistanceLoss：计算batch内两两样本之间距离与标定距离的差异

二、Baseline

主要问题：

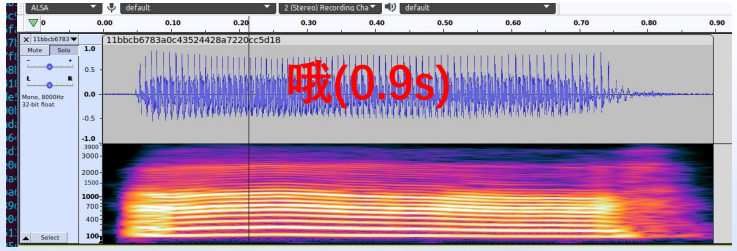
- 特征、模型结构较为简单；
- 数据量少、方言类别少：

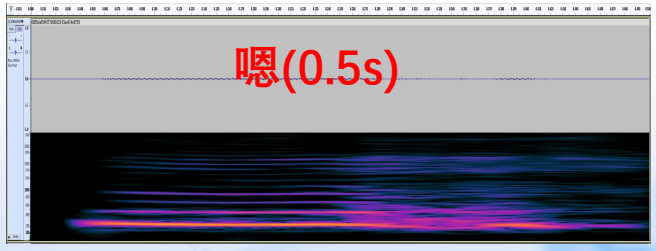


- Baseline 训练目标主要优化集内-集内:（AAMSoftmax>StandardSimilarityLoss>PairDistanceLoss）(集外-集外 > 集内-集外 > 集内-集内)，模型过拟合在已知类别上；

ECAPA-TDNN不同训练目标不同数据下的指标	集内-集内	集内-集外	集外-集外
AAMSoftmax	**38.17**	37.21	23.22
StandardSimilarityLoss	**32.03**	30.78	22.81
PairDistanceLoss	**8.74**	**14.20**	**21.61**

- 部分数据过短，人也很难分辨；



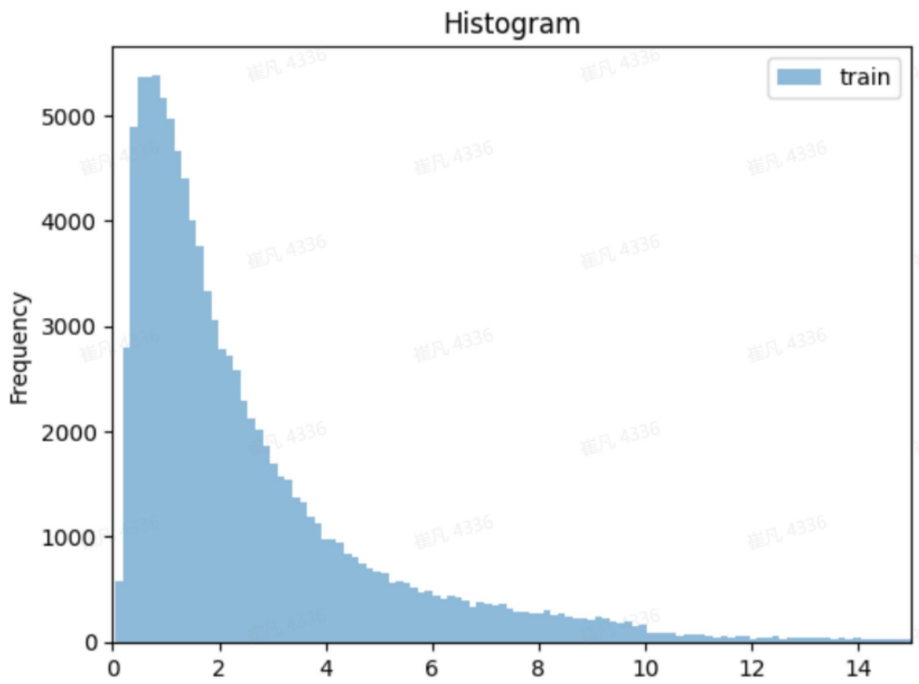


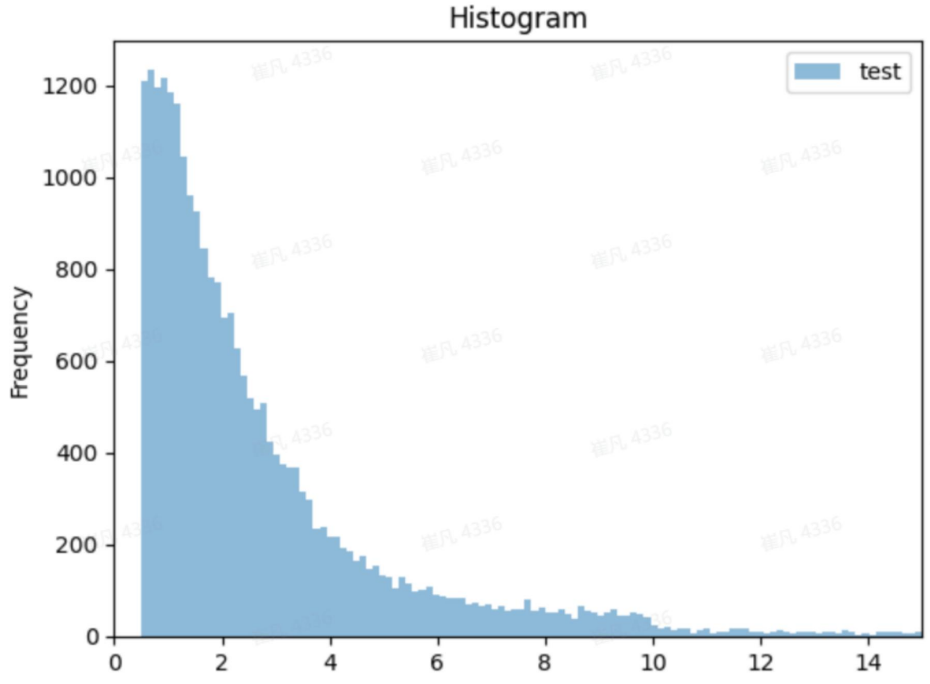
二、Baseline

算法优化方向：

- 结合各种数据增强优化技术提升模型的泛化性；
- 训练数据和训练数据类别有限，利用预训练模型提升模型的表征能力；
- 优化算法训练目标，提升对未知类别数据的建模能力；

三、数据优化——短数据生成与拼接

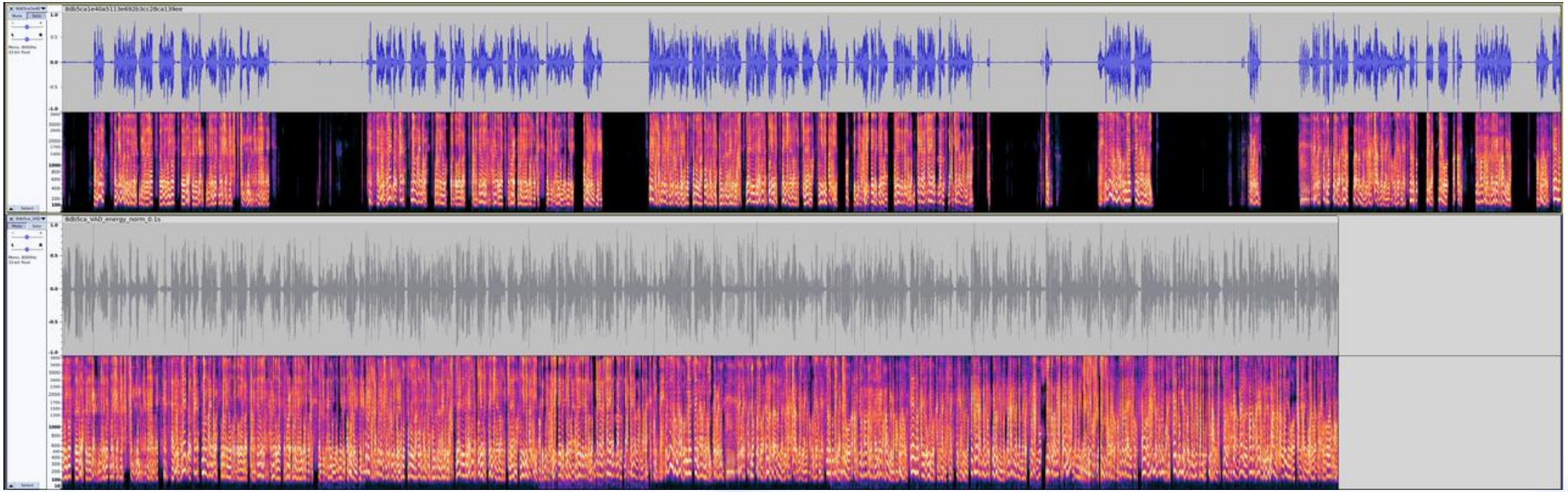




- 训练时长度大于3s的数据有50%的概率random clip成0.5s-3s，充分利用长音频不同位置的信息，也提升对短音频的表征能力；
- 训练时长度小于3s的数据有50%的概率用自身循环拼接（wrap padding），与测试时策略保持一致；另外50%随机采样同一方言类别数据进行concat augmentation，提升类别内表征的一致性；

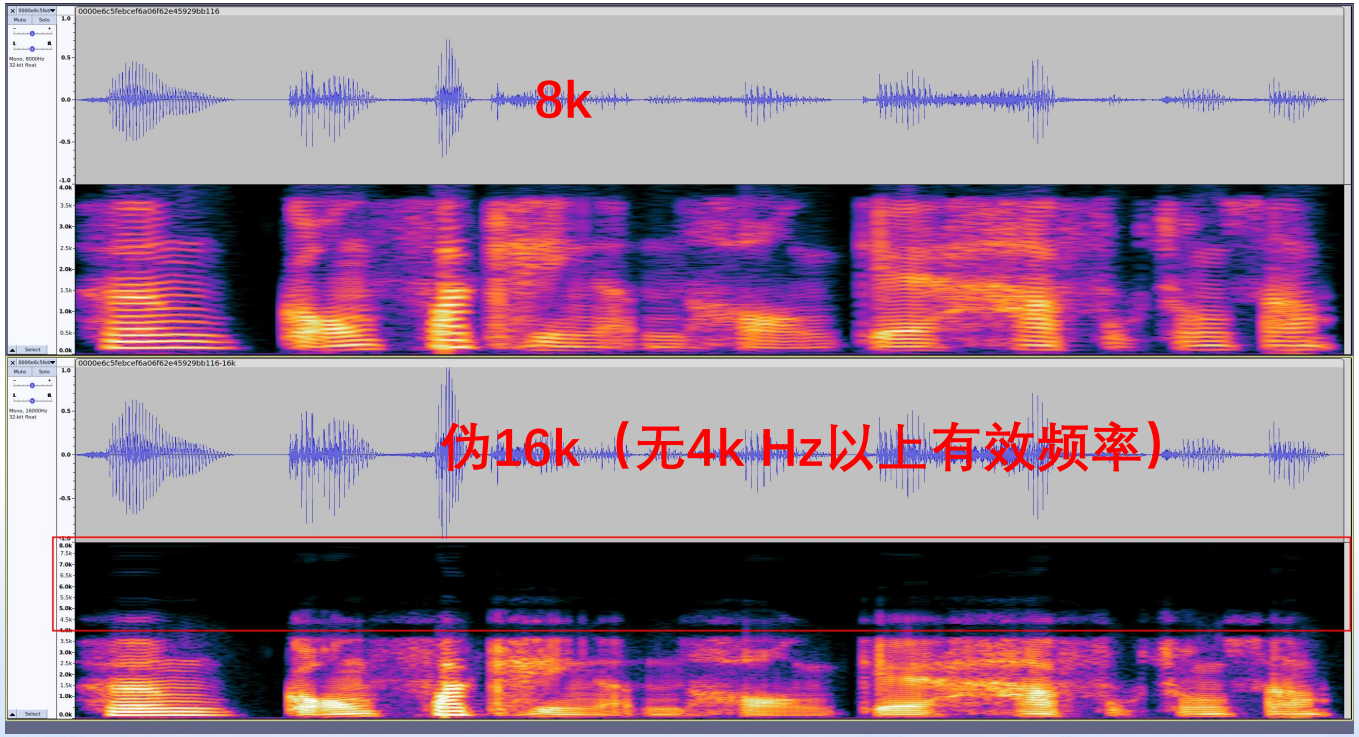
三、数据优化——VAD

采用基于能量的VAD去除长静音段，减少无信息静音段的干扰

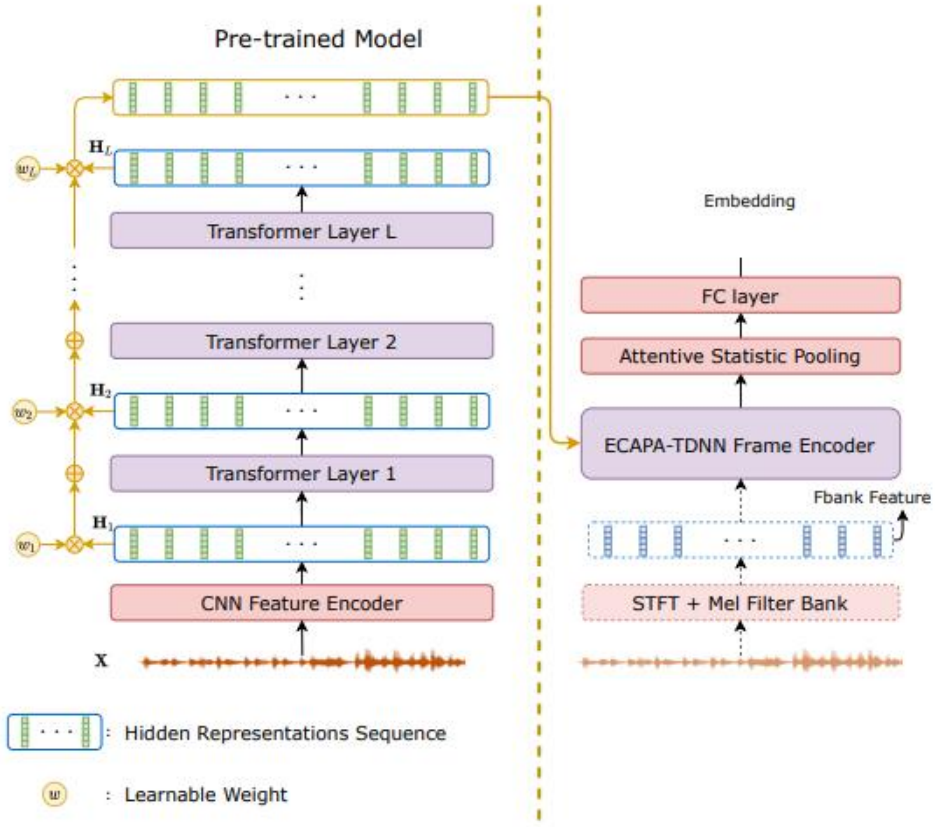


三、数据优化——数据增广

- 加噪声：SLR17_MUSAN
- 加混响：SLR28_RIR_NOISES
- 语速扰动（变速变调、变速不变调）
- 音量扰动
- SpecAug
- Resample：8k->伪16k (与开源预训练模型采样率保持一致)



四、模型优化——使用开源预训练模型



利用预训练模型的优势：

- 提升对unseen & unknown集外数据的表征能力；
- 减少数据分布等干扰影响。

预训练模型(hugging face):

- WavLM-Base
- WavLM-Large
- HuBert-Base
- HuBert-Large
- MMS
- Chinese-Wav2Vec2-Base
- Chinese-Wav2Vec2-Large
- Wav2Vec2-Large-xlsr-53-chinese-zh-cn
- Chinese-Hubert-Base
- Chinese-Hubert-Large

不同预训练模型初赛数据结果	ECAPA-TDNN	WavLM-Base	WavLM-Large	Hubert-Base	Hubert-Large	MMS	Chinese-Wave2Vec2-Base	Chinese-Wave2Vec2-Large	Wav2Vec2-Large-xlsr-53	Chinese-Hubert-Base	Chinese-Hubert-Large
指标	10.6	8.27	7.9	7.7	7.6	7.6	7.4	7.3	7.4	**7.3**	**7.12**

- **中文数据预训练的模型效果更好；**
- **Large模型相比Base模型结果更好；**

四、模型优化——预训练模型调优

针对8K方言数据重新训练预训练模型

- 整理了约2.7万小时中文数据训练集：
Tencent Chinese hubert使用1万小时wenetspeech数据进行预训练，其中含有大量的播客数据，这限制了预训练模型的泛化能力。我们整理了更多的开源数据，包含更多口语风格的对话数据，各种方言口音数据，丰富的数据类型可以提高模型的鲁棒性。

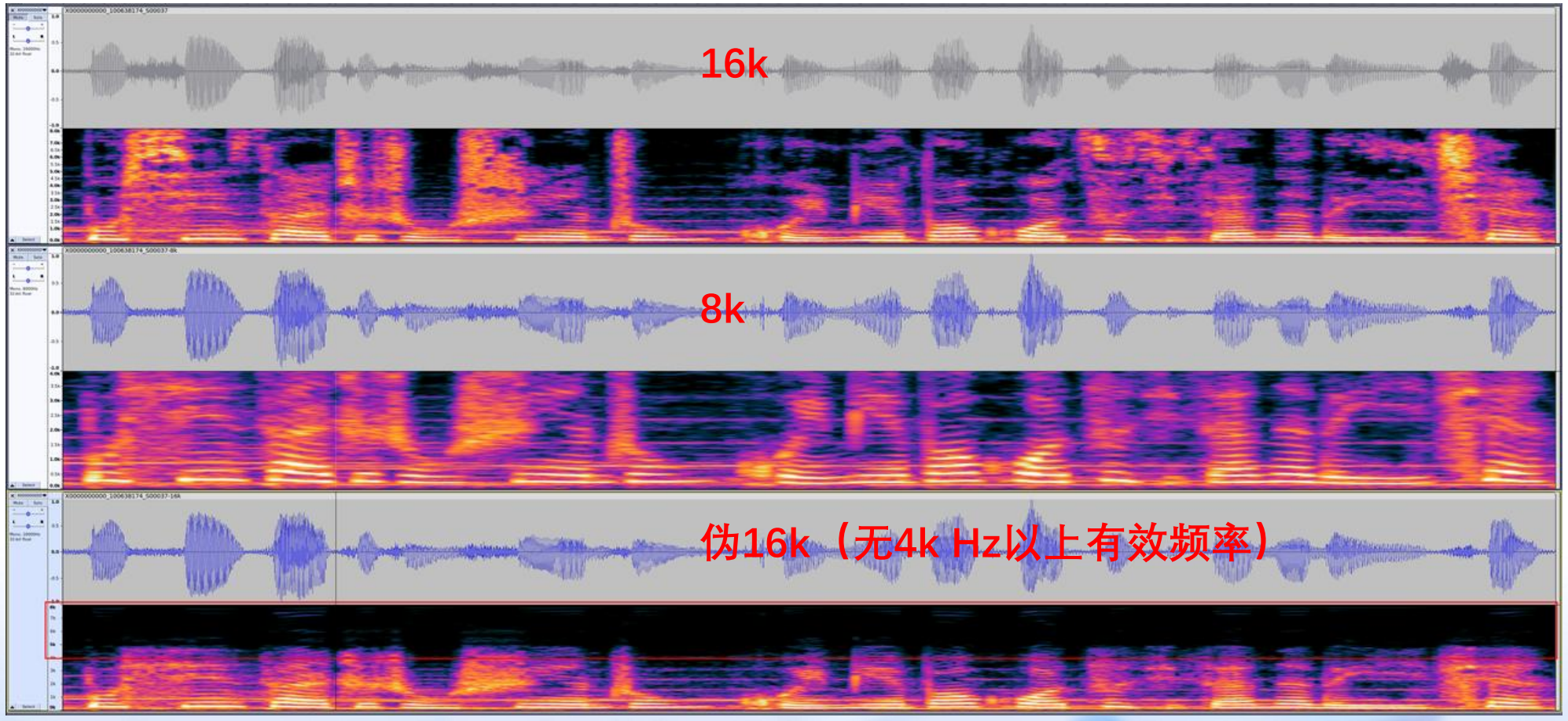
Training Sets	SLR18	SLR33	SLR38	SLR47	SLR62	SLR68	SLR82	SLR87	SLR93	SLR111	SLR119	SLR121	SLR123	SLR124	SLR133
	THCHS-30	aishell	ST-CMDS	Prime-words	Aidatatang	magicdata	CN-Celeb	mobvoi	Aishell3	aishell4	alimeeting	Wenetspeech	MAGICDATA	Tibetan	XBMU
Hours	34	178	110	99	200	755	1358	224	86	178	1386	12482	155	85	31

	kespeech	3D-Speaker	VoxCeleb1	VoxCeleb2	HAL	CommonVoice	Finvcup								TOTAL
Hours	1542	1275	352	2460	645	3600	106								27341

四、模型优化——预训练模型调优

针对8K方言数据重新训练预训练模型

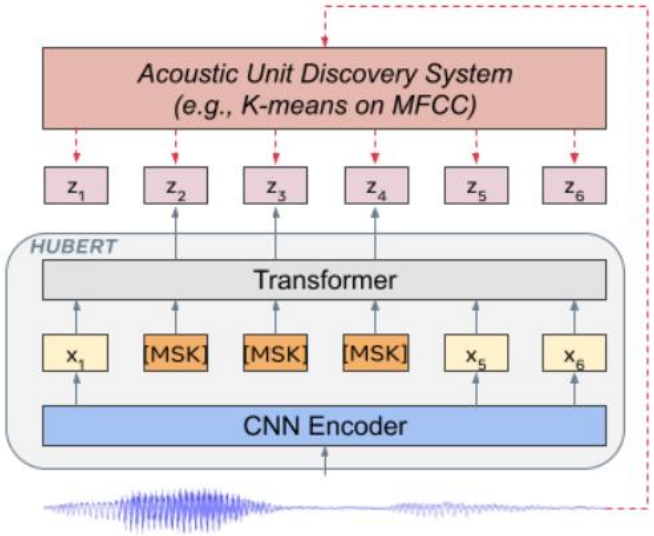
- 对16k开源数据进行重采样（16k->8k->伪16k），重新训练适配竞赛数据（8k->伪16k）的hubert-base、hubert-large 模型



四、模型优化——预训练模型调优

针对8K方言数据重新训练预训练模型

- 自训练HuBERT模型



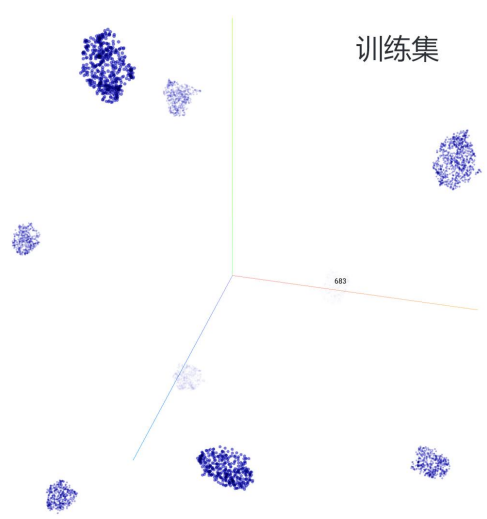
Model	Num. of Params	Pseudo label
HuBERT-it1	95M	K-Means {50,100,500} on the 39D MFCC+delta+ddelta
HuBERT-it2	95M	K-Means 500 on the features from the 6th transformer layer of the HuBERT-it1
HuBERT-base	95M	K-Means 500 on the features from the 9th transformer layer of the HuBERT-it2
HuBERT-large	317M	K-Means 500 on the features from the 9th transformer layer of the HuBERT-it2

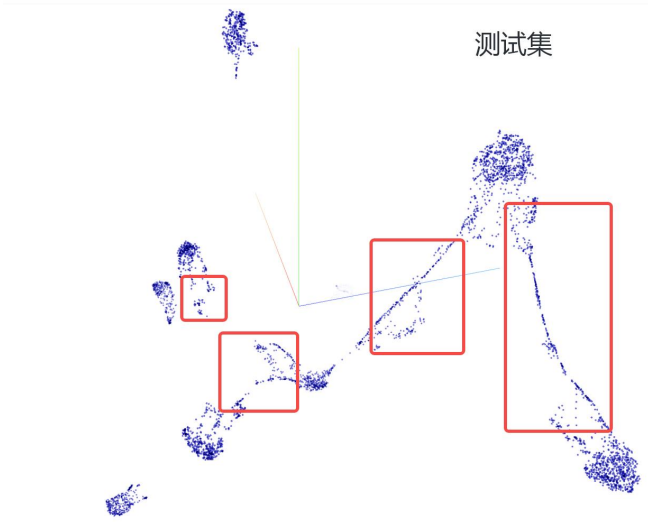
- Fine-tune结果

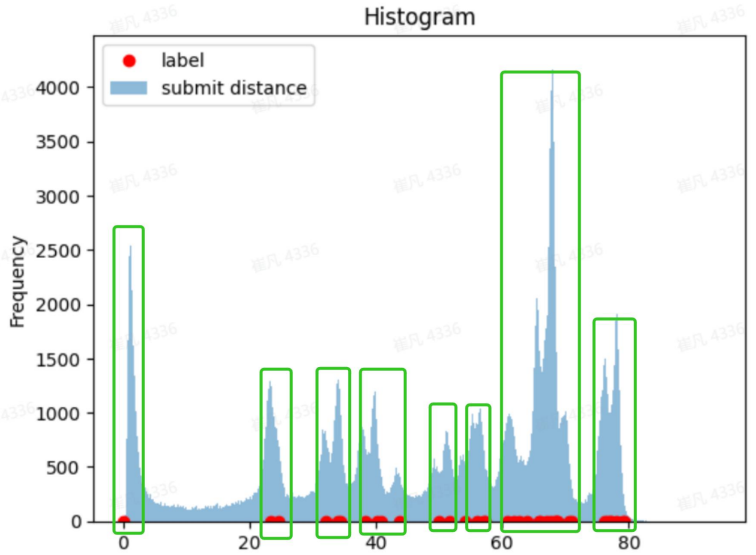
不同预训练模型初赛测试结果	Chinese-Hubert-Base	自训练Hubert-Base (1w小时)	自训练Hubert-Base (2.7w小时)
指标	7.3	7.28	7.22

五、训练目标优化——Embedding分析

训练集样本embedding聚类分布**类内距离小，类间距离大**，符合预期；



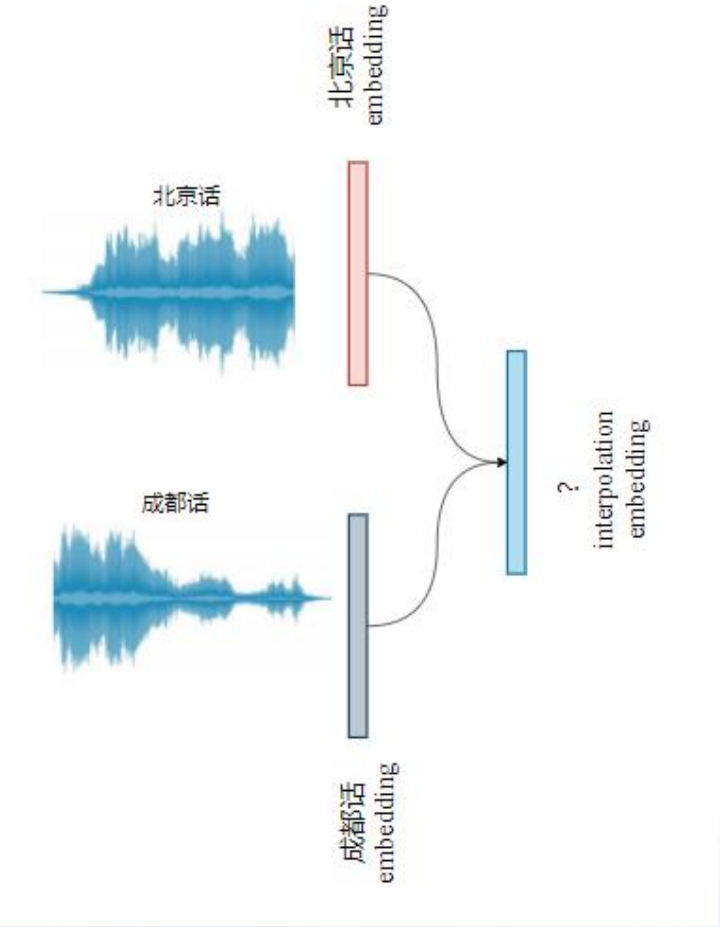


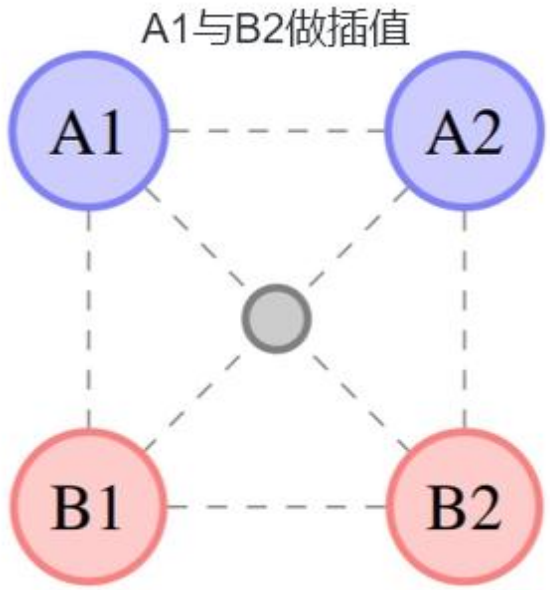


测试集样本embedding聚类分布**类间区分度不高，且大部分聚集在9个已知类别上**，不符合预期；

五、训练目标优化——Embedding Mixup

针对集外数据结果较差的问题，利用集内数据生成未知类别数据：





其中，灰色为生成的新数据，当A1、A2分布相近，B1、B2分布相近，在embedding层面上进行插值生成新类别的embedding。

五、训练目标优化

- 通过对预训练模型的hidden embeddings进行插值生成多阶混合embedding:

$$二阶混合embedding: E' = E_i * a + E_j * a,\ (a = 1/2)$$

$$三阶混合embedding: E'' = E_i * a + E_j * a + E_k * a,\ (a = 1/3)$$

- 多阶混合方言embedding与其他方言embedding的标定距离为:

$$D(E_0,E') = D(E_0,E_i) * a + D(E_0,E_j) * a,\ a = 1/2$$

$$D(E_0,E'') = D(E_0,E_i) * a + D(E_0,E_j) * a + D(E_0,E_k) * a,\ a = 1/3$$

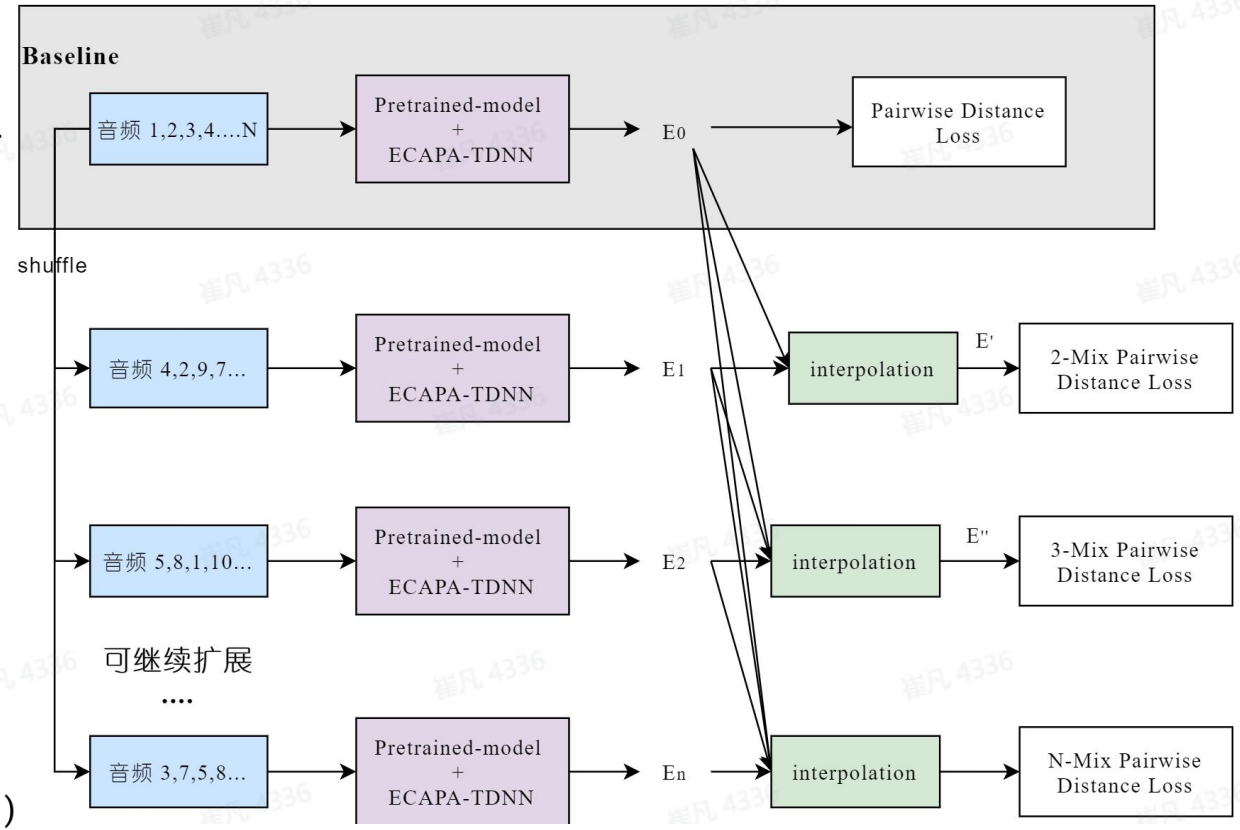
- Multi-Mix PairDistanceLoss:

$$PairDistanceLoss:\ Loss_0 = MSE(CDIST(E_0,E_0,p=2),D(E_0,E_0))$$

$$2MixPairDistanceLoss:\ Loss_1 = Loss_0 + MSE(CDIST(E_0,E'),D(E_0,E'))$$

$$3MixPairDistanceLoss:\ Loss_2 = Loss_1 + MSE(CDIST(E_0,E''),D(E_0,E''))$$

其中，D(E0，E0) 表示batch内两两数据的聚类矩阵。



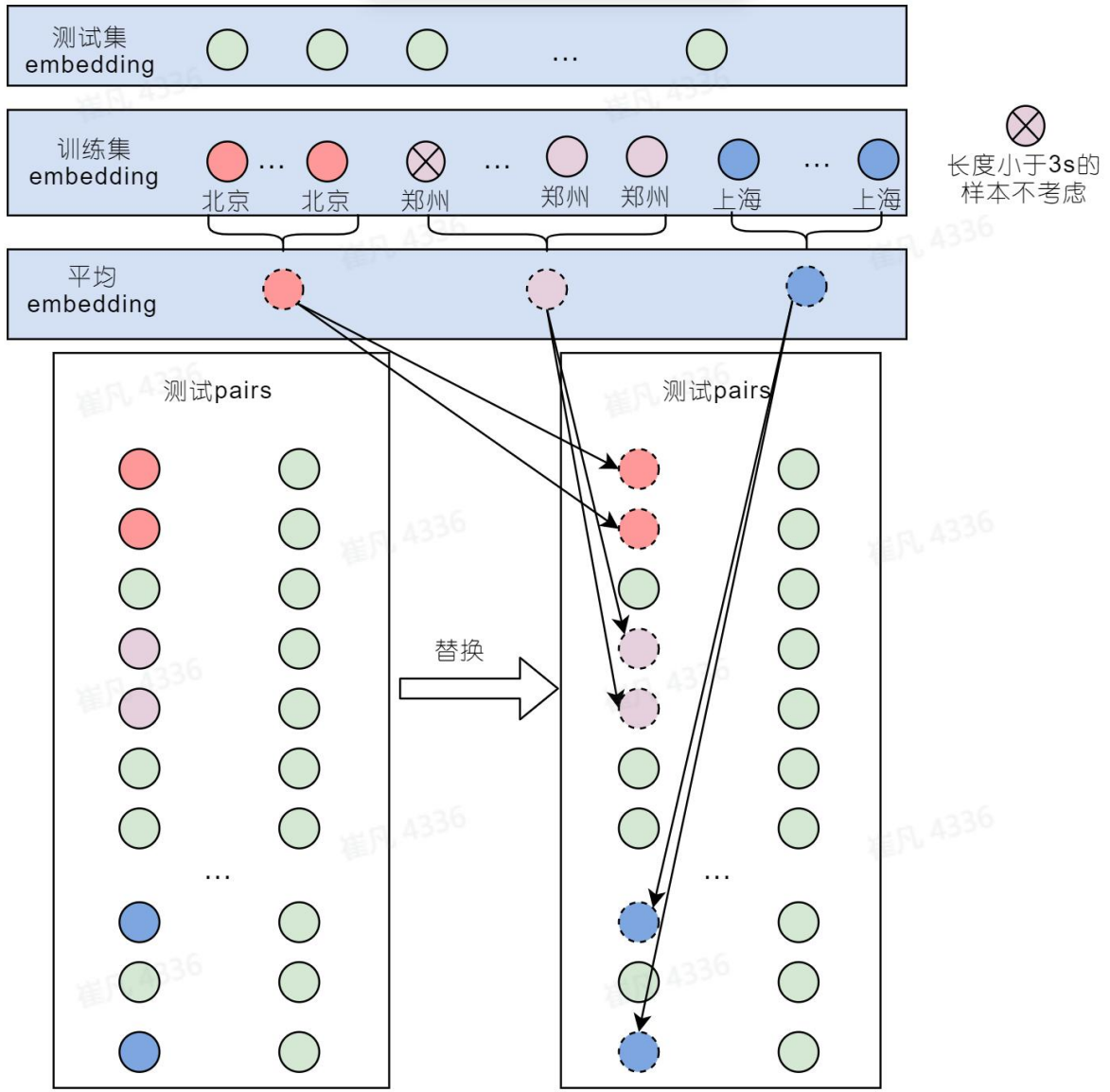
离线测试结果	集内-集外	集外-集外
PairDistance	12.20	23.58
2MixPairDistance	10.41	23.01
3MixPairDistance	8.93	23.11
4MixPairDistance	**8.11**	**22.84**
5MixPairDistance	**8.14**	**23.67**

六、后处理——计算方言平均embedding

对于测试样本中已知类别的情况用方言平均embedding替换（音频长度小于3s的样本不用于计算方言平均embedding）；

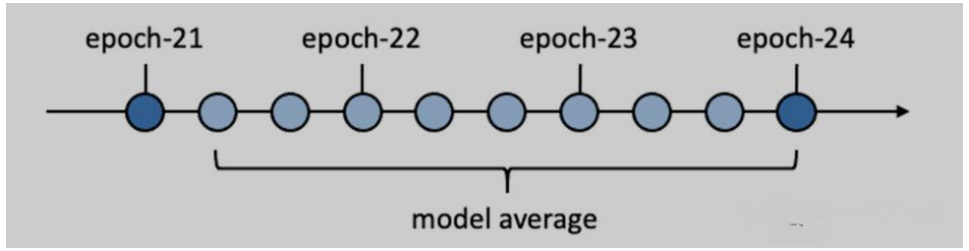
- 减少推理时间（13W->7W）；
- 减少噪声样本影响（短音频、或错误样本）提升结果；

离线测试结果	集内-集外	集外-集外
未替换	12.20	23.58
替换为平均embedding	11.89	23.58



六、后处理——更细粒度的模型平均

将两个模型之间的所有更新模型进行平均，利用更细粒度的模型平均获得采样点更加密集，即噪声（随机性）更低的平均模型。最终实现结果，只需要加载两个模型，即可计算二者之间的所有模型平均值。具体参考： https://shorturl.at/gnprN



$$model\_avg_{[1,n]} = \frac{1}{n}\sum model_j$$

$$model\_avg' = model\_avg_{[1,n]} \times \frac{n}{n+1} + model_{n+1} \times \frac{1}{n+1}$$

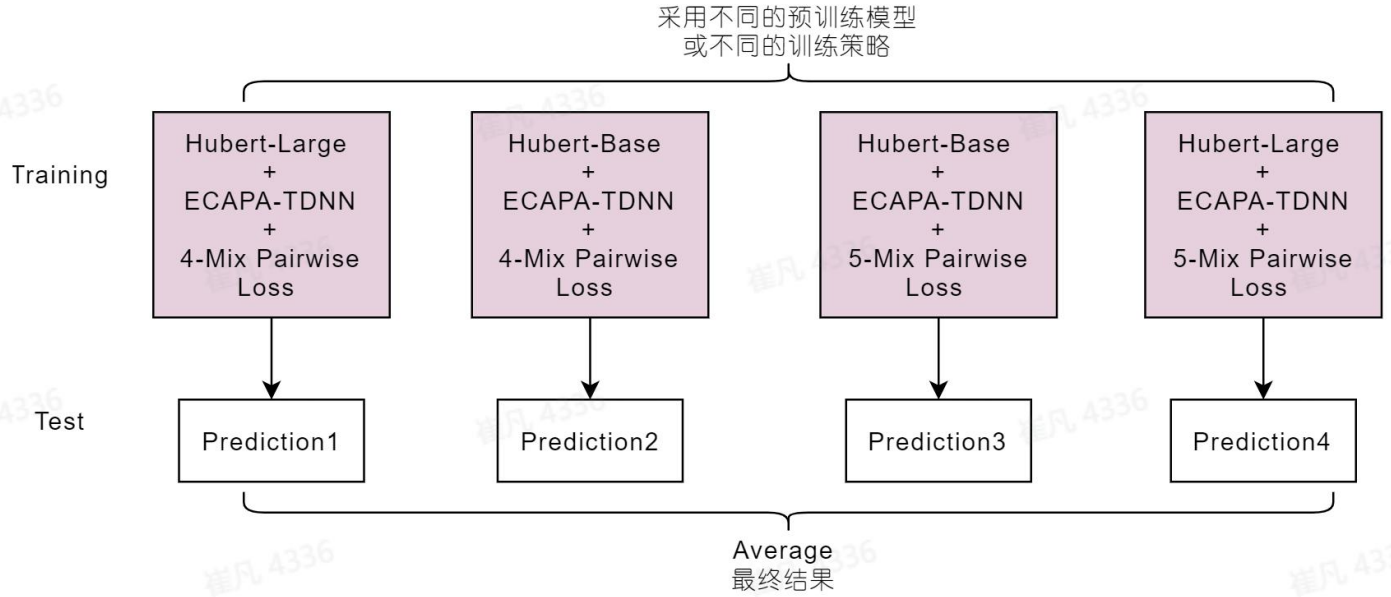
$$model\_avg_{[p+1,q]} = \frac{1}{q-p}(model\_avg_{[1,q]} \times q - model\_avg_{[1,p]} \times p)$$

离线测试结果	集内-集外	集外-集外
epoch10	11.89	23.58
epoch15	12.68	25.15
ave epoch10-15	11.87	23.50
K2 ave epoch10-15	11.84	23.45

- **需要保存的模型更少；**
- **模型平均后更平滑，结果更好；**

六、后处理——多模型ensemble

融合多个训练好的模型，实现测试数据的多模型融合，使最终的结果能够“取长补短”，提高最终模型的泛化能力：



复赛测试结果	Hubert-base + 4Mix PairLoss	Hubert-Large + 4Mix PairLoss	Hubert-Large + 4Mix PairLoss + Hubert-base + 4Mix PairLoss	Hubert-Large + 4Mix PairLoss + Hubert-base + 4Mix PairLoss + Hubert-base + 5Mix PairLoss
Distance	11.16	11.10	10.57	10.52

七、参赛总结

算法方面主要提升点：

- 预训练模型
- 数据优化: 数据增广、数据拼接
- 训练准则优化: Multi-Mix PairDistanceLoss
- 后处理：模型平均、模型ensemble、平均embedding替换

工程方面优化：

- Baseline 训练准则优化训练时间降低为原来1/2;
  - 减少循环遍历
  - 数据优化
- 测试代码优化单模型推理时间降低为原来1/3;
  - 短音频优化
  - batch推理

信也科技杯 FINVOLUTION BIG DATA COMPETITION | 信也 FINVOLUTION

THANK YOU