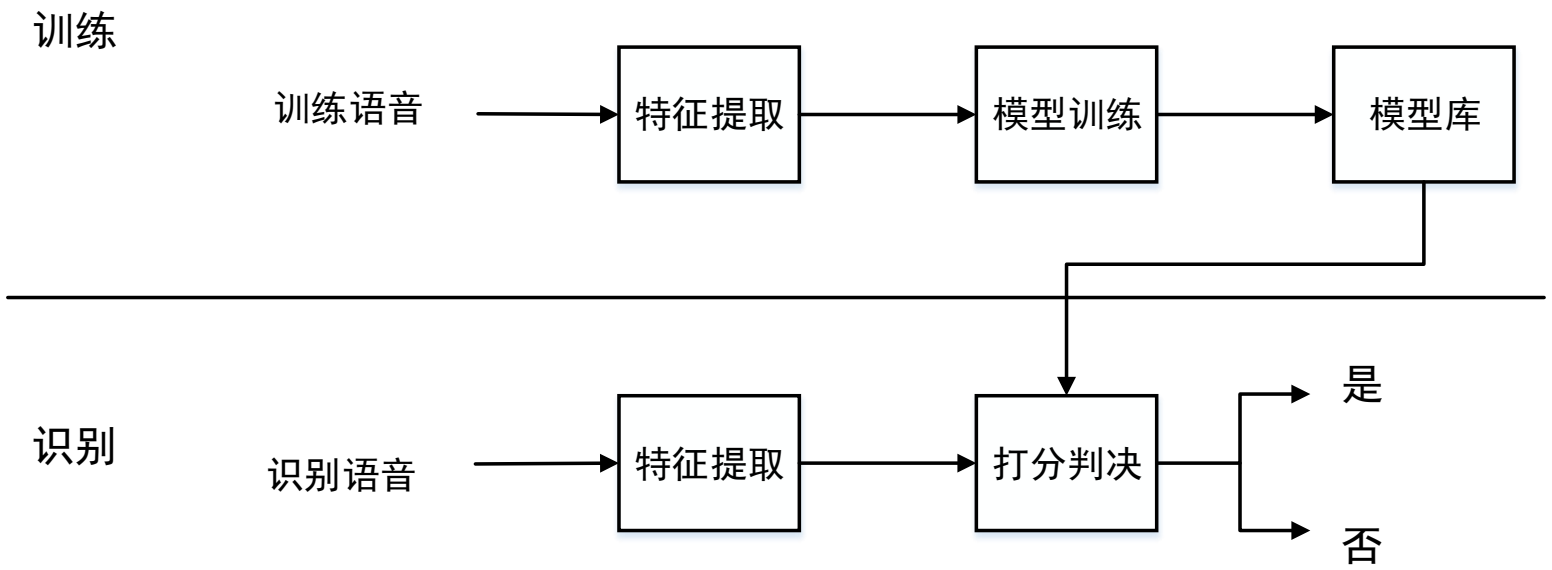
Text-Independent Speaker Verification Based on Triplet Convolutional Neural Network Embeddings

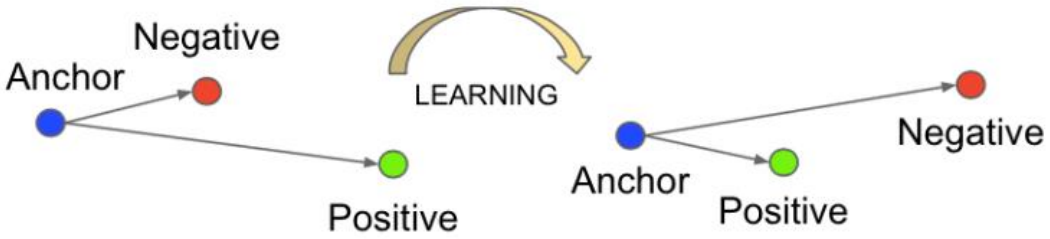
1. 声纹识别



声纹识别系统框架

声纹识别的系统框架如上图所示：整个系统分为训练阶段和识别阶段。在训练阶段，首先对训练语音进行预处理和特征提取，然后构建能代表说话人的模型，将其存放在模型库中。在识别阶段，同样需要对待识别的语音进行预处理、特征提取和模型构建，然后把得到的模型与模型库中的模型进行一对一的相似性比较，判断是否属于同一个人。

2. triplet loss



Loss:

$$\Delta_i = \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha,$$

$$L = \sum_{i=1}^{N} \max(0, \Delta_i), (x_i^a, x_i^p, x_i^n) \in \mathcal{T}$$

Triplet loss: 首先选择一个样本Anchor，Anchor属于身份A，再选择一个样本Positive，同样属于A，然后再选择一个样本Negative，该样本不属于A，triplet loss 的目的就是使属于同一个身份的样本之间的距离尽可能地近，不同身份样本之间的距离尽可能的远，最终得到能够代表一个说话人的embedding向量。

3. 论文提出的方案

该方案主要分为三个阶段。

训练阶段：

（1）. 从数据集中每次选取一个batch的triplet三元组（anchor、positive、negative）

（2）. 对每条语音提取fbank特征或FFT频谱特征，每条语音得到一个二维的特征矩阵。

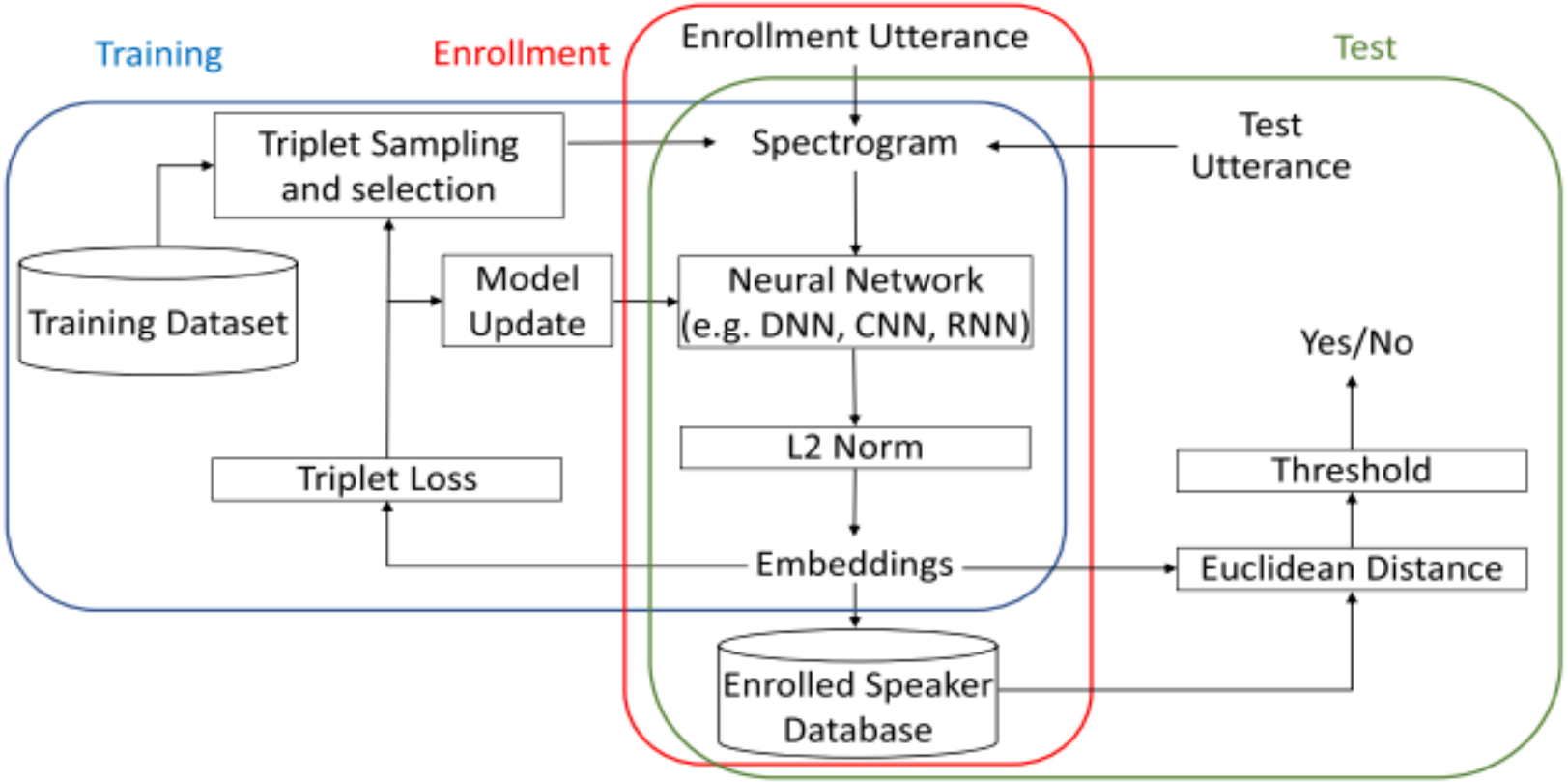
（3）. 把一个三元组的每个特征矩阵分别输入到网络中，得到输出结果，输出为128维向量。

（4）. 对输出的结果进行L2归一化处理。

（5）. 计算一个三元组的triplet loss，并以同样的方式计算出一个batch三元组的triplet loss

（6）.根据triplet loss计算出梯度并更新网络模型的参数。

（7）.回到第一步，直到模型训练完成



3. 论文提出的方案

注册阶段：

（1）. 采集说话人的语音

（2）. 对语音进行有效性检测，去除静音并分成4s的定长语音片段

（3）. 分别提取语音片段的fbank特征或者FFT变换的频谱特征。

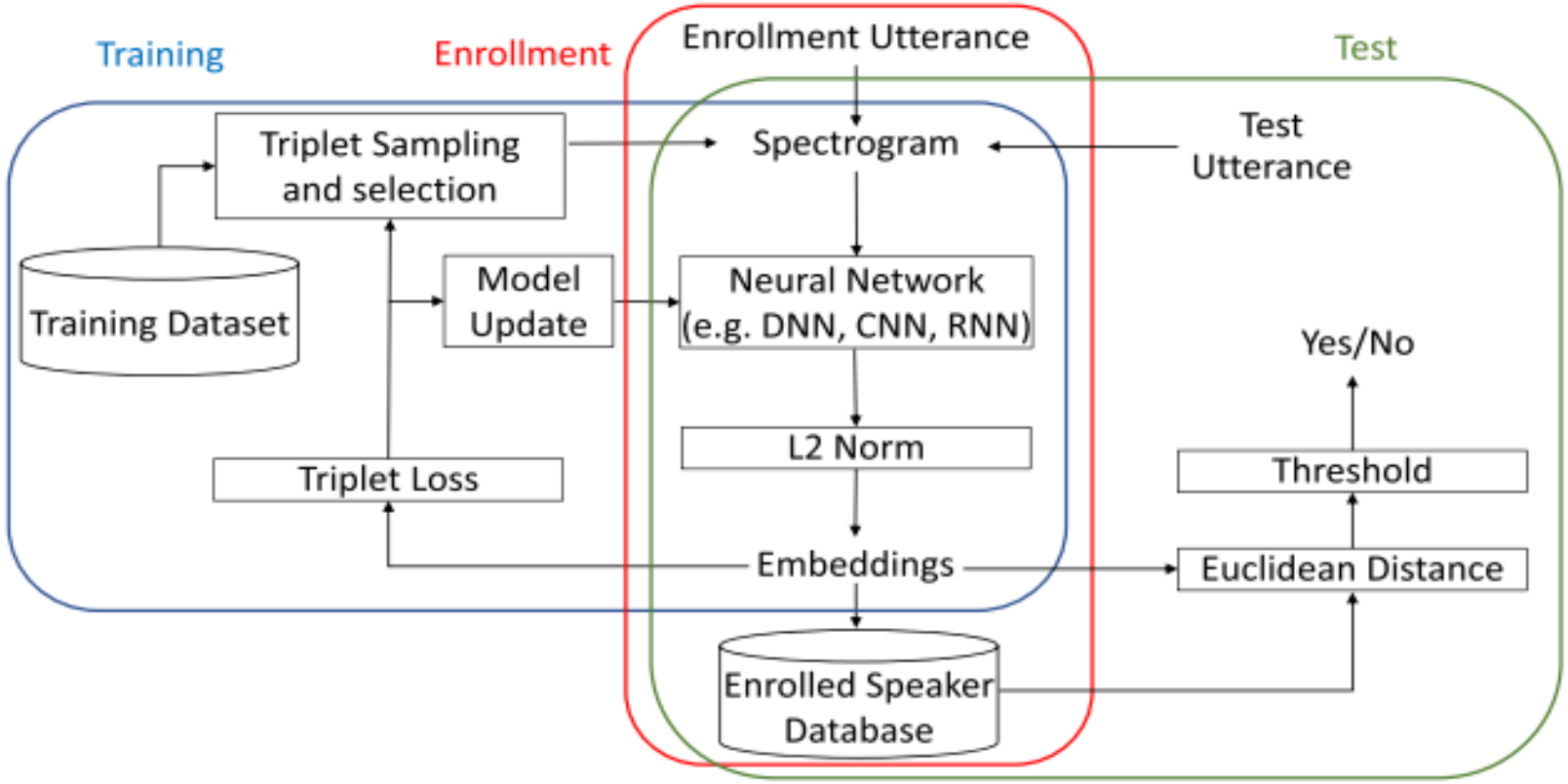
（4）. 把提取的特征矩阵输入到训练好的网络中得到多个128维的向量

（5）. 对输出的向量分别做L2归一化，然后取平均值，再进行L2归一化，得到代表一个说话人的embedding，并存入数据库中。

测试阶段：

（1）.按照与注册阶段相同的过程，得到该语音的embedding

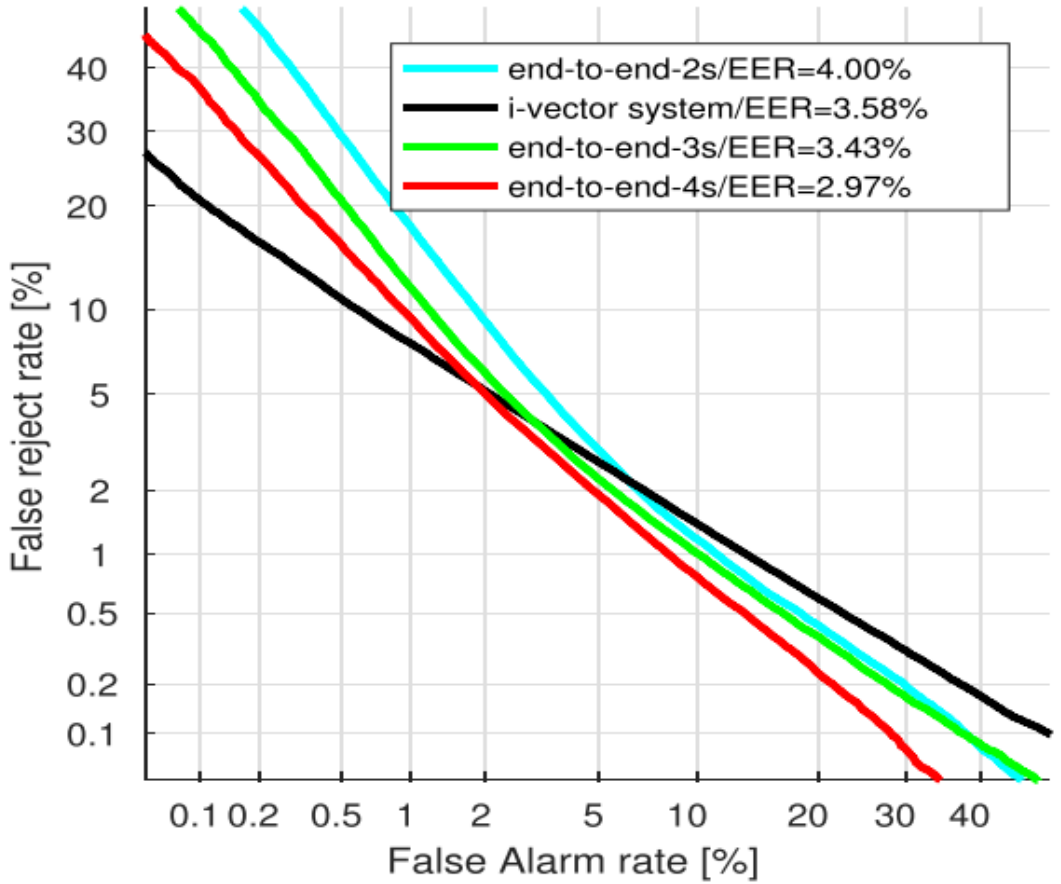
（2）.把该条embedding分别与数据库中的embedding做相似性比较，判断是否属于同一个人



4. 实验结果

i-Vector/PLDA	e2e 4s	e2e variable length
3.58%	2.97%	2.72%

# enroll utts	1	2	5	10
i-Vector/PLDA	3.58%	2.76%	2.03%	1.97%
end-to-end	**2.97%**	**2.41%**	**1.94%**	**1.84%**



谢谢！