目标

在当今智能科技的潮流中，语音和声纹识别技术的重要性愈发凸显。作为人机交互的核心手段之一，它们已经深入到我们生活的方方面面，从智能手机到智能家居，再到金融安全等领域。然而，在嵌入式设备这一领域，如何在有限的计算资源下实现高准确度的语音和声纹识别仍然是一项挑战。因此，本次实践的主旨在于设计一个综合语音和声纹识别功能的模型，以在嵌入式设备上实现计算能力与模型准确度的平衡。我们的目标不仅仅是实现一个具备语音和声纹识别功能的模型，更重要的是在这个过程中寻找一种创新的方法，即如何在保持准确性的同时，最大程度地节约计算资源。在完成这一设定的目标后，我们将致力于对已有的模型进行尽可能的优化。通过对算法、模型结构以及计算流程的优化，我们期望能够进一步提高模型的性能，以适应各种嵌入式设备的需求。

语音识别背景

自动语音识别(Automatic Speech Recognition, ASR)，简称语音识别，是人与人、人与机器顺畅交流的关键技术。随着智能通信设备的蓬勃发展，语音识别技术早已转换成产品，并被广泛应用于会议、客服电话、出行驾驶、教育医疗等各种场景。主流的语音识别技术主要包括：基于机器学习的方法 (如 GMM-HMM)和基于深度学习的方法(如 DNN-HMM)。但由于 GMM-HMM 不仅要求帧内元素之间相互独立，而且没有利用帧间上下文信息，致使模型无法充分刻画声学特征的空间状态分布，识别率较低。而 DNN-HMM 是有监督训练，由于训练数据人工无法标注，所以需要借助 GMM-HMM 来实现帧与状态的对齐，因此 DNN-HMM 模型依然存在一定局限性。在训练架构上，以上两种主流语音识别模型在声学模型、语言模型和发音词典三大组件上都需要单独设计、训练，步骤比较繁琐。而且这种分阶段系统还需 要声学、语言学等专业知识和技术的积累，存在入门门槛高、开发成本高和难维护等问题。近年来，随着计算能力的快速发展，出现了将传统语音识别技术的三大组件融合成一个模型的端到端语音识别技术，实现了语音到文本的直接映射。为解决语音序列和输出序列长度不一致问题，端到端语音识别技术可分为：连接时序分类(Connectionist Temporal Classification，CTC)、循环神经 网络转换器(RNN-Transducer，RNN-T)以及基于注意力机制(Attention)的方法。Wang 等人不仅对这三种模型的发展趋势进行了详细总结，而且深入分析了相关技术的优缺点。随着 Transformer 在机器翻译领域的广泛应用，Dong 等人首次将 Transformer 模型架构引入到语音识别领域，进一步提升了语音识别的准确率。谢旭康等人提出了一种 TCN-Transformer-CTC 模型 ，通过时序卷积(TCN)加强 Transformer 对位置信息的捕捉能力。尽管 Transformer 在捕获长距离上下文信息上具有较大的优势，但提取局部特征的能力较弱。为解决这个问题，Gulati 等人提出了 Conformer 模型，该模型在 Transformer 编码器的基础上加入卷积模块，通过卷积捕获局部细粒度特征，同时保留了 Transformer 的全局表征能力。Burchi 等人提出了一种更为高效的 Conformer 模型,进一步降低了计算复杂度。Gao 等人提出了一种快速并行的 Transformer 模型--Paraformer，将模型的解码速度提升了 10 倍以上。Peng 等人提出了一种 Branchformer 模型，进一步研究了局部特征和全局特征的关系以及对语音识别准确率的影响。Radford 等人提出了一种 Whisper 模型，该模型支持多任务学习，在解码器里通过引入 prefix prompt 来支持任务切换，从而实现多种语言到文本的转换。

声纹识别背景

声纹识别技术的发展总体来说可以分为四个阶段。其第一阶段可以追溯到上个世纪的三四十年代，1945 年，劳伦斯·科斯塔（L.G.Kersta）等人在美国的 Bell 实验室以语谱图为基础展开实验，并且进行相关匹配研究，从而使“声纹”的理论得以萌芽。随着时间的推移，“声纹”的概念也被广泛熟知。

第二阶段发生在 20 世纪 40 年代到 70 年代，在这个阶段初步建立了声纹识别理论体系，声纹识别技术的研究集中在如何从声音中提取出能够准确反映身份信息的特征参数上。线性预测倒谱系数（LPCC）由 BSAtal 提出，这种参数的稳定性较高，可以显著提高该技术的精确性。随着数字信号处理技术的发展，实验人员陆续引入了线性预测编码系数（LPC）和 LSP 谱系数等间接特征参数。1963 年，贝尔实验室的 S.Pruzansky 及其团队开发出一种新型的声纹识别技术，该技术采用模板匹配（Template Matching），并且运用了方差分析的方法，这一创新成果迅速引发了行业的广泛关注。1969 年，Luck JE 在声纹识别中首次使用倒频谱技术，并将说话人确认问题作为二分类问题进行实验，最后获得了不错的识别效果。

第三阶段发生在 20 世纪 70 年代到 80 年代末之间，在此阶段，实验者们致力于深入挖掘数字信息的内涵，有助于更加有效的建立模型，同时也在探索新的模式匹配方法，从而大大提升了模型建立的准确性与效率。1971 年，BS Atal 开始将线性预测倒谱系数（Linear Prediction Cep-strum Coefficient，LPCC）应用在声纹识别领域中，从而大大改善了该技术的精度和可靠性。1972 年，Doddington 及其合作伙伴首次尝试利用共振峰参数来实现声音辨认，从而开创性地推动了一种全新的声音辨认技术。这一年，BS Atal 也进一步发展了基于基音轮廓的声纹识别技术。

第四阶段发生在 20 世纪 90 年代到现在，声纹识别领域的研究人员正在更加关注从声音中提取出关键的特征参数。在 1999 年，Vergin R 和他的团队首次提出了梅尔频率倒谱系数（Mel Frequency Cep-strum Coefficient，MFCC），这一发现为声音识别带来了重大突破。随着互联网技术的不断进步，声纹识别领域引入了许多先进的技术，如动态时间规整（Dynamic Time Warping，DTW）、矢量量化法（Vector Quantization，VQ）、隐马尔可夫（Hidden Markov Model，HMM）、人工神经网络（Artificial Neural Network，ANN）等，进一步提升了声纹识别的准确性。D.Reynolds 等人提出了高斯混合模型（Gaussian Mixture Model，GMM），在各个领域有着广泛的应用。到了 21 世纪，P.Kenny 和 N.Dehak 等人研究出了基于联合因子分析（Joint Factor Analysis，JFA）方法，之后又提出了基于 i-vector 的技术，促进了声纹识别系统性能的提高。在声纹识别领域，深度学习被提出来之前，它们一直是主要方法。

深度学习已成为人工智能领域中的一项重要技术，其对于人工智能的发展和应用有着重要的推动作用。未来，随着人工智能应用的不断拓展和深度学习理论的不断完善，深度学习将在更多的应用场景中发挥其巨大的潜力和价值。2014 年，谷歌公司的 VarianiE 及其团队开发出一种基于 DNN 的声纹识别技术，该技术利用全连接神经网络来提取语音信号的特征，并将每一帧信息输入到深度学习网络中，通过多层处理，可以获得每个说话者的深度说话人嵌入，这些值来自于隐藏层的平均值，这一技术对于基于深度学习的声纹识别具有重要的参考价值。2017 年，x-vector 系统由 D.Snyder 及其研究小组提出，该系统利用时延网络技术，将声学特征映射到池化层，声学特征是帧级别的，池化层是语句级别的，从而学习到语句级的深度说话人嵌入，以实现对说话人的识别。NagraniA 和牛津大学团队研发了 VGGVOX 系统，该系统由孪生网络和对比损失来建立，在研究方面获得了不错的成效。

百度团队研发了 DeepSpeaker 系统，该方法利用深度残差网络和循环神经网络提取语音信号的特征，用平均池化获取说话人的嵌入特征，通过交叉熵损失预训练和三元组损失微调进行声纹识别的训练，最后通过打分融合方法提高声纹识别的准确率。Hu 及其团队研究出了 SE-Net（Squeeze and Excitation Network，SENet），利用 SE 块的概念，可以有效地识别和调整不同的通道，从而改变其中的信息传输模式，并且能够有效地控制信息传输。在 2018 年，谷歌公司的 W.Li 等人引入了与众不同的声纹识别方法，该方法利用 GE2E 损失函数，将最新值和语音库中所有人一一对比，从而加速模型训练速度，同时也提高了准确率。2021 年，肖等人提出在改进 ResNet34 模型上使用 AAM-Softmax 损失函数，针对 ResNet34 作为背景模型时输入语谱图维度过大以及泛化能力差两大问题做出改进。2022 年，荣等人利用知识蒸馏技术来处理 ResNet 的数据，并利用深度学习技术来处理这些数据。这种技术通过用蒸馏损失 MSE 对 ResNet 声纹特征和 I-Vector 的差异进行约束，并且能够更精细地分析数据。经过两种独特的数据增强方式的引入，使得该模型能够更好地抵御外界的噪音干扰，从而大大改善了其稳定性和可靠性。此外，该实验也显示出，此技术能够更好地支持声音识别的需求。

ASR 综述

ASR 任务

ASR 任务表示为：

$$W^* = argmax_w P(W/X) W^* = argmax_w P(X|W) P(W)$$

其中X 为语音提取后的特征向量，$W*$ 为语句输出，P (X/W) 使用声学模型建模，P (W) 使用语言模型建模。

发音字典 Lexicon

考虑对单词/单字建模，英语与中文常用单词/单字的数目十分庞大，并且训练集的规模限制导

致极易出现 OOV(Out Of Vocabulary) 问题。

考虑对音素建模，英语中音素总量约 50 个，这极大的减少了模型数量。

同时，每个音素的声学特征在随发音过程而变化，通常将每个音素划分为 3 状态，即开始，稳

定与结束状态。

问题在于模型如何确定某个单词由哪些音素构成？事实是模型无法确定，这种映射需要引入人

工发音字典。

Acoustic Model 声学模型

输入语音，输出句子，只考虑语音与文字的对应关系，不考虑文字间的概率，即语义语法等关

系。

GMM+HMM (Monophone)

使用 GMM(高斯混合模型) 与 HMM(隐马尔可夫模型) 构成单音素模型，GMM 对每个状态建模，HMM 对每个音素建模。

1.将输入语音切分提取 MFCC 特征帧

2.GMM 模型预测特征帧属于某音素某状态的概率

3.使用 Viterbi 算法计算每个单词的 HMM 生成该序列的概率

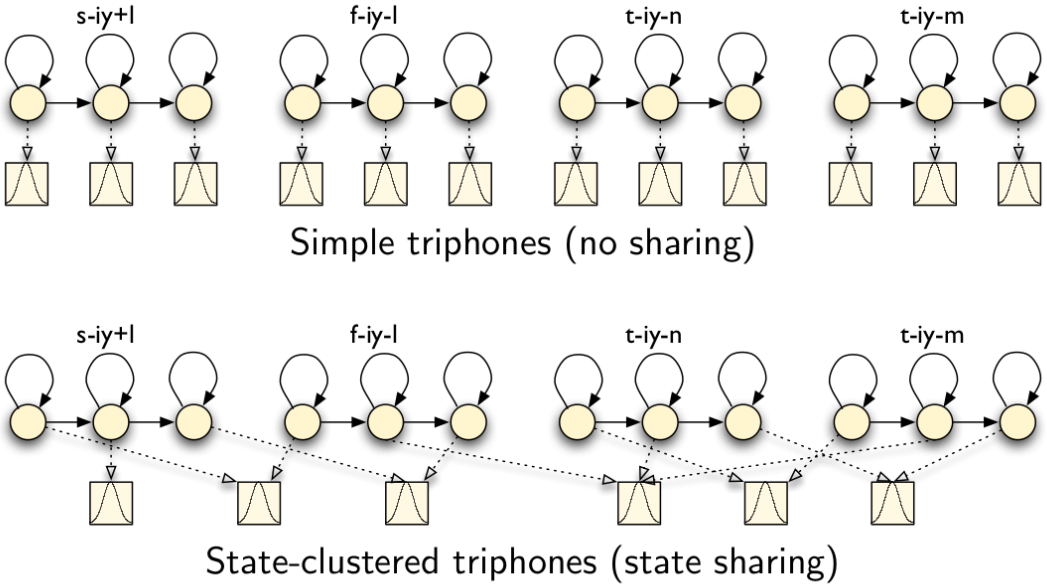
4.选择最大概率的单词作为输出

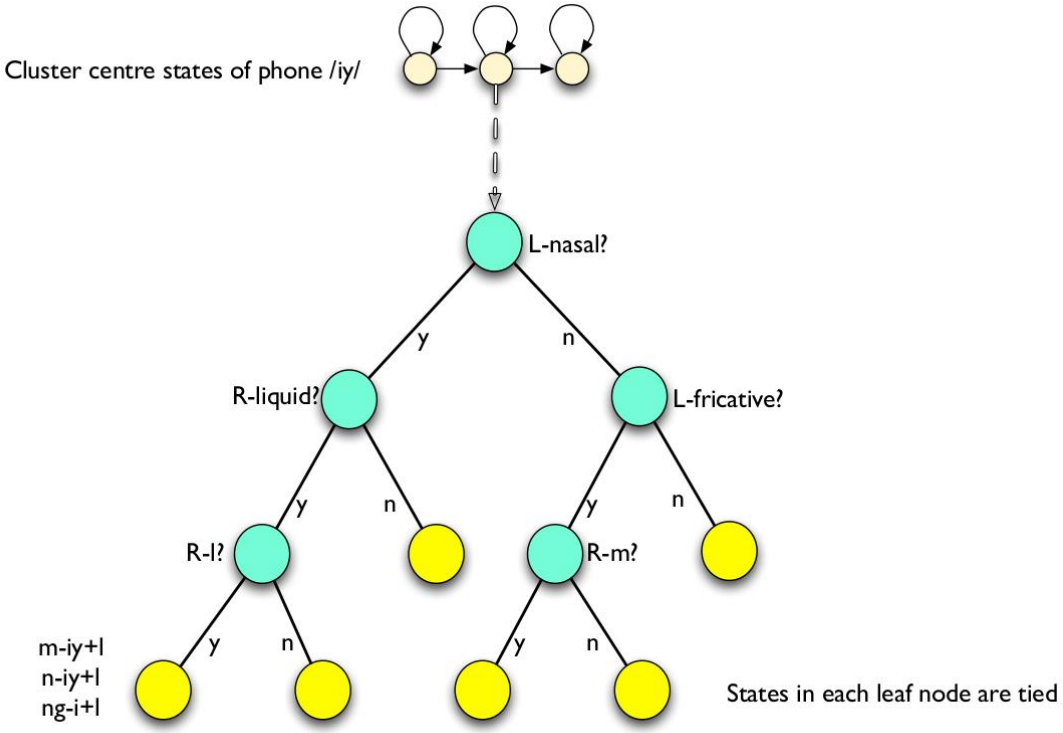
对 N 个单词三状态的语料库，需要N 个 HMM 与$N*3$ 个 GMM，GMM 分量则需要人工设定。

GMM+HMM (Triphone)

Triphone 三音素预测考虑了相邻音素对该音素发音的影响，N 个音素的集合理论上可构成N 3 个 triphone，假设每个 triphone 有三个状态，直接建模需要N 3 $*$ 3 个 GMM，随着状态划分的增长，音素规模的增大，GMM 模型的数量极其庞大。

考虑引入状态共享，即不同音素的状态可以共享同一个 GMM 模型，这种划分通常使用决策树完成。



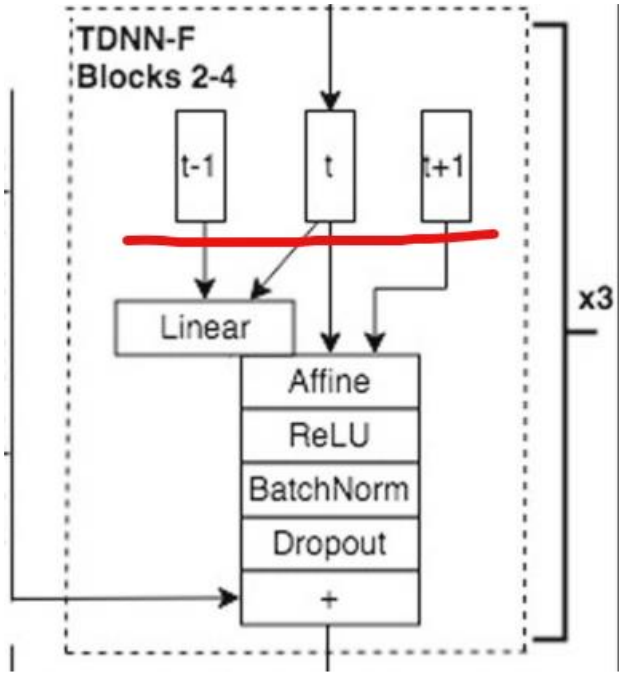


NN+HMM (Monophone)

使用神经网络代替 GMM 输出观测概率。

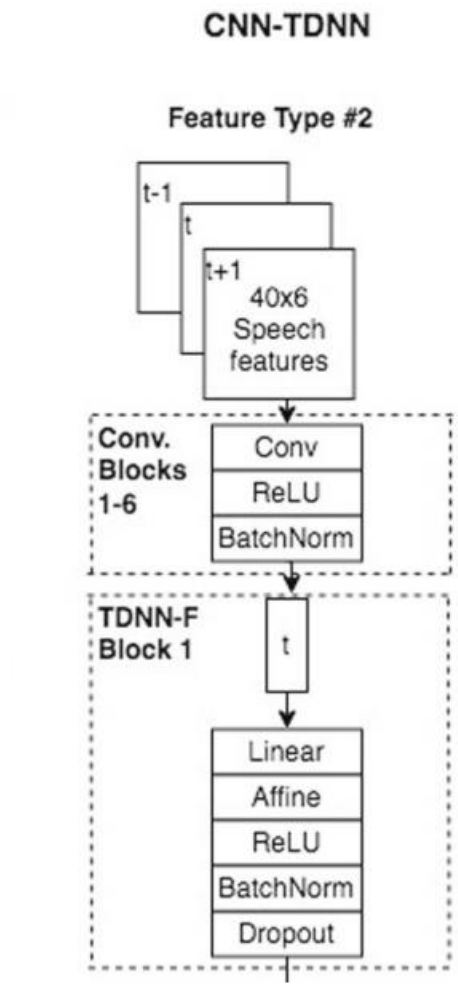
TDNN+HMM (Triphone)

TDNN 引入时序依赖，与 Triphone 相似，可以建模相邻音素$Vt-1, Vt+1$ 之间的影响，同时得益于神经网络权值共享的概念，可以对更长时间依赖的音素$Vt-3, Vt+3$ 进行建模。



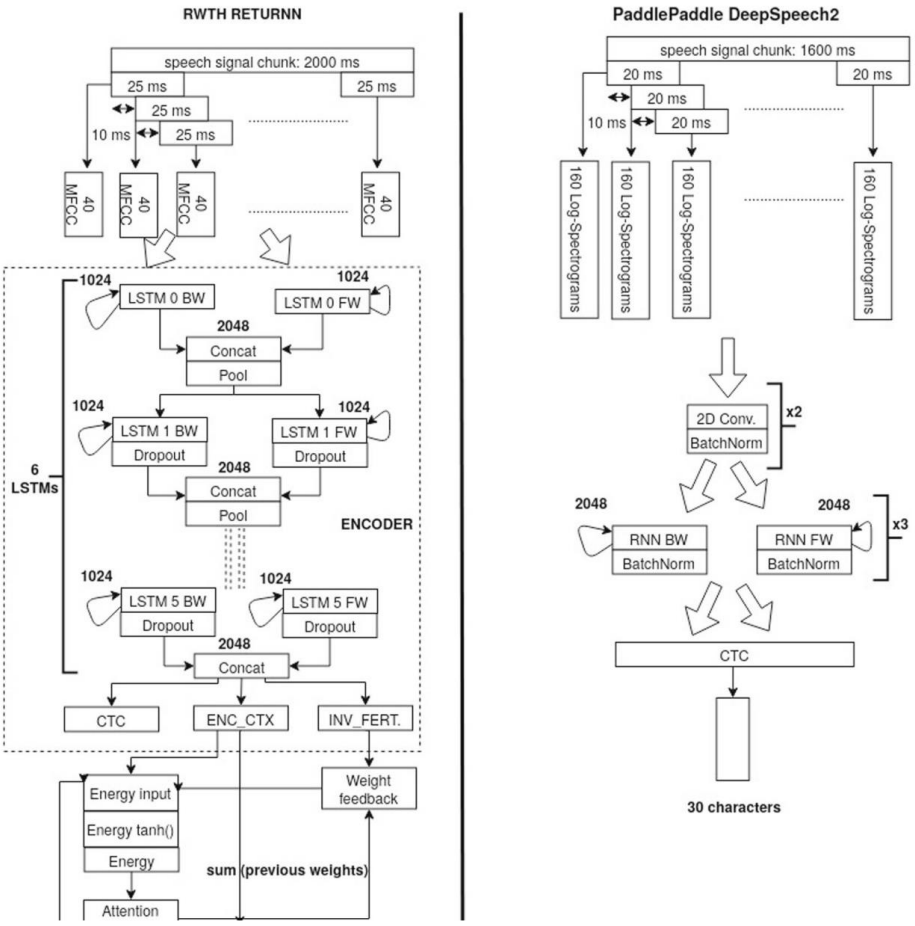
CNN+TDNN+HMM(Triphone)

使用 CNN 权值共享的特征，针对每个输入特征对$Vt-1, Vt, Vt+1$ 进行卷积建模。



NN+CTC(Triphone)

这部分模型使用 RNN 替代 TDNN/CNN 等进行更强的时序建模，拓展时序依赖关系，同时将 HMM 替换为了 CTC 模型，来解决变长序列的建模问题。



CNN

使用纯 CNN 架构, 参数量与计算量均较大。

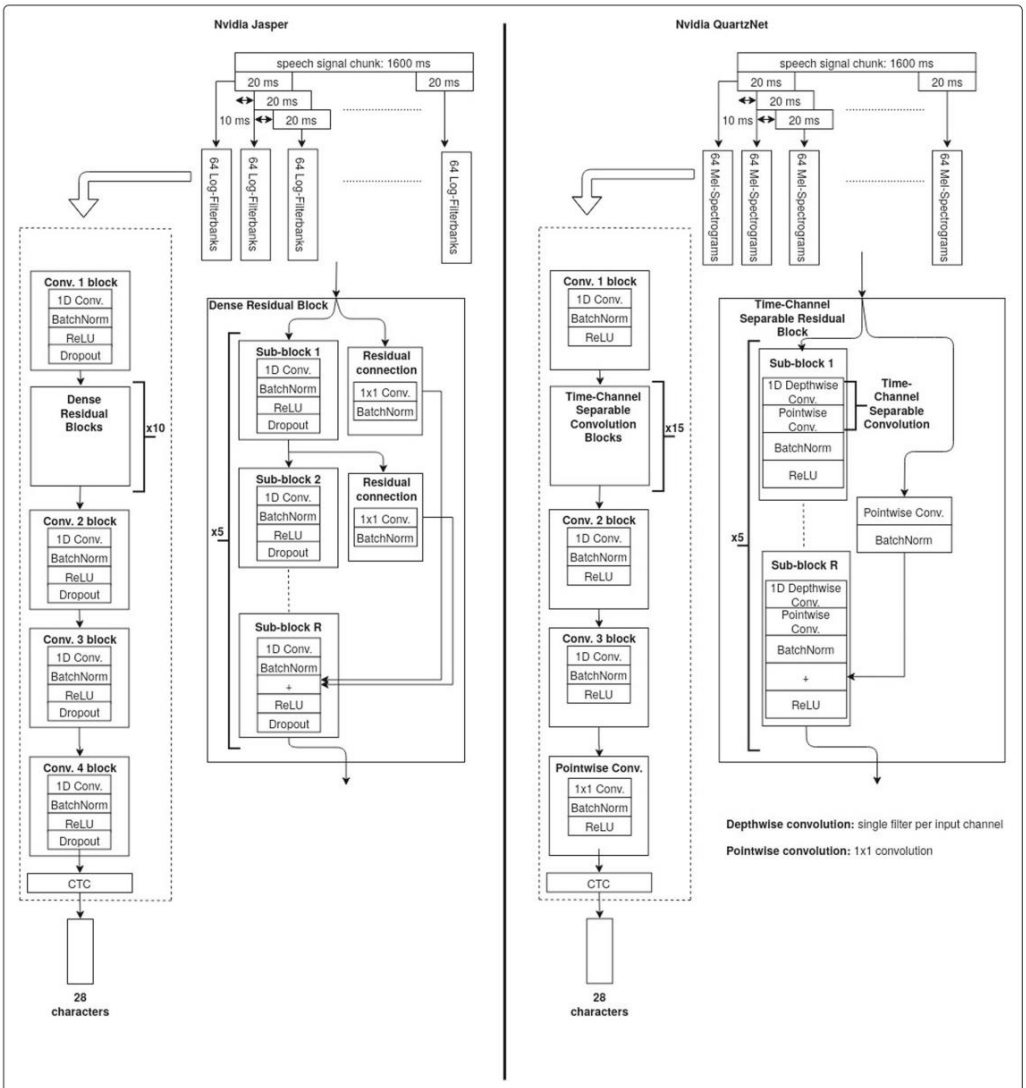
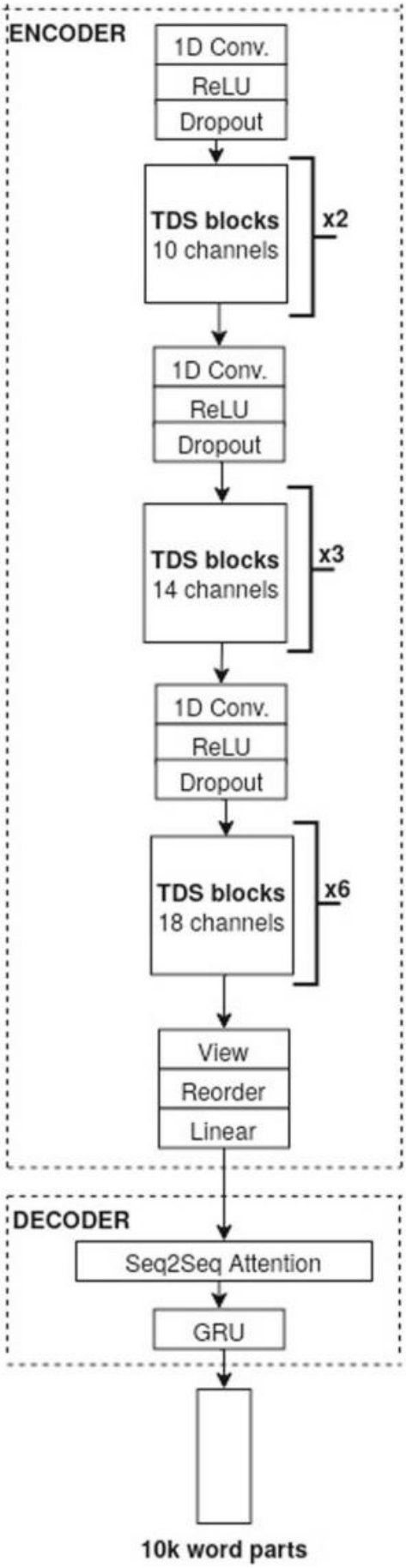


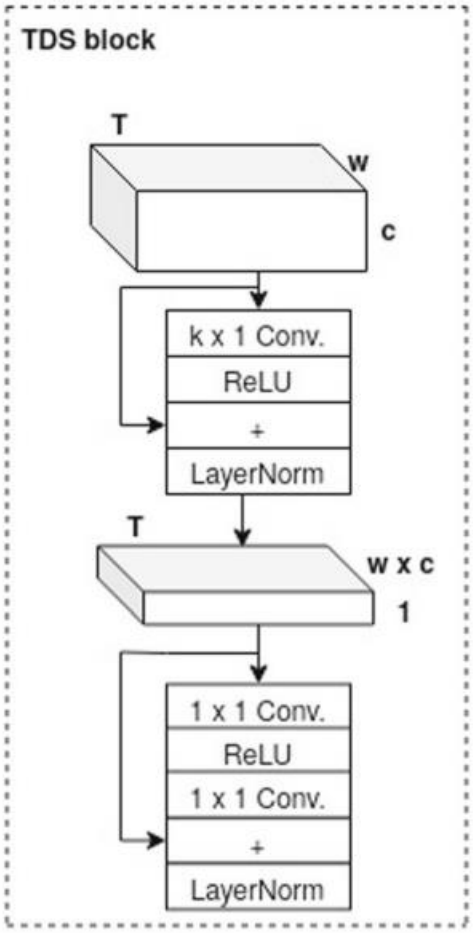
Fig. 6 Nvidia time convolutional networks: Jasper with dense residual blocks (left) and QuartzNet with time-channel separable residual blocks (right)

RNN / RNN

这部分模型往往使用 Encoder/Decoder 结构，RNN 替代 TDNN/CNN 等拓展时序依赖关系进行更强的时序建模作为 Encoder，Decoder 同样使用 RNN。

输入仍然是 MFCC 特征提取后的特征向量。


ENCODER
1D Conv.
ReLU
Dropout
TDS blocks
10 channels
x2
1D Conv.
ReLU
Dropout
TDS blocks
14 channels
x3
1D Conv.
ReLU
Dropout
TDS blocks
18 channels
x6
View
Reorder
Linear
DECODER
Seq2Seq Attention
GRU
10k word parts


TDS block
T
w
c
k x 1 Conv.
ReLU
+
LayerNorm
T
w x c
1
1 x 1 Conv.
ReLU
1 x 1 Conv.
+
LayerNorm

Attention-Based / Transformer
计算量过大，嵌入式可行性存疑。
Language Model 语言模型
建模文字间的依赖关系，语言模型与声学模型相对独立。
N-Gram
引入N 阶马尔可夫假设, 即当前文字先验概率只与前N 个文字概率相关。
使用 N-Gram 计算当前单词序列/句子W = W1...WK 的先验概率为：

$$P(W_1...W_K) = \Pi_{i=1}^{K} P(W_i/W_{i-1}...W_{i-N})$$

N-Gram 基于统计
RNN
Seq2Seq 方式建模条件概率

$$P(W_1...W_K) = \Pi_{j=1}^{K} \Pi_{i=1}^{j} P(W_i/W_{i+1}...W_{i+j})$$

神经网络输出为

$$\{P(W_2/W_1), ..., P(W_K/W_1...W_{K-1})\}$$

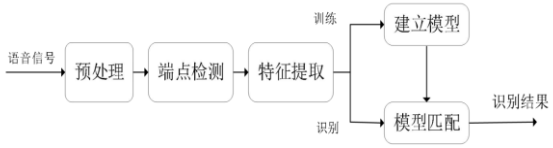
SR 综述

声纹识别任务包括

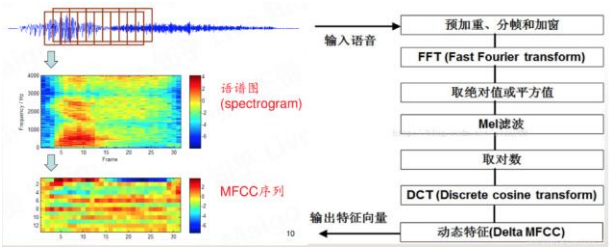
- 说话人辨认：在多人语音中，辨认某一段语音是谁说的；
- 说话人确认：在一段语音和一个人的情况下，确认它是不是他说的；
- 文本无关：对每个人的发音内容没有要求；
- 文本相关：要求训练数据与测试数据的内容相同或前者包含后者；
- 开集辨认：假定待识别说话人可以在集合外；
- 闭集辨认：假定待识别说话人在集合内；

本工程实践的应用实际当属于文本无关的开集说话人辨认任务。

声纹识别的基本流程：

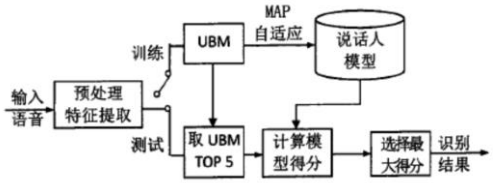


特征提取：
经典方法是 MFCC，还包括语音端点检测、预加重、分帧、加窗和提取声纹模型参数几个步骤。



- GMM_UBM（高斯混合模型-通用背景模型，对单纯 GMM 的改进，2000 年提出）：

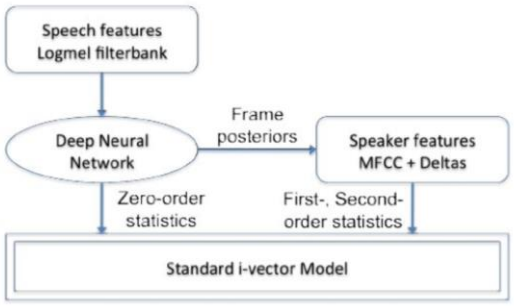
GMM 仍然是常用的简单模型之一，而 UBM 通过最大后验估计(Maximum A Posterior，MAP)的算法对模型参数进行估计，避免过拟合的发生、减少了训练时对说话人语音数据量的需求。



其缺陷在于仍需要较多参数和训练数据、不抗噪声和设备干扰。

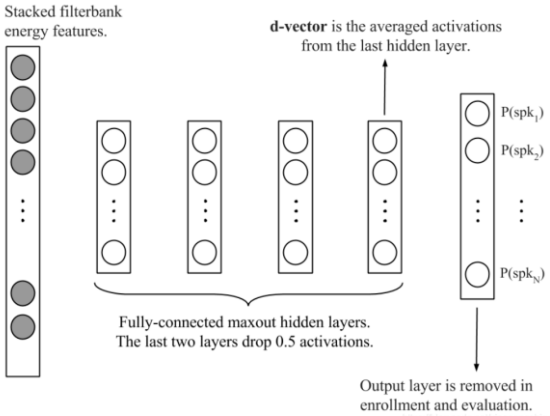
- i-vector 及其延申（2009 年）

将说话人的声纹信息和信道信息映射到低维 i-vector 上，利用其空间方向区分性，使用各种聚类方法实现说话人识别，典型的如 DNN/i-vector、 TVM-i-vector（Total Variability Modeling 全局差异空间建模）



- d-vector （《Deep neural networks for small footprint text-dependent speaker verification》，2014 年提出）

深度网络的特征提取层（隐藏层）输出帧级别的说话人特征，将其以合并平均的方式得到句子级别的表示，这种语篇级的表示即深度说话人向量，简称 d-vector。计算两个 d-vectors 之间的余弦距离，得到判决打分。类似主流的概率统计模型 i-vector，可以通过引入一些正则化方法 (线性判别分析 LDA、概率线性判别分析 PLDA 等)，以提高 d-vector 的说话人区分性。



- x-vector Embedding（《X-VECTORS: ROBUST DNN EMBEDDINGS FOR SPEAKER RECOGNITION》，2018 年提出）
  使用数据增广来提高深度神经网络（DNN）embedding 对于说话人识别的性能。经过训练以区分说话者的 DNN 将可变长度的语料映射到我们称为 x-vector 的固定维度 embedding。

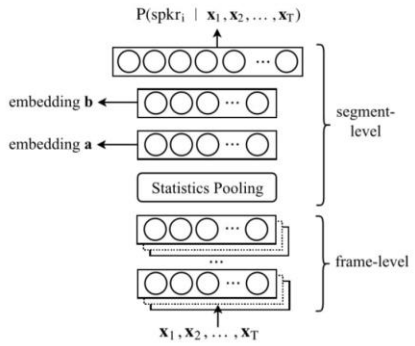


Figure 1: *Diagram of the DNN. Segment-level embeddings (e.g.,* **a** *or* **b***) can be extracted from any layer of the network after the statistics pooling layer.*

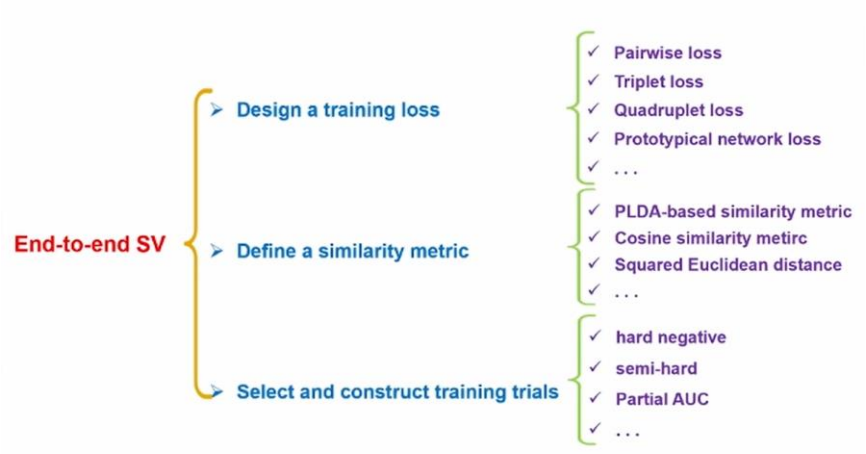
有点事有效且训练稳定，缺点是

训练标签需要精确到每句话对应的说话人身份；

使用 Softmax 不一定是最优的损失函数；

输出层随着说话人数量增大而变大，因而不适应与变化较多的开集任务；

End-to-end（2019 年至今的多种模型，关键点集中在损失函数确认、定义相似度指标、选择和构建训练方法）



有学者提出了基于 LSTM 神经网络的说话人识别模型，对比三元组、TE2E、GE2E 三个损失函数，起的了较好的准确率。

总结：

非端到端的分类损失需要引入减小类内方差的 margin

端到端确认损失引入类中心学习可以增加训练稳定性、提高性能

实践进度

已完成了哪些任务。。。