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

团队名称：WeLearnNLP

团队成员：汪超、杨玉舒、李青、王梓聪、李璐

团队单位：同济大学

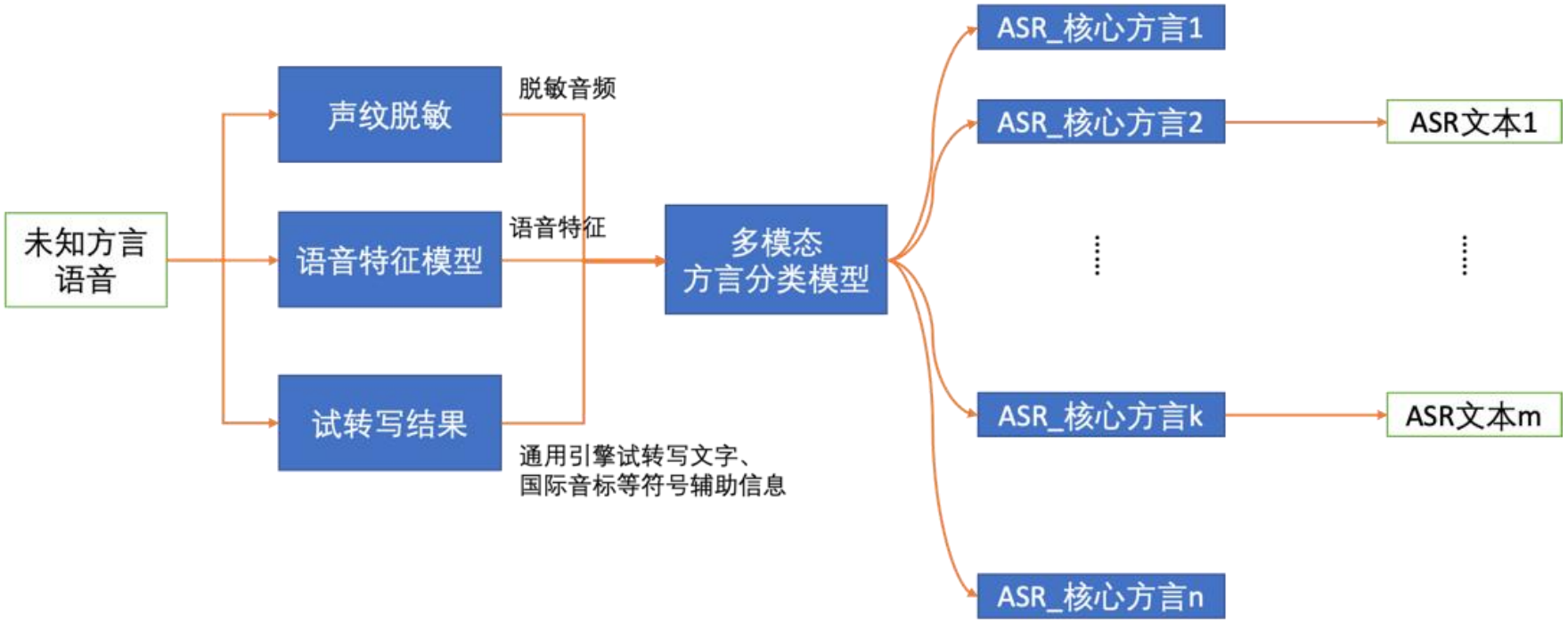
CONTENTS

目录

2. 数据探索

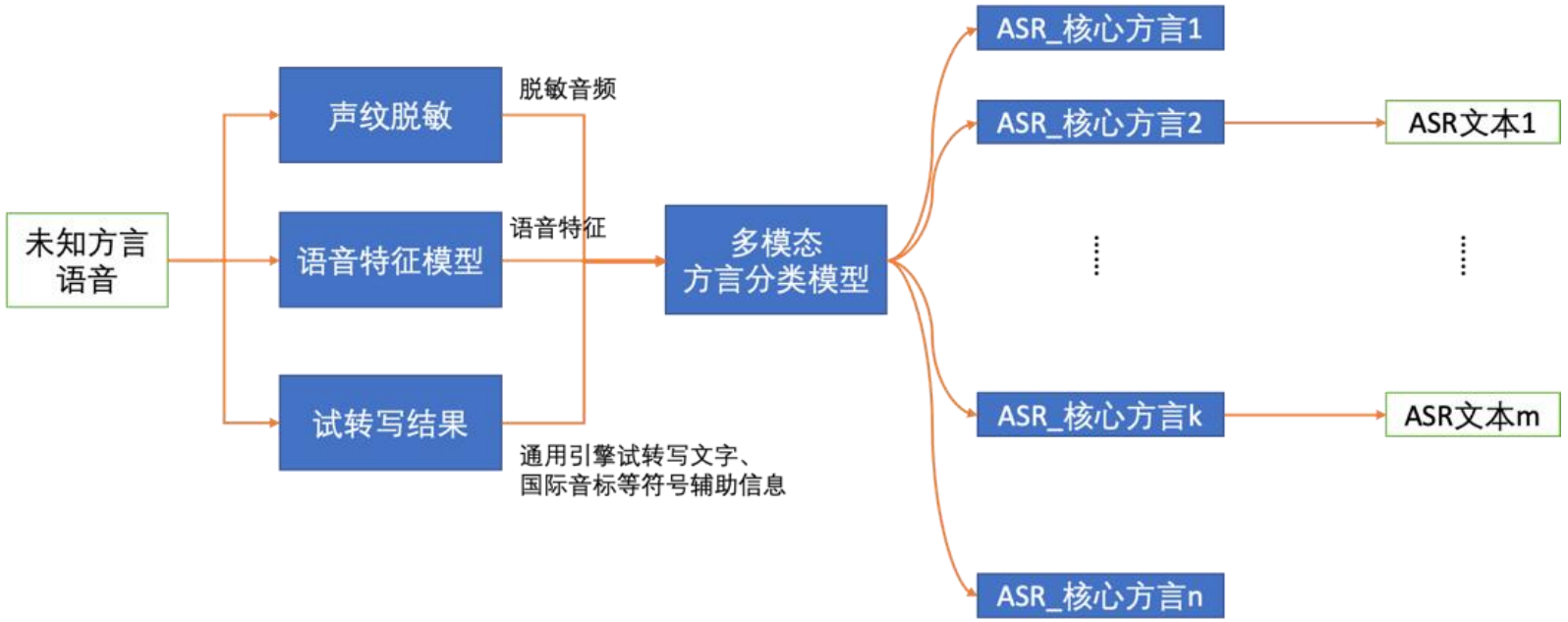
2. 数据探索

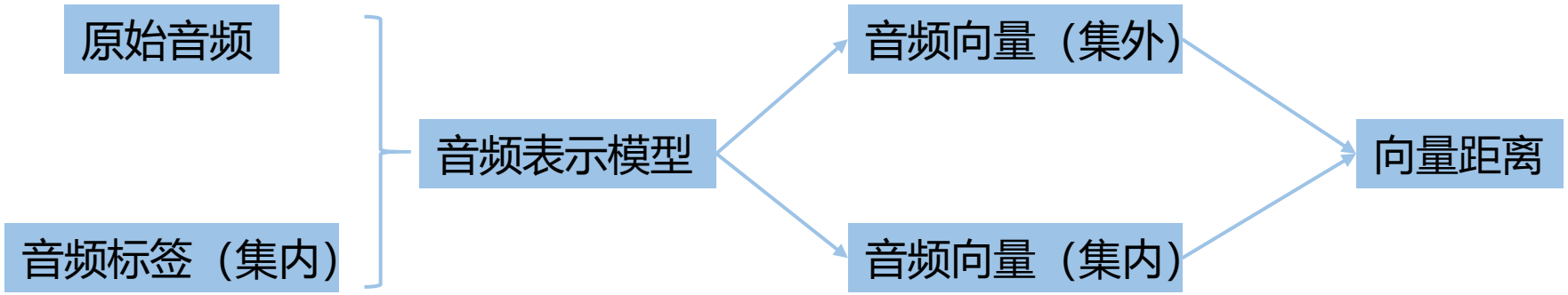
- 在ASR转写中部分客户使用方言进行交谈，这样通用ASR模型就无法转写得到正确的文字。
- 大多数可用的汉语ASR模型不支持方言或覆盖很有限的几种方言。
- 对于国内仍广泛使用、大量存在且种类繁多的方言语音，商业解决方案还不能满足大部分方言的转写。



设置一系列核心方言并建ASR模型，对未知的方言进行鉴别，确定距离其最近的m (m >=1)种核心方言，再尝试用这m种ASR引擎转写该未知方言，转写的不完美结果可用于支持下游任务。

2. 数据探索





训练一个音频表示模型，对音频向量间的距离进行表示，其中部分音频来自标签集合外部。

2. 数据探索

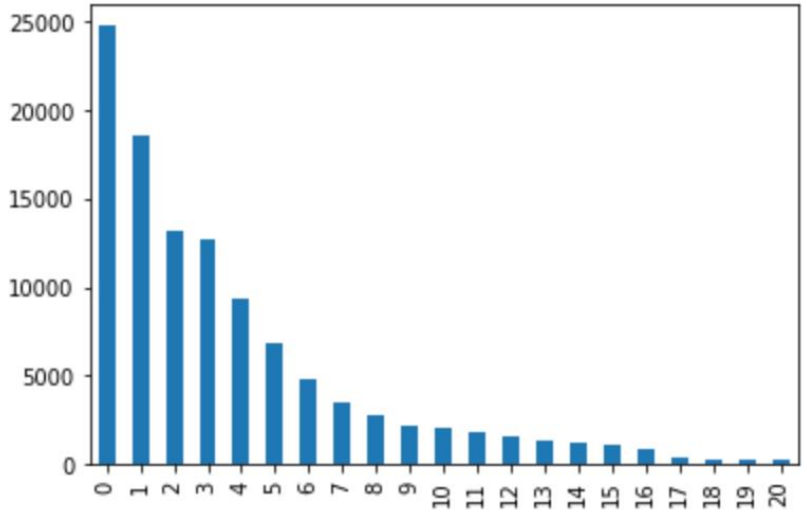
•数据格式：8kHz 16bit

•训练集：该方言数据集包含9种已知方言，截取自各个方言的自然对话语音的切片，分别采集自以下城市的中心城区标准方言：北京（指标准普通话或北京口音）、成都、郑州、武汉、广州、上海、杭州、厦门、长沙，每种方言均含多个说话人，时长约9小时，共计约81小时。

•测试集：除训练数据，还为初赛和复赛提供了额外的d（10 < d < 100）种测试方言的语音。初赛提供约18小时的测试数据，其中包括部分训练方言集外的方言；复赛阶段测试集不予公开。

组委会指定一个从语音集合中抽取的样本对列表，预测结果是各样本对之间的方言距离，构成一个距离向量。评分将采用预测距离和已知距离的差值，并作L1范数。

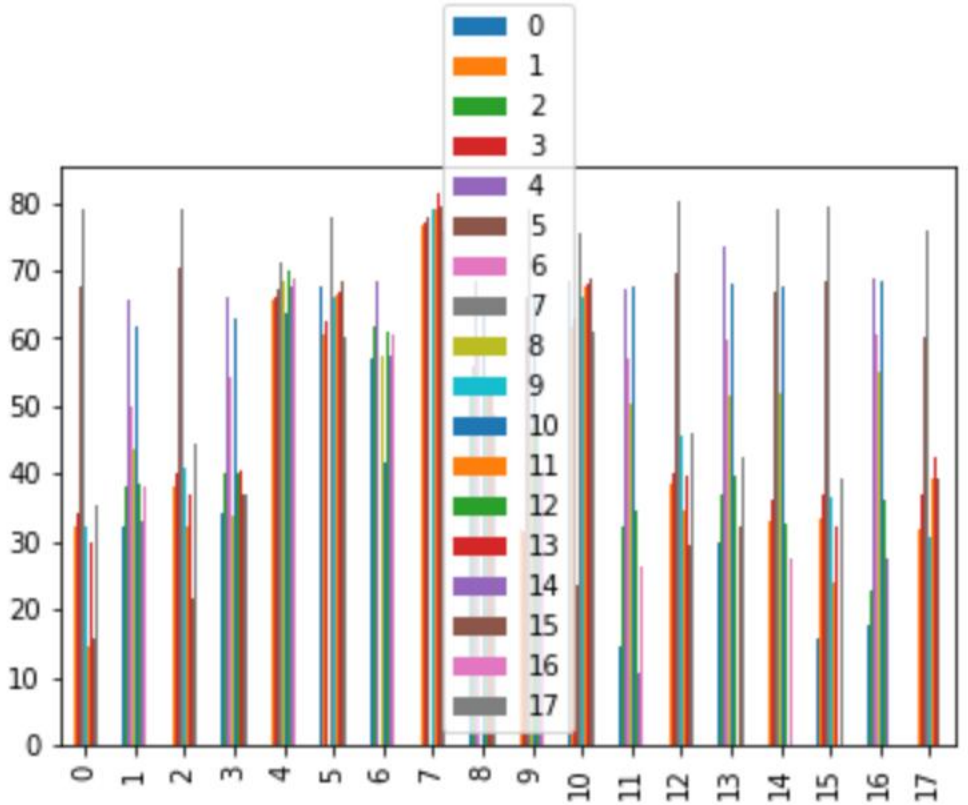
2. 数据探索



音频长度分布

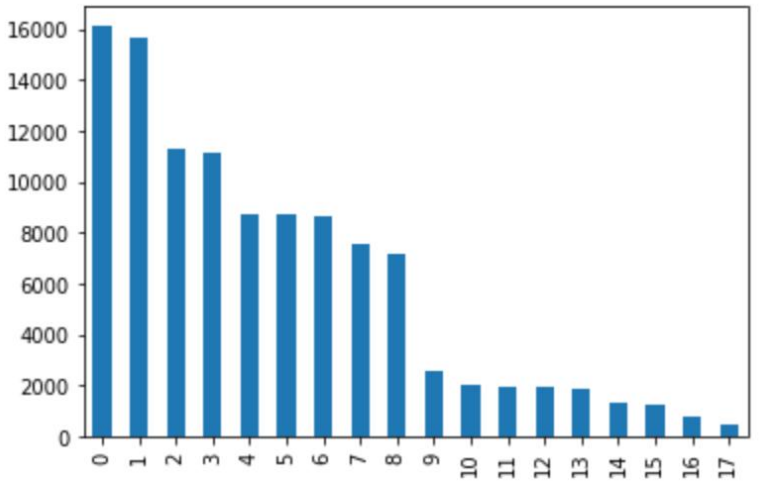
样本对		训练语音集									测试语音集	
		方言 1	方言 2	方言 3	方言 4	方言 5	方言 6	方言 7	方言 8	方言 9	方言 1-9	方言 10-18
测试集	方言 1-9	集内-集内	集内-集内	集内-集内	集内-集内	集内-集内	集内-集内	集内-集内	集内-集内	集内-集内	集内-集内	集内-集外
	方言 10-18	集内-集外	集内-集外	集内-集外	集内-集外	集内-集外	集内-集外	集内-集外	集内-集外	集内-集外	集内-集外	集外-集外

测试集方言对分布



方言对距离分布

2. 数据探索



音频类别分布

样本对	file_2	训练语音集	测试语音集
file_1		方言 1-18	方言 19-...
测试语音集	方言 19-...	集内-集外	集外-集外

复赛测试集样本对分布

	北京	成都	郑州	武汉	广州	上海	杭州	厦门	长沙	昆明	苏州	沈阳	西安	兰州	天津	济南	石家庄	南京
北京	0	32	23	34	69	68	57	79	52	32	68	15	33	30	15	16	18	36
成都	32	0	38	25	66	61	50	77	44	32	62	33	39	38	33	34	38	32
郑州	23	38	0	40	69	71	62	79	56	41	70	32	32	37	32	22	23	44
武汉	34	25	40	0	66	63	54	77	34	31	63	34	40	41	36	37	41	37
广州	69	66	69	66	0	67	68	71	68	69	64	67	70	73	68	69	69	64
上海	68	61	71	63	67	0	41	78	64	66	24	66	69	67	67	68	70	60
杭州	57	50	62	54	68	41	0	76	57	57	42	57	61	60	57	59	61	51
厦门	79	77	79	77	71	78	76	0	77	79	76	79	80	81	79	79	79	76
长沙	52	44	56	34	68	64	57	77	0	47	64	51	53	52	52	53	55	49
昆明	32	32	41	31	69	66	57	79	47	0	66	39	46	40	40	37	41	31
苏州	68	62	70	63	64	24	42	76	64	66	0	68	70	68	68	69	69	61
沈阳	15	33	32	34	67	66	57	79	51	39	68	0	35	33	11	24	26	39
西安	33	39	32	40	70	69	61	80	53	46	70	35	0	40	33	29	36	46
兰州	30	38	37	41	73	67	60	81	52	40	68	33	40	0	32	32	37	42
天津	15	33	32	36	68	67	57	79	52	40	68	11	33	32	0	25	27	39
济南	16	34	22	37	69	68	59	79	53	37	69	24	29	32	25	0	19	40
石家庄	18	38	23	41	69	70	61	79	55	41	69	26	36	37	27	19	0	44
南京	36	32	44	37	64	60	51	76	49	31	61	39	46	42	39	40	44	0

复赛训练集已知类别对距离矩阵

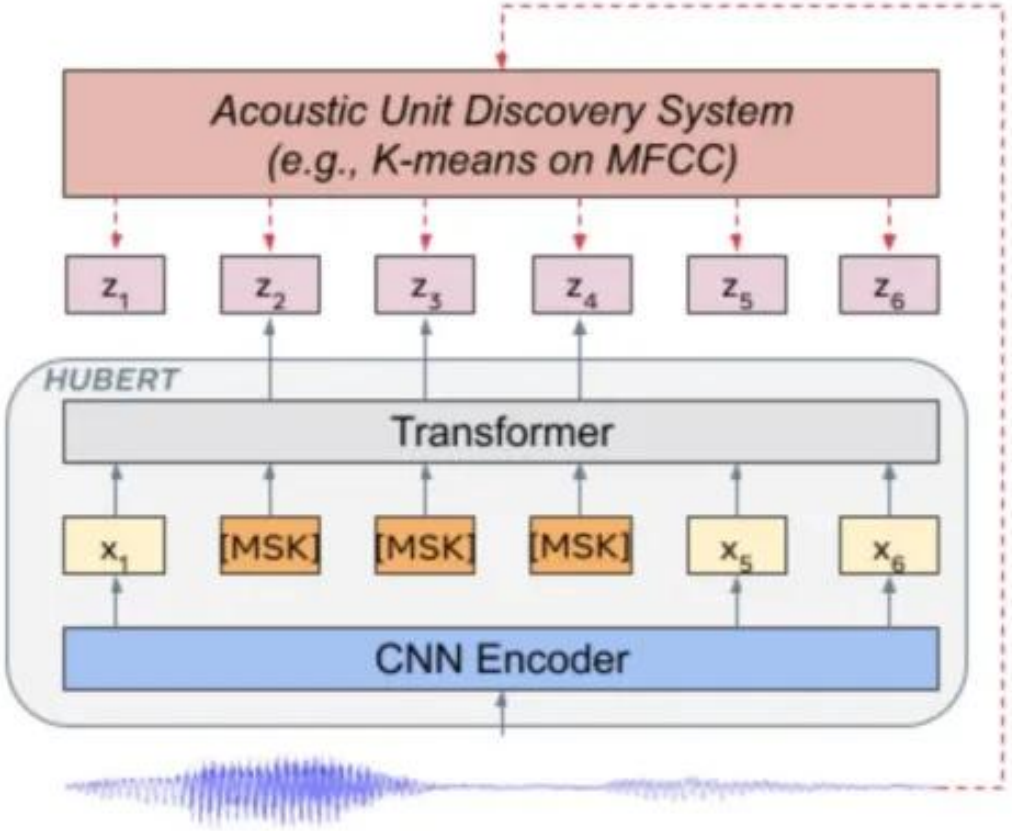
3. 模型简介

3. 模型简介

采用HUBERT模型做语音表示

采用余弦相似度作为距离度量指标

使用多种训练tricks提升模型的稳健性



难点分析

音频分布不均衡

ood

距离度量

解决方案

过采样+batch内采样

鲁棒性增强

余弦相似度

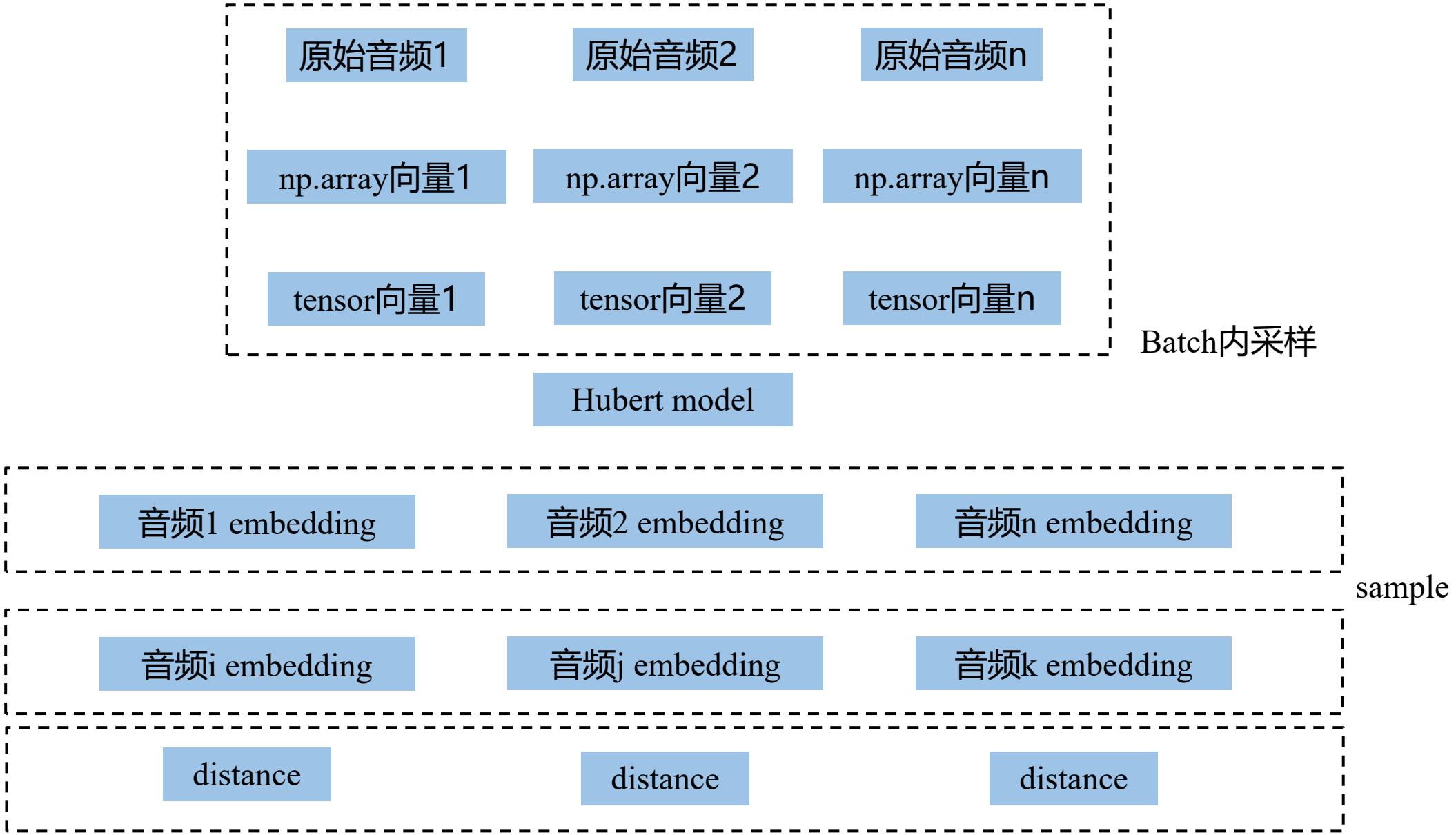
训练tricks

FGM对抗训练

音频截断

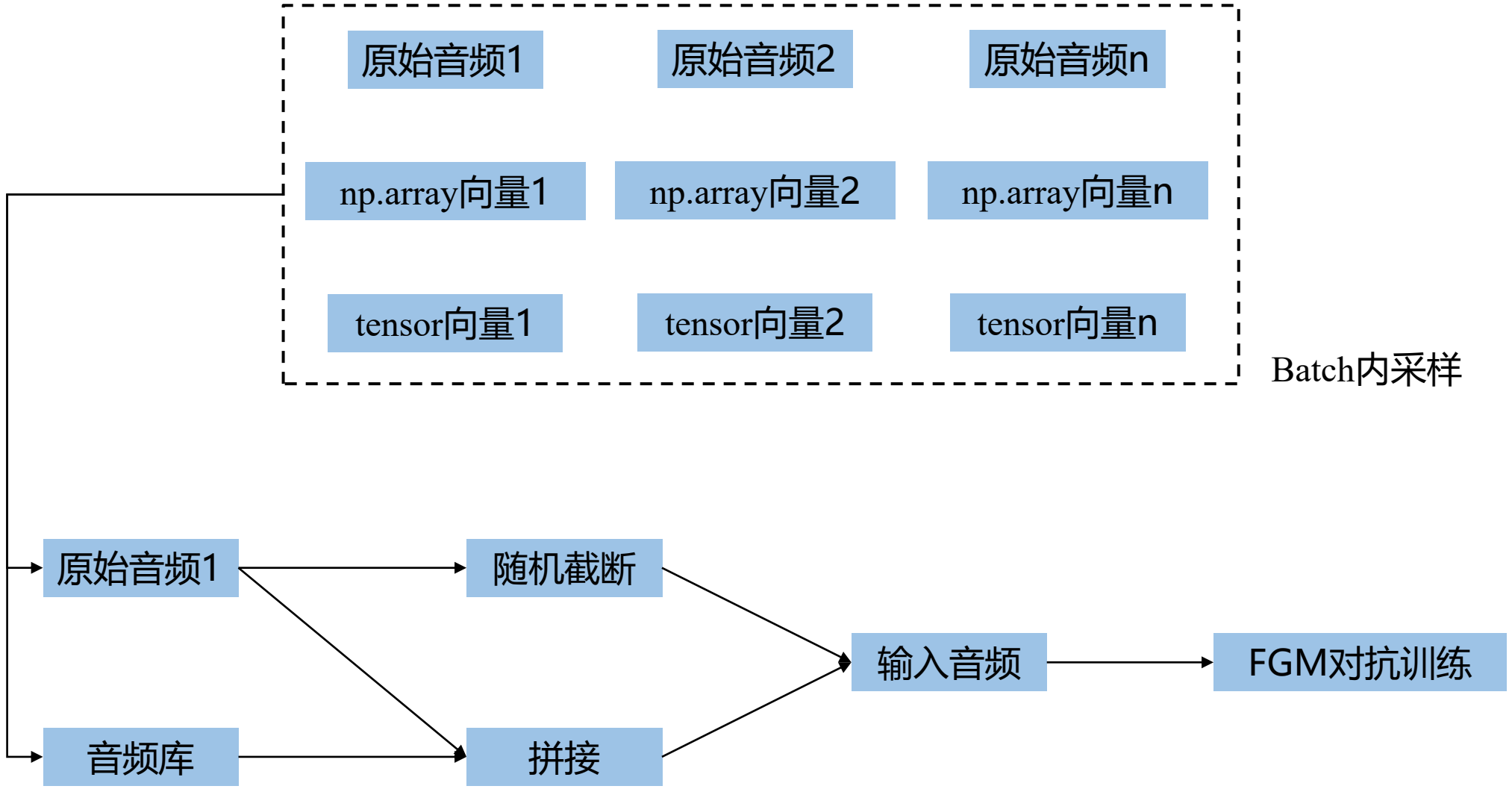
随机替换

3. 模型简介

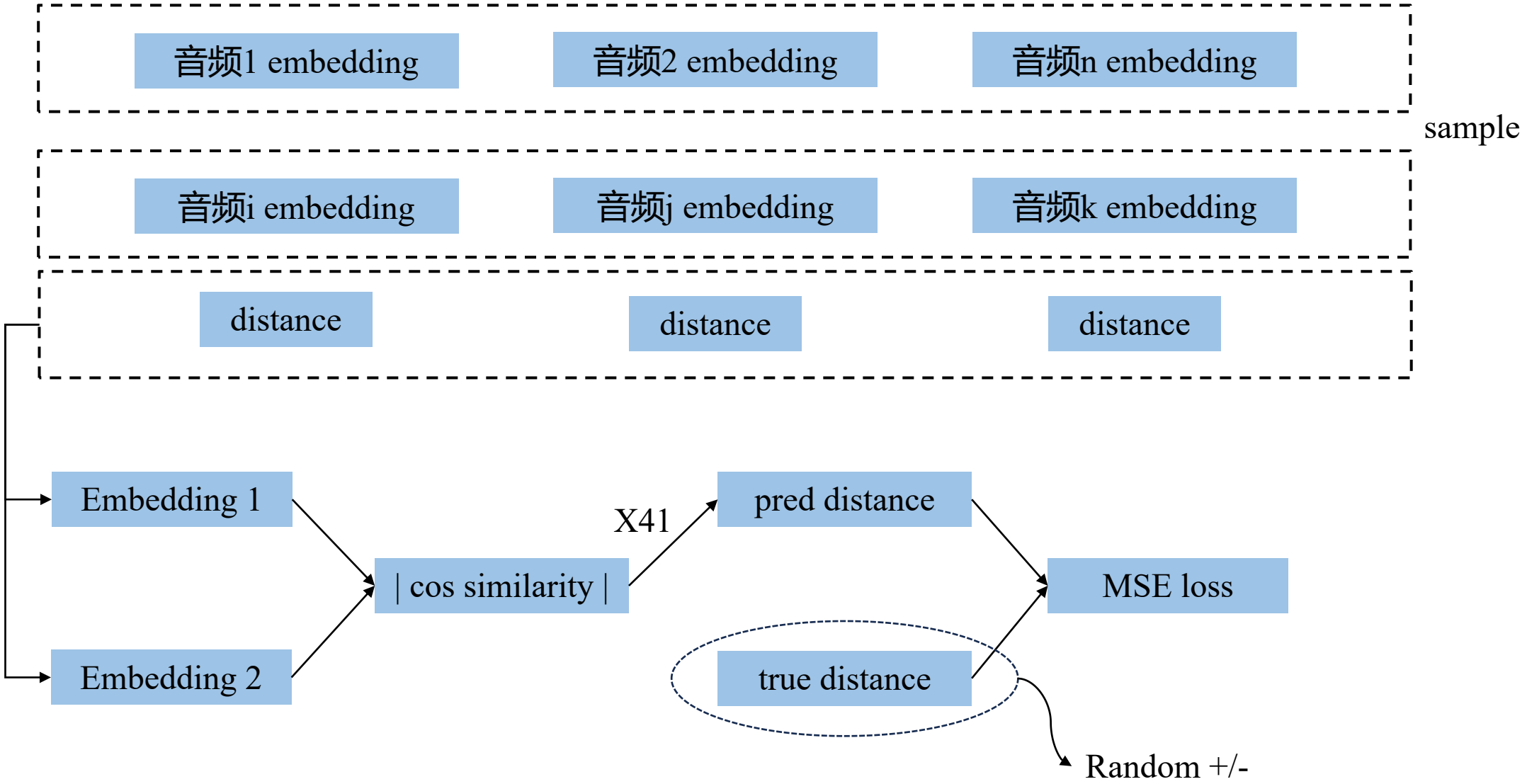


HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Unit

3. 模型简介


原始音频1
原始音频2
原始音频n
np.array向量1
np.array向量2
np.array向量n
tensor向量1
tensor向量2
tensor向量n
Batch内采样
原始音频1
随机截断
音频库
拼接
输入音频
FGM对抗训练

3. 模型简介


音频1 embedding
音频2 embedding
音频n embedding
sample
音频i embedding
音频j embedding
音频k embedding
distance
distance
distance
Embedding 1
Embedding 2
| cos similarity |
X41
pred distance
true distance
MSE loss
Random +/-

4. 模型优化

4. 模型优化

针对模型的优化：

冻结卷积层参数

FGM对抗训练

针对参数的优化：

逐渐衰减的学习率

训练初期距离指标在原有指标基础上随机增减

针对loss的优化：

采用余弦相似度作为距离度量

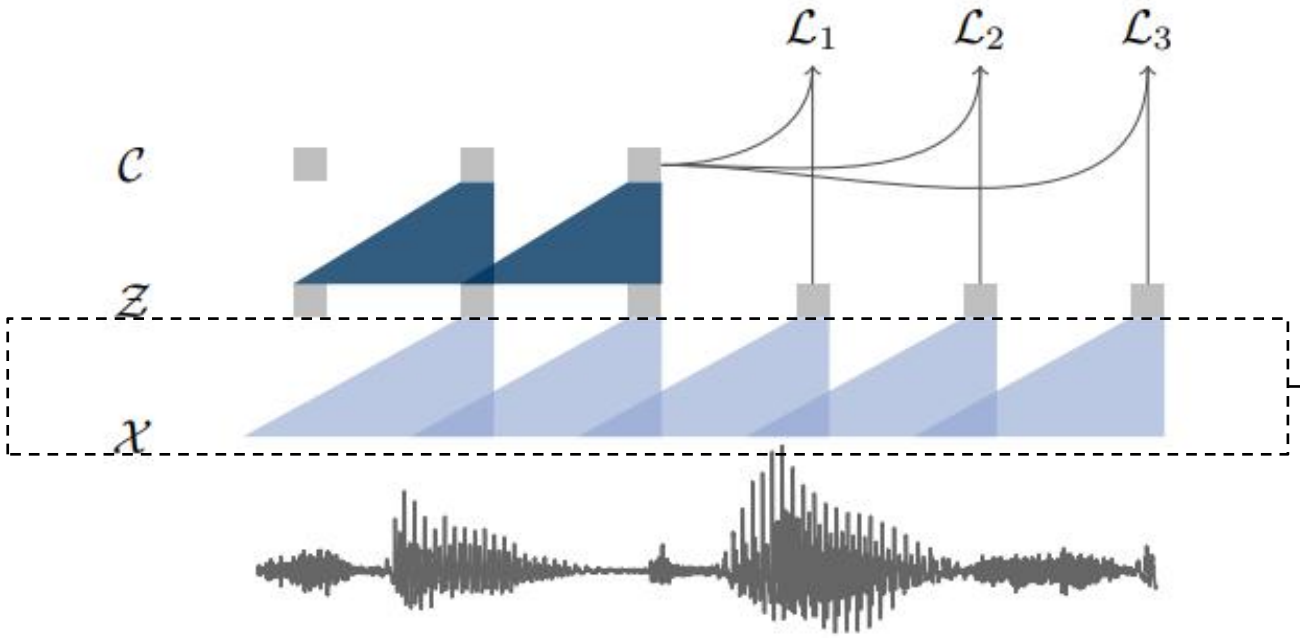
平衡距离为0的样本占比

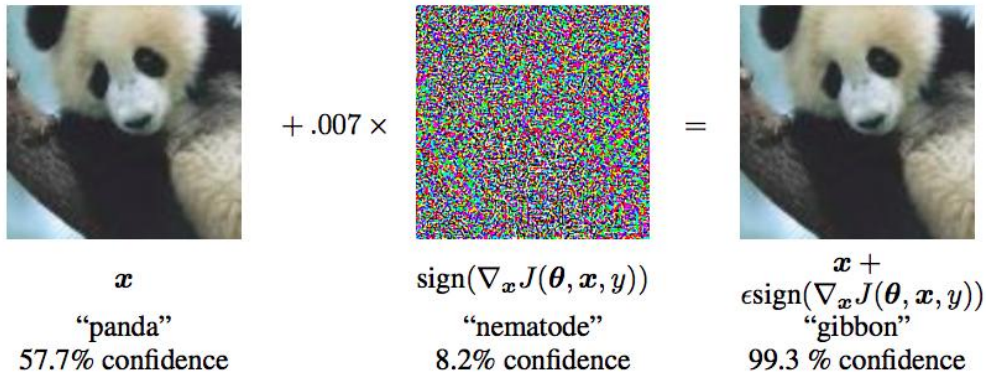
针对采样方式的优化：

在每个batch内做负样本的采样

训练阶段单样本随机截断，推理阶段取两次截断的平均值

4. 模型优化





采用对抗训练（FGM）的思想，在Hubert模型的特征映射层中引入噪声，在尽量不改变原样本的分布，对样本增加扰动，使得模型能够忽视这种扰动，从而提升模型的鲁棒性。

对于每个x:

1.计算x的前向loss、反向传播得到梯度

2.根据feature projection矩阵的梯度计算出r，并加到当前feature projection上，相当于x+r

3.计算x+r的前向loss，反向传播得到对抗的梯度，累加到(1)的梯度上

4.将embedding恢复为(1)时的值

5.根据(3)的梯度对参数进行更新

5. 模型创新点

5. 模型创新点

轻量级

单模，推理速度快，训练时间短

易迁移

迁移能力强，在不同方言上具有鲁棒性

端到端

端到端模型，不需要特别的特征工程

实时性强

推理速度快、能够实现实时更新

模型灵活

鲁棒性强，能适应不同方言的语音

性能稳定

在初赛、复赛都取得了非常好的预测性能

6. 总结与展望

6. 总结与展望

- 给到的音频是8kHz的，Hubert模型是16kHz的，可以对Hubert模型进行8kHz的二次预训练
- 音频长度分布不均衡，可以对音频长度进行截断产生更多的训练样本
- 基于给定的距离进行预测对于集外的距离预测效果不显著，可以采用相似度指标进行处理
- 结合更多音频语音特征上的信息对相似度进行测度
- 更大的模型是否会产生更好的结果?

信也科技杯 FINVOLUTION BIG DATA COMPETITION | 信也 FINVOLUTION

THANK YOU!

团队名称：WeLearnNLP

团队成员：汪超、杨玉舒、李青、王梓聪、李璐

团队单位：同济大学