深度说话人表征学习

理论、应用与实践

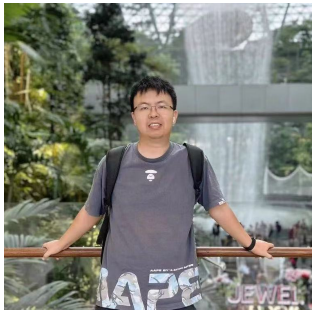
王帅

2025 年 10 月

南京大学 NANJING UNIVERSITY

智能科学与技术学院
School of Intelligence Science and Technology

关于讲者



王帅（副教授，特聘研究员）
南京大学智能科学与技术学院

- 研究方向：说话人建模，语音 & 音乐生成
- 开源项目：WeSpeaker, WeSep, West, DiffRhythm, SongBloom ...
- Email: shuaiwang@nju.edu.cn
- 个人主页：https://shuaiwang-nju.github.io/

目录

什么是说话人建模？

定义

说话人建模旨在通过分析语音信号中嵌入的语音模式来表征和识别个体的独特特征。

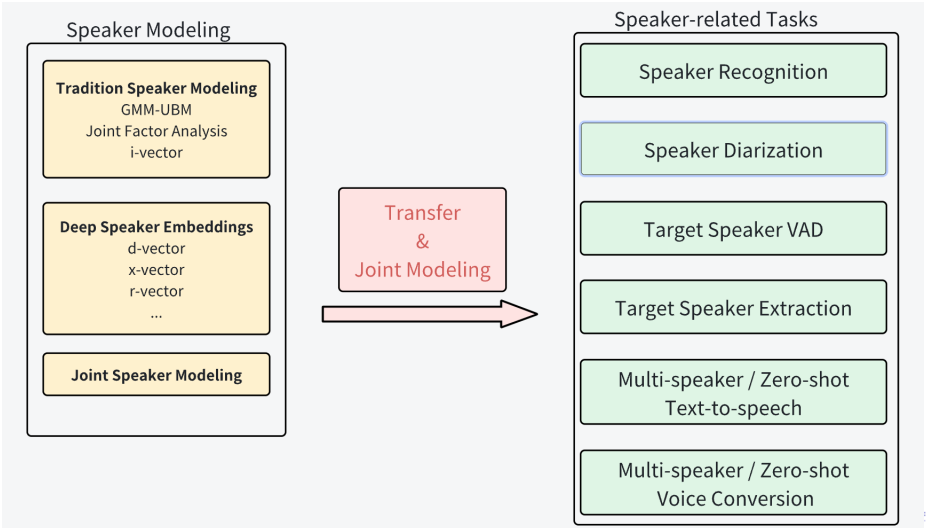
主要应用：

- 说话人识别
- 说话人分离
- 语音克隆
- 语音合成
- 目标说话人提取

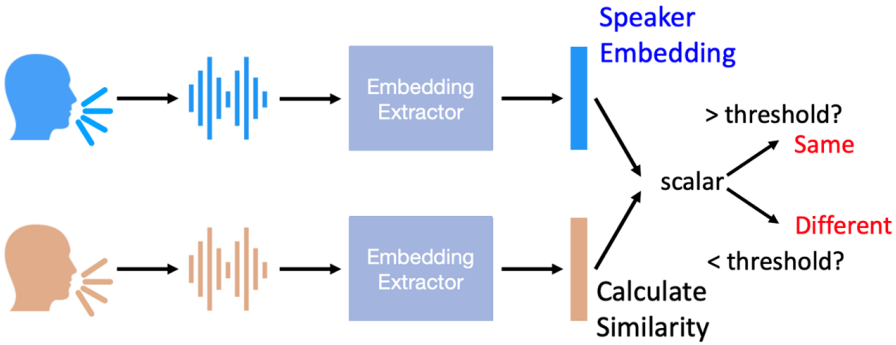
实际场景：

- 生物识别认证
- 监控系统
- 个性化服务
- 法医分析
- 隐私保护

说话人建模的应用

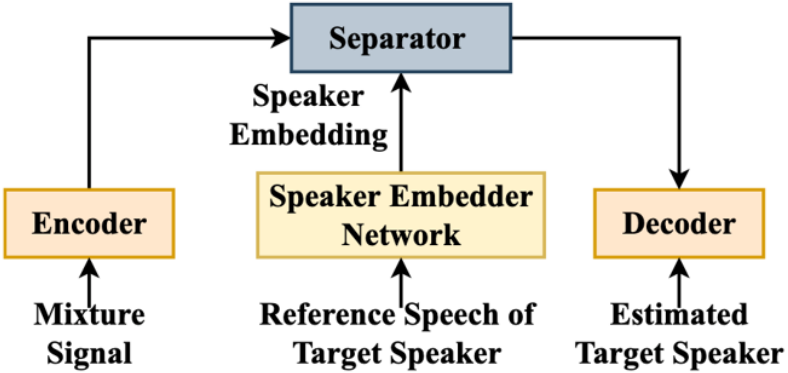

Speaker Modeling
Tradition Speaker Modeling
GMM-UBM
Joint Factor Analysis
i-vector
Deep Speaker Embeddings
d-vector
x-vector
r-vector
...
Joint Speaker Modeling
Transfer
&
Joint Modeling
Speaker-related Tasks
Speaker Recognition
Speaker Diarization
Target Speaker VAD
Target Speaker Extraction
Multi-speaker / Zero-shot
Text-to-speech
Multi-speaker / Zero-shot
Voice Conversion

说话人验证：声音即密码

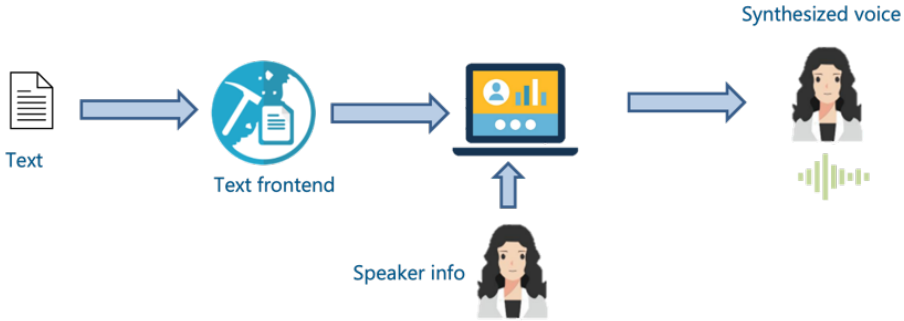


图片改编自李宏毅的 DLHLP20 幻灯片[1]

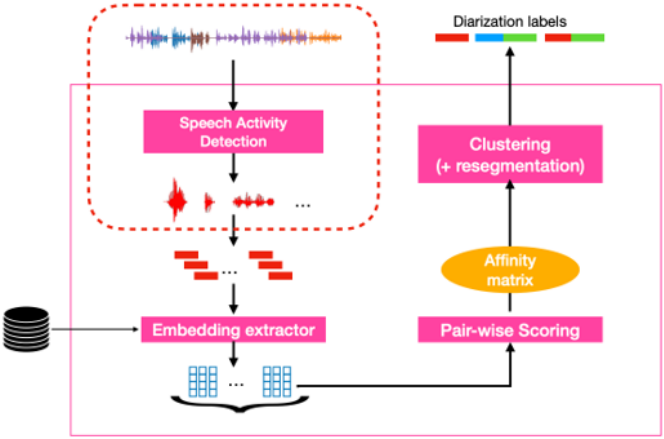
目标语音提取：倾听目标人的声音



目标语音的说话人建模。



说话人日志：谁在什么时候说话？



历史演进：范式转变

1. 从 **VQ** 到 **GMM** (1980s-1990s)
   - 向量量化 → 高斯混合模型
   - 关键进展：用协方差矩阵进行不确定性建模
2. 从 **GMM-EM** 到 **GMM-UBM** (2000s)
   - 通用背景模型，提高泛化能力
   - 用 MAP 自适应替代 EM 训练
3. 从超向量到 **i-vector** (2010s)
   - 降维和信道补偿
   - 低维说话人表征
4. 从生成式到判别式 (2015-至今)
   - 深度神经网络用于说话人嵌入
   - 端到端判别式训练

时代 1：从 VQ 到 GMM (1980s–1990s)

方法和表征

- 向量量化 (VQ)：离散码本表征；基于距离的匹配。
- 高斯混合模型 (GMM)：连续密度生成式建模，通过 EM 训练。
- 基于似然的评分替代启发式距离；更好地拟合声学特征分布。

VQ 的局限性

- 没有明确的不确定性建模；在信道和噪声变化下脆弱。
- 固定、手工设计的码本；容量和自适应能力有限。

为什么 GMM 占主导

- 软分配和协方差建模捕获说话人内部变异性。
- 有原则的最大似然训练；可扩展到自适应和补偿。

时代 2：从 GMM-EM 到 GMM-UBM (2000s)

关键创新

- 通用背景模型 (UBM)：所有说话人共享的说话人无关声学空间。
- MAP 自适应：使用有限的说话人数据从 UBM 推导说话人模型（相关性因子控制）。
- 似然比评分：$\log p(\mathbf{X} \mid \text{spk}) - \log p(\mathbf{X} \mid \text{UBM})$ 用于验证。

实际影响

- 在稀缺注册数据下的鲁棒性；改进跨信道泛化。
- 为大规模评估建立了可重现、可扩展的基线（电话、麦克风语音）。

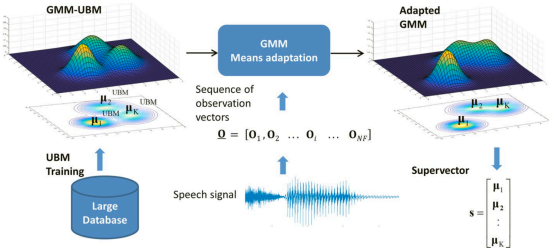


图: GMM-UBM[a] 说话人建模模型

高斯混合模型 (GMM)[a]

- $p(\mathbf{x}) = \sum_1^K c_k \mathcal{N}(\mathbf{x} \mid \_k, \_k)$ s.t. $\sum_1^K c_k = 1$
- 任何分布都可以通过几个高斯分布的加权线性组合来近似
- 当使用 GMM 对说话人的声学特征建模时，高斯数量可以视为产生的声音类型

通用背景模型 (UBM)

- 通常，一个人的注册语音是有限的（几秒钟），很难用这些数据训练 GMM
- 可以先在大规模数据集上训练 UBM，然后适应特定说话人的数据

GMM-超向量

- 连接每个高斯的均值向量来表示说话人

[a] Zheng, Zhang, and Xu, "Text-independent speaker identification using gmm-ubm and frame level likelihood normalization".

[a] Reynolds et al., "Gaussian mixture models."

基于高斯混合模型的说话人建模-测试评分

- 在 GMM-UBM 系统中，通常使用似然比来计算测试音频的分数。给定一段音频 Y，有如下两个基本假设：

$$H_0 : Y\text{来自目标说话人}S$$
$$H_1 : Y\text{不是来自目标说话人}S$$

- 那么评分 Λ 由以下对数似然比决定：

$$\Lambda = \frac{1}{T}\log\frac{p(Y \mid H_0)}{p(Y \mid H_1)} = \begin{cases} \geq \theta & \text{接受}H_0 \\ < \theta & \text{接受}H_1 \end{cases}$$

其中 T 为该段测试语音的总帧数，θ 为预先设好的阈值

- 在具体计算时，$p(Y \mid H_0)$ 为语音 Y 的特征向量在说话人 S 的 GMM 上的概率密度，$p(Y \mid H_1)$ 为语音 Y 的特征向量在冒认者模型上的概率密度。在 GMM-UBM 系统中，使用 UBM 作为冒认者模型来计算测试语音不属于目标说话人的概率。

时代 3：从超向量到 i-vector (2010s)

从高维到低维空间

- 超向量 SVM：将自适应的 GMM 均值连接成非常高维的向量；强大但对信道敏感。
- 总变异性 (T) 模型：联合因子分析；话语级潜在变量 (i-vector) 总结说话人和信道。
- i-vector 提取：在 $\mathcal{N}(\mathbf{m}+\mathbf{T}\mathbf{w},\Sigma)$ 中的后验推理，产生紧凑的 $\mathbf{w}\in\mathbb{R}^d$。

后端和补偿

- 长度归一化，LDA/WCCN 用于类间/类内方差控制。
- PLDA 或余弦评分用于校准验证。

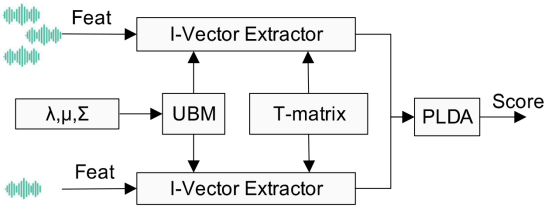


图: 基于 i-vector 的说话人识别系统框图

GMM-超向量的缺点

- 超向量是极高维的（通常数万维），使其在计算上具有挑战性

将超向量分解为低维 i-vector[a]:

$$M(s) = m + Tw(s)$$

- $M(s)$: 说话人 s 的 GMM-超向量
- m: 说话人无关超向量
- T: 总变异性矩阵，捕获语音样本中所有变异性来源（说话人相关和信道相关）
- $w(s)$: 说话人 s 的 i-vector

[a] Dehak et al., "Front-end factor analysis for speaker verification".

i-vector 提取过程

1. **UBM** 训练：使用大量数据训练通用背景模型
2. 总变异性矩阵训练：学习 $\mathbf{T}$ 矩阵
3. 后验推理：计算 $\mathbf{w} = E[\mathbf{w}|\mathbf{X}]$
4. 长度归一化：$\mathbf{w}_{norm} = \frac{\mathbf{w}}{||\mathbf{w}||}$

后端补偿技术

- **LDA**：线性判别分析
  - 最大化类间方差
  - 最小化类内方差

评分方法

- **PLDA**：概率线性判别分析
  - 考虑说话人内变异性
  - 提供概率解释

时代 4：从生成式到判别式 (2014–至今)

神经嵌入和训练

- x-vector/TDNN，ResNet/ECAPA 编码器；注意力统计池化用于时间聚合。
- 大边距分类损失 (AM-Softmax，AAM-Softmax) 用于判别式说话人空间。
- 数据增强和域自适应用于鲁棒性（混响、噪声、编解码器、信道）。

趋势和性能

- 自监督预训练（如 wav2vec 2.0，HuBERT）作为前端或端到端微调。
- 简单的余弦/PLDA 后端在良好校准的嵌入下保持竞争力。
- 在开放基准测试（如 VoxCeleb）上 EER/minDCF 在约束评分下持续提升。

说话人表征学习：
给定语音话语 $\mathbf{O} = \{\mathbf{o}_1, \cdots, \mathbf{o}_T\} \in \mathbb{R}^{T \times D}$，学习映射函数 $\mathcal{F}$ 提取说话人表征：

$$\mathbf{v} = \mathcal{F}(\mathbf{O}) \in \mathbb{R}^d$$

其中：

- $\mathbf{o}_t$：时间 t 的帧级声学特征
- T：帧数
- D：特征维度
- $\mathbf{v}$：固定长度说话人嵌入
- d：嵌入维度

目标：相同说话人的嵌入应该接近，不同说话人应该远离。

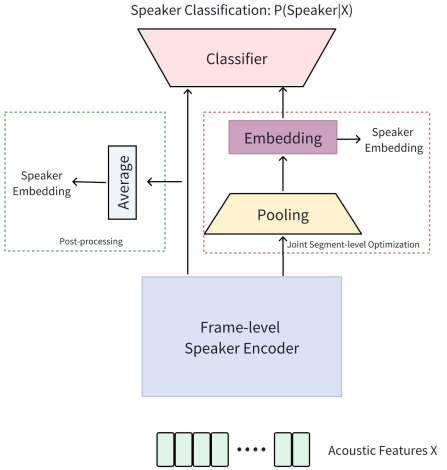
帧级 **(d-vector)**

- 在帧级标签上训练
- 训练后聚合
- 简单但性能有限

$$\mathbf{v} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t \quad (1)$$

段级 **(x-vector)**

- 在话语级标签上训练
- 在网络中集成聚合
- 更好的性能



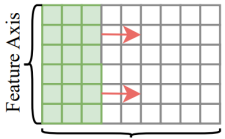
1D 卷积

- 沿时间维度应用
- 计算成本较低
- 可能的大感受野
- 简单架构
- 频率建模有限

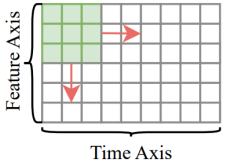
2D 卷积

- 沿时间和频率应用
- 更好的时频建模
- 更高的计算成本
- 更好的性能潜力

a) 1D Convolution for raw wav input



b) 1D Convolution for spectrogram input



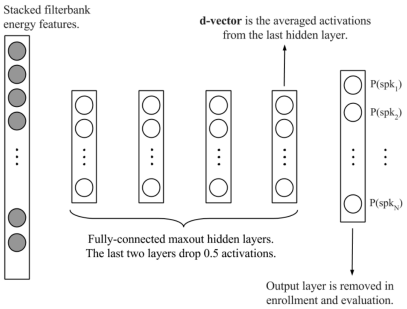


图: d-vector 架构

d-vector[a] 的标签:

- 将深度神经网络应用于说话人信息建模的早期尝试
- 展示了与 i-vector 的良好互补性
- 帧级嵌入平均为话语向量；用说话人 ID 的 CE 训练
- 典型编码器：TDNN/LSTM/CNN 与时间平均池化（无注意力统计）
- 优点：简单训练流程，适用于短话语，低延迟
- 局限性：音素内容泄漏；段级聚合比 x-vector 弱

[a] Variani et al., "Deep neural networks for small footprint text-dependent speaker verification".

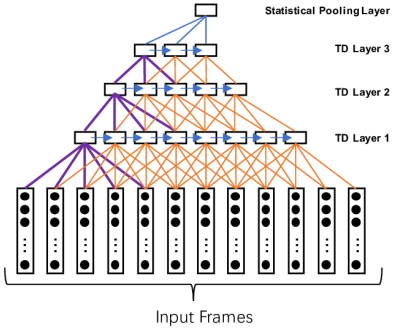


图: x-vector 采用的 TDNN 架构

x-vector[a] 的标签:

- 第一个在公认数据集 (NIST SRE) 上击败传统方法的深度说话人嵌入
- 第一个引入段级优化的工作
- 强大的变体 ECAPA-TDNN[b]

时间池化:

$$= \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t \quad (\mathbf{h}_t \in \mathbb{R}^D)$$

统计池化:

$$= \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\mathbf{h}_t -)^{\odot 2}}$$

$$\mathbf{v} = [\ ;\ ] \in \mathbb{R}^{2D}$$

[a] Snyder et al., "X-vectors: Robust dnn embeddings for speaker recognition".

[b] Desplanques, Thienpondt, and Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification".

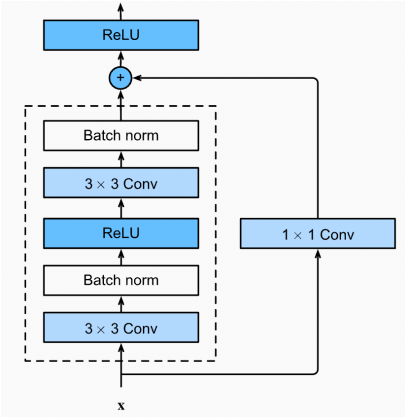


图: r-vector 采用的 ResNet 架构

r-vector[a] 的标签:

- VoxSRC 2019 两个赛道的获胜系统
- 在 DIHARD 2019 所有 4 个赛道的获胜系统中用作嵌入

[a] Zeinali et al., "But system description to voxceleb speaker recognition challenge 2019"; Wang et al., "Discriminative neural embedding learning for short-duration text-independent speaker verification".

标注数据：

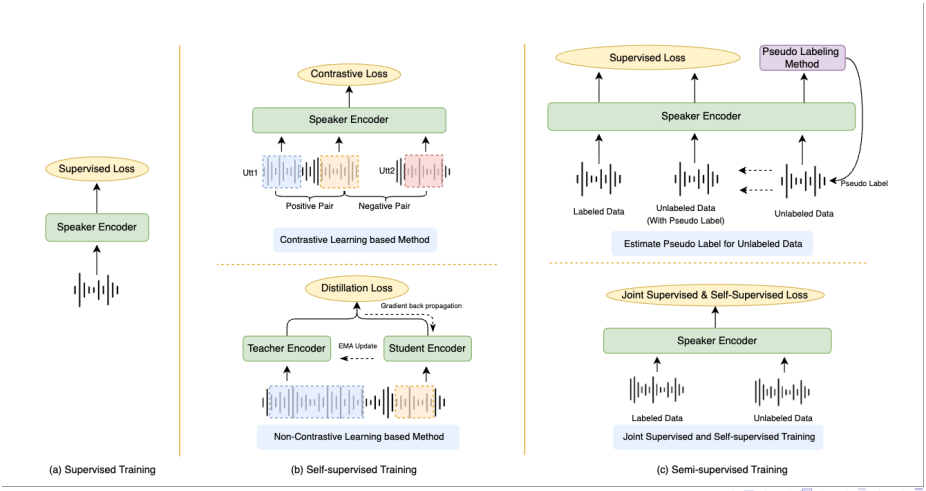
- 标注数据成本高昂
- 使用说话人信息自动收集的数据，如 voxceleb，存在隐私问题
  - voxceleb 数据集不再可从官方网站访问

未标注数据：

- 容易获得
- 覆盖更广泛的真实数据范围
- 无隐私问题

说话人建模

训练范式：从监督学习到无监督学习/半监督学习/自监督学习



(a) Supervised Training (b) Self-supervised Training (c) Semi-supervised Training

- GMM，i-vector 可视为单层 MLP
- d-vector，j-vector，x-vector：少于 10 层
- 基于 ResNet 的模型（常见设置：34 层，在挑战中扩展到 293 层甚至 500+ 层）

- 纯音频模态
- 音频-视觉说话人嵌入

- 从零开始训练说话人判别模型
- 利用大型预训练语音模型如 WavLM
- 半监督：迭代聚类和监督微调

说话人表征学习：趋势

任务：从单任务到跨任务

- 预训练嵌入用于不同任务
- 与特定任务的显式联合优化
- 相关任务中的隐式说话人建模

目录

问题设置

给定话语 $\mathbf{O} = \{\mathbf{o}_t\}_{t=1}^{T} \in \mathbb{R}^{T \times D}$，学习 $\mathcal{F}$ 产生固定长度嵌入 $\mathbf{v} = \mathcal{F}(\mathbf{O}) \in \mathbb{R}^{d}$。

- 相同说话人的嵌入应该接近；不同说话人的嵌入应该远离。
- 使用与验证时使用的余弦/角度对齐的分类驱动目标。
- 引入显式边距以加强开放集泛化。

ASR（经典声学建模）

- 推理时封闭集：音素/音子标签是固定的。
- 目标：在已知标签集上最大化分类准确率。

说话人验证

- 推理时开放集：说话人在训练时未见。
- 需要在余弦/角度下紧凑聚类和清晰边距。

模型

嵌入 $\mathbf{v} = \mathcal{F}(\mathbf{O}) \in \mathbb{R}^d$，分类器 $W = [\mathbf{w}_1, \dots, \mathbf{w}_C] \in \mathbb{R}^{d \times C}$ 带偏置 $\mathbf{b}$。

- 逻辑：$s_j = \mathbf{w}_j^\top \mathbf{v} + b_j$。

$$\mathcal{L}_{\mathsf{CE}} = -\log \frac{e^{s_y}}{\sum_{j=1}^{C} e^{s_j}}.$$

监督损失函数

从分类到嵌入：核心挑战

理想嵌入应具备：

- 类内紧致性：同类样本聚集
- 类间分离性：不同类样本远离

Softmax 的局限性

- 目标错位：只要求正确分类，未显式约束紧致性和分离性
- 开放集问题：对未见身份的泛化能力有限，决策边界缺乏足够"安全边际"

解决方案

引入边际损失来直接优化嵌入质量

归一化

强制 $\|\mathbf{v}\| = 1$，$\|\mathbf{w}_j\| = 1$，并设置 $b_j = 0$。然后

$$s_j = s \cos \theta_j,$$

其中 θ_j 是 $\mathbf{v}$ 和 $\mathbf{w}_j$ 之间的角度，$s > 0$ 是尺度（逆温度）。

$$\mathcal{L}_{\text{Norm-CE}} = -\log \frac{e^{s \cos \theta_y}}{\sum_{j=1}^{C} e^{s \cos \theta_j}}.$$

几何视图

分类等于在单位超球面上选择最小角度，与 $\mathbf{v}$ 的余弦评分一致。

形式

$$\mathcal{L} = -\log \frac{e^{s\cos(m\theta_y)}}{e^{s\cos(m\theta_y)} + \sum_{j\neq y} e^{s\cos\theta_j}}, \quad m \geq 1.$$

- 决策边界：从 $\theta_1 = \theta_2$ 到 $m\theta_1 = \theta_2$。
- 优点：开创性的角度边距公式。

注意事项

高度非线性；训练可能不稳定，通常需要退火/特殊调度。

形式

$$\mathcal{L} = -\log \frac{e^{s(\cos\theta_y - m)}}{e^{s(\cos\theta_y - m)} + \sum_{j \neq y} e^{s\cos\theta_j}}, \quad m > 0.$$

- 解释：要求 $\cos\theta_y \geq \cos\theta_j + m$（余弦空间中的常数边距）。
- 优点：简单稳定的优化；无需退火。

注意

边距在余弦中是常数但在角度中不是常数；角度间隙随 θ 变化。

形式

$$\mathcal{L} = -\log \frac{e^{s\cos(\theta_y + m)}}{e^{s\cos(\theta_y + m)} + \sum_{j \neq y} e^{s\cos\theta_j}}, \quad m > 0.$$

- 解释：要求 $\theta_y + m \leq \theta_j$（常数角度间隙）。
- 优点：清晰的几何；强大的开放集性能。

实现技巧

确保数值稳定性（将 $\cos\theta$ 裁剪到 $[-1, 1]$；避免不稳定的反三角函数操作）。

数值稳定技巧的适用边界

为何仅 ArcFace 重点强调？

核心结论

数值稳定技巧（裁剪 $\cos\theta$、规避 arccos）并非 ArcFace 独有，但仅 ArcFace 因「计算逻辑强制依赖角度操作」，必须重点解决；其他损失（SphereFace/CosFace）因「计算路径不同」，对数值异常的敏感度更低或可自然规避。

数值异常敏感度对比

损失函数	需 arccos 反推 θ？	对 $\cos\theta$ 超范围敏感？
CosFace	否（直接用 $\cos\theta$）	低（超范围不影响减法操作）
SphereFace	否（倍角公式绕开）	中（超范围仅影响 $\cos(m\theta)$ 精度）
ArcFace	是（必须反推 θ）	高（超范围直接返回 NaN）

- ArcFace 痛点：arccos 是「必选项」，$\cos\theta$ 超范围即崩溃；
- 其他损失：arccos 是「可选项」，数值异常仅影响效果，不中断训练。

监督损失函数

三种边距一览

方法	边距域	决策边界 (1 vs 2)
SphereFace	角度（乘法）	$\cos(m\theta_1) = \cos(\theta_2)$
CosFace	余弦（加法）	$\cos(\theta_1) - m = \cos(\theta_2)$
ArcFace	角度（加法）	$\cos(\theta_1 + m) = \cos(\theta_2)$

常数性

CosFace：余弦中常数。
ArcFace：角度中常数。
SphereFace：取决于 θ。

实践

为了稳定性和性能，优先选择 ArcFace

形式

$$\mathcal{L}_{\text{center}} = \frac{1}{2}\sum_i \|\mathbf{v}_i - \mathbf{c}_{y_i}\|_2^2.$$

- 通过显式减少说话人内部方差来补充边距 softmax。
- 结合为 $\mathcal{L} = \mathcal{L}_{\text{ArcFace/CosFace}} + \lambda\mathcal{L}_{\text{center}}$，使用小的 λ。

形式

$$\mathcal{L} = -\log \frac{e^{s\cos(\theta_y + m_y)}}{e^{s\cos(\theta_y + m_y)} + \sum_{j \neq y} e^{s\cos(\theta_j - m_j)}}$$

其中 $m_y, m_j > 0$ 为类别自适应边际参数

- 核心思想：为不同类别分配不同边际，动态适应数据特性
- 边际调整策略：
  - 样本量少的类别使用更大边际
  - 类内离散度高的类别使用更大边际
  - 难分样本比例高的类别使用更大边际
- 优点：处理数据不平衡问题；对难分类别优化更精准

实现考虑

需要额外计算类别统计信息；边际调整策略需仔细设计

- 利用大型预训练模型
  - 自监督预训练语音模型
  - ASR 模型初始化
  - 高效微调
- 自监督学习方法
  - SimCLR/MoCo/DINO
  - 阶段性迭代训练

- 自监督预训练语音模型
  - Wav2Vec[a]
  - HuBERT[b]
  - WavLM[c]
  - UniSpeech[d]

[a] Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations".

[b] Hsu et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units".

[c] Chen et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing".

[d] Chen et al., "Unispeech-sat: Universal speech representation learning with speaker aware pre-training".

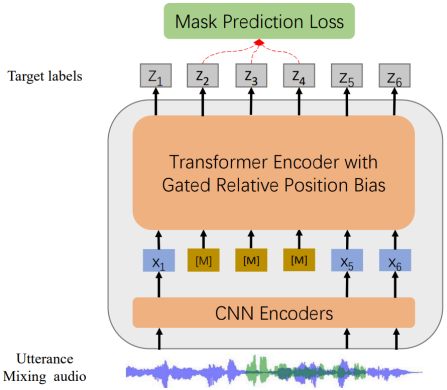
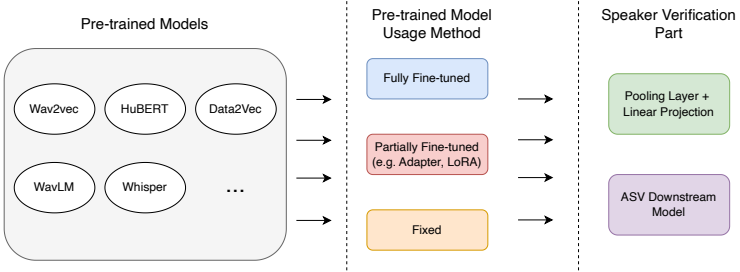


图: WavLM 模型架构



集成策略：

1. 特征提取：使用预训练特征作为输入
2. 微调：针对说话人任务适配预训练模型
3. 多任务学习：联合优化多个任务

在说话人验证任务上微调 **SSL** 语音模型[a]

- 用预训练模型的表征替换 Fbank
- 可学习加权求和

[a]Chen et al., "Large-scale self-supervised speech representation learning for automatic speaker verification".

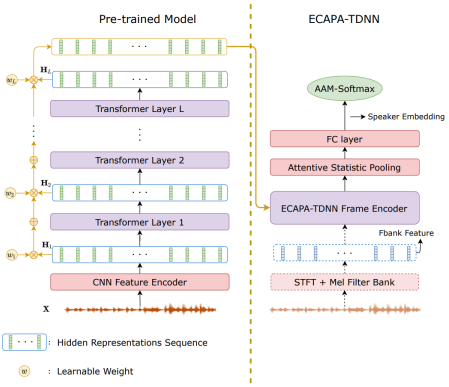
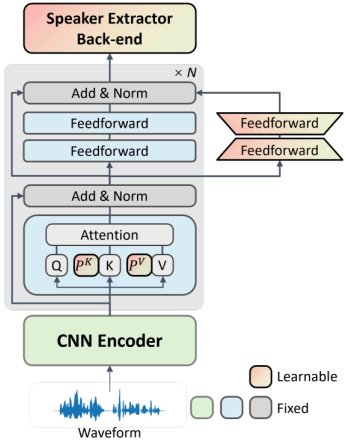


图: 利用预训练模型表征

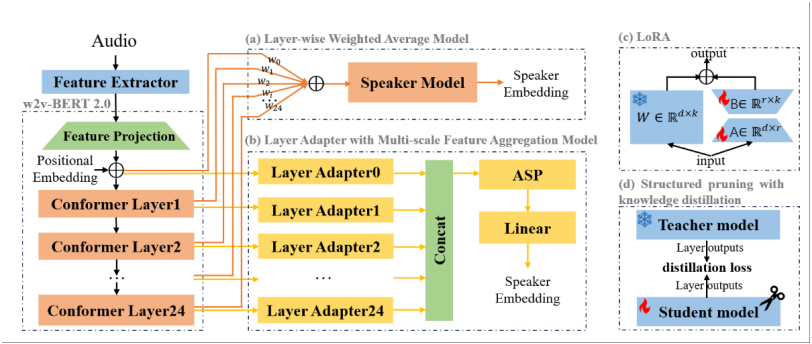
在说话人验证上高效微调自监督模型与适配器[a]

- 冻结大型预训练模型
- 使用适配器在说话人任务上高效微调

[a]Peng et al., "Parameter-efficient transfer learning of pre-trained Transformer models for speaker verification using adapters".



采用大型自监督预训练模型 w2v-BERT 2.0，SOTA 性能[2]



图：利用 w2v-BERT 2.0 和知识蒸馏引导的结构化剪枝增强说话人验证性能

[2]Li, Cheng, and Li, "Enhancing Speaker Verification with w2v-BERT 2.0 and Knowledge Distillation guided Structured Pruning".

在说话人验证任务上微调 **ASR** 模型[ab]

- 用 ASR 数据集预训练模型
- 为说话人任务训练初始化

[a]Liao et al., "Towards a unified conformer structure: from asr to asv task".
[b]Cai et al., "Pretraining Conformer with ASR for Speaker Verification".

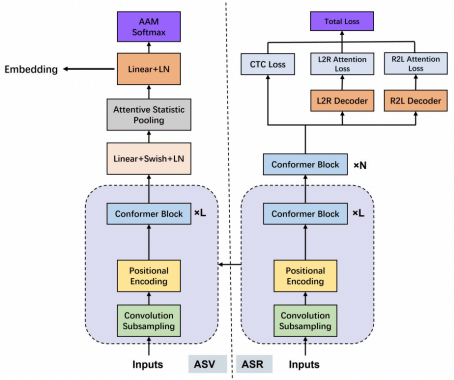


图: ASR 迁移示意图

说话人验证任务上自监督学习的假设[a]

- 来自相同话语的片段属于相同说话人
- 来自不同话语的片段属于不同说话人

[a]Huh et al., "Augmentation adversarial training for self-supervised speaker recognition".

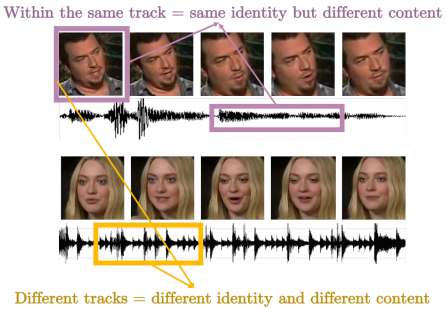


图: 假设示意图

基于度量学习的损失函数提供对比监督信号，如 Triplet、Prototypical、GE2E[3]和 Angular Prototypical[4]。

$$L_{\text{Triplet}} = \frac{1}{N}\sum_{j=1}^{N}\max(0, \|\mathbf{x}_{j,0} - \mathbf{x}_{j,1}\|_2^2 - \|\mathbf{x}_{j,0} - \mathbf{x}_{k\neq j,1}\|_2^2 + m)$$

$$L_{\text{Prototypical}} = -\frac{1}{N}\sum_{j=1}^{N}\log\frac{e^{\mathbf{S}_{j,j}}}{\sum_{k=1}^{N} e^{\mathbf{S}_{j,k}}}, where\ \mathbf{S}_{j,k} = \|\mathbf{x}_{j,M} - \mathbf{c}_k\|_2^2$$

[3]Wan et al., "Generalized end-to-end loss for speaker verification".

[4]Chung et al., "In defence of metric learning for speaker recognition".

基于 SimCLR[a]框架，适应说话人任务[b]

- 从话语中裁剪两个片段并构建正负对
- 使用度量损失吸引正对并排斥负对

[a]Chen et al., "A simple framework for contrastive learning of visual representations".

[b]Zhang, Zou, and Wang, "Contrastive self-supervised learning for text-independent speaker verification".

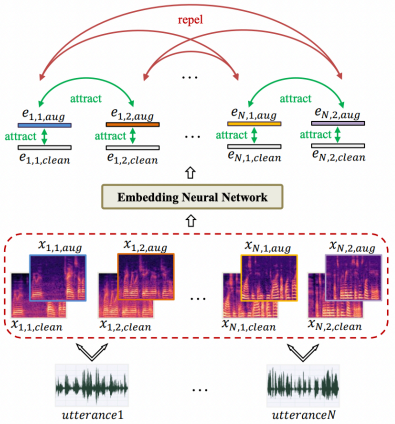


图: SimCLR 在说话人任务上的示意图

基于 DINO[a]框架，适应说话人任务[bc]

- 从一个话语中裁剪几个片段并只构建正对
- 使用交叉熵损失吸引正对

[a] Caron et al., "Emerging properties in self-supervised vision transformers".

[b] Han, Chen, and Qian, "Self-supervised speaker verification using dynamic loss-gate and label correction".

[c] Chen et al., "A comprehensive study on self-supervised distillation for speaker representation learning".

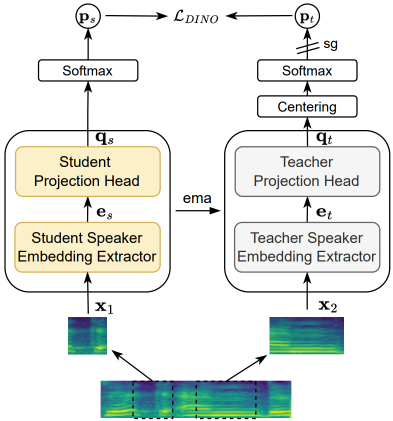


图: DINO 在说话人任务上的示意图

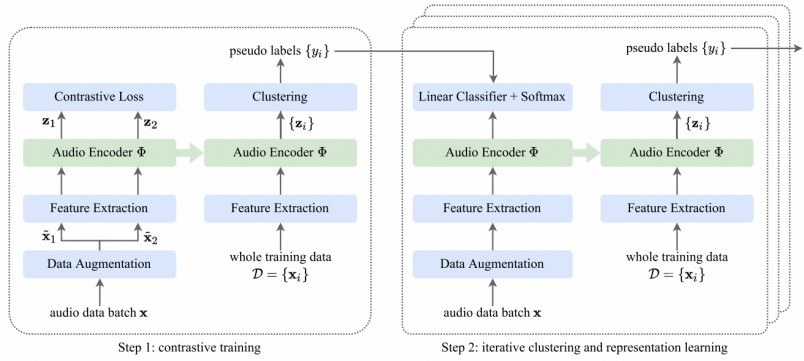
基于两阶段的迭代框架[abc]

- I：对比训练
- II：迭代聚类和表征学习

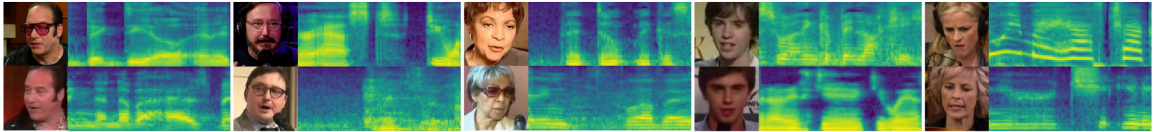
[a]Cai, Wang, and Li, "An iterative framework for self-supervised deep speaker representation learning".

[b]Han, Chen, and Qian, "Self-Supervised Learning with Cluster-Aware-DINO for High-Performance Robust Speaker Verification".

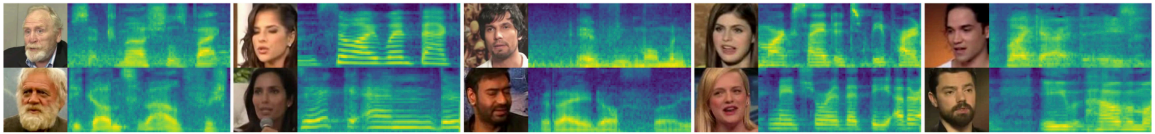
[c]Tao et al., "Self-supervised speaker recognition with loss-gated learning".



图：SSL 说话人验证迭代框架示意图



(a)

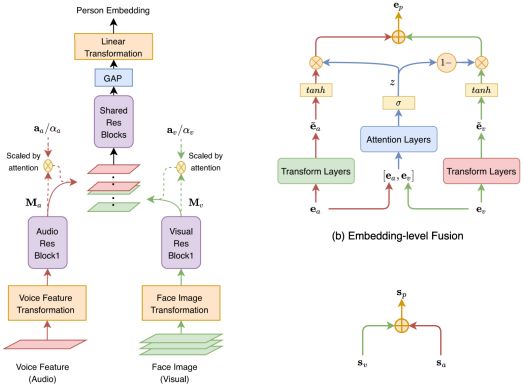


(b)

图: 基于音频或视觉信息的说话人相似度[5]

- 上半部分显示说话人与同一个人的相似度
- 下半部分显示说话人与不同人之间的相似度

[5]Qian, Chen, and Wang, "Audio-visual deep neural network for robust person verification".



图：不同级别的音频-视觉信息融合[a]

- 嵌入级融合比低级融合表现更好
- 嵌入级融合中的注意力机制使其比分数级融合更抗噪声

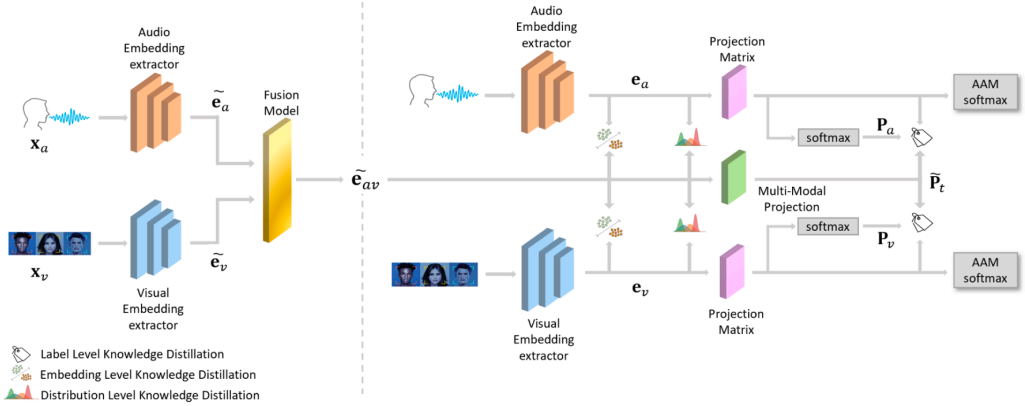


图: 从音频-视觉系统到单模态系统的知识蒸馏[6]

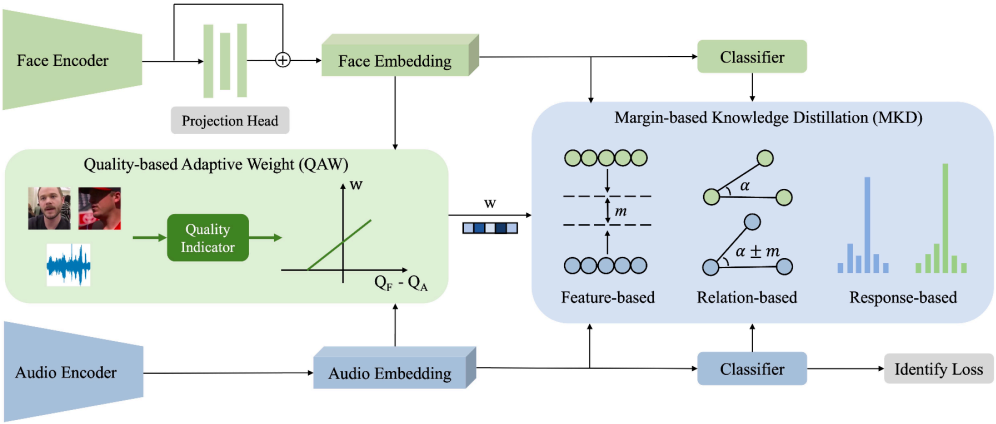


图: 从视觉系统到音频系统的知识蒸馏[7]

在说话人表征学习中，我们主要从两个角度优化模型的效率：计算效率和内存效率

- 计算效率
  - 知识蒸馏
  - 网络量化
  - 高效架构设计
- 内存效率
  - 可逆神经网络

说话人验证任务上的知识蒸馏

- 从教师模型到学生模型的知识蒸馏[a]
- 通过特征增强的自知识蒸馏[b]
- 从多模态到单模态的知识蒸馏[c]

[a]Wang et al., "Knowledge Distillation for Small Foot-print Deep Speaker Embedding".

[b]Liu et al., "Self-Knowledge Distillation via Feature Enhancement for Speaker Verification".

[c]Zhang, Chen, and Qian, "Knowledge Distillation from Multi-Modality to Single-Modality for Person Verification".

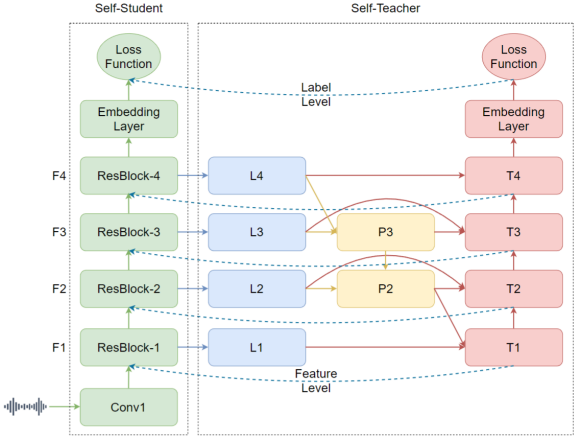
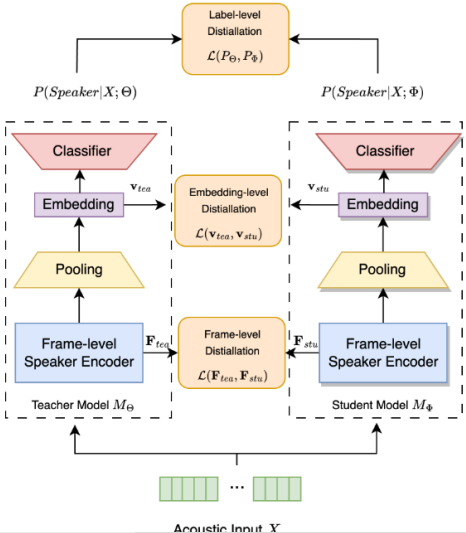


图: 通过特征增强的自知识蒸馏示意图

知识蒸馏在不同层级的实现

- 特征级蒸馏
- 嵌入级蒸馏
- 标签级蒸馏



量化通过降低参数精度实现模型压缩

- 二元神经网络[a]
- 线性和 PoT(2 的幂) 量化[b]
- 基于 K-Means 的量化[c]
- 静态和自适应量化器用于二元量化[d]

[a] Zhu, Qin, and Li, "Binary Neural Network for Speaker Verification".

[b] Liu et al., "Self-Knowledge Distillation via Feature Enhancement for Speaker Verification".

[c] Wang et al., "Adaptive Neural Network Quantization For Lightweight Speaker Verification".

[d] Liu, Wang, and Qian, "Extremely Low Bit Quantization for Mobile Speaker Verification Systems Under 1MB Memory".

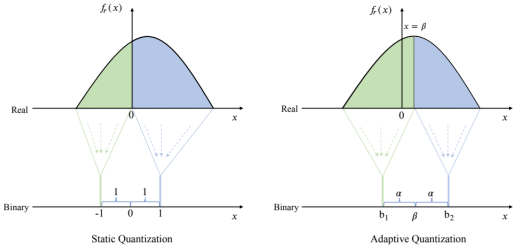


图: 静态和自适应二元量化概述

说话人验证任务上的高效架构设计

- 深度优先神经架构与注意力特征融合[a]
- CS-CTCSCONV1D[b](信道分割时间-信道-时间可分离 1 维卷积)
- 非对称注册-验证结构 (ECAPA-TDNNLite[c])

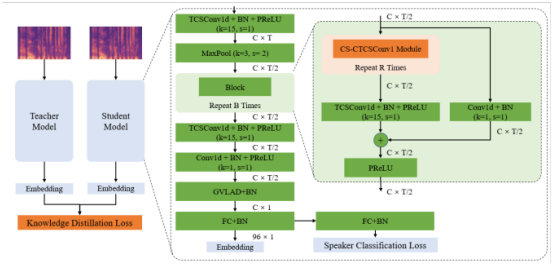
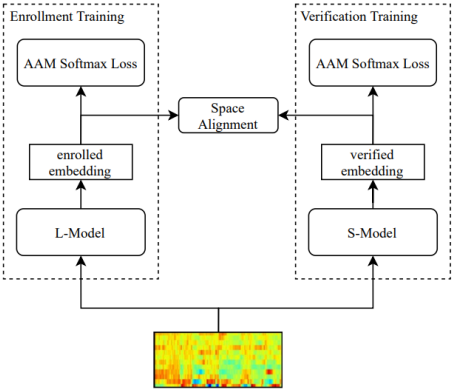


图: CS-CTCSCONV1D 示意图

[a] Liu, Chen, and Qian, "Depth-First Neural Architecture With Attentive Feature Fusion for Efficient Speaker Verification".

[b] Cai et al., "CS-CTCSCONV1D: Small footprint speaker verification with channel split time-channel-time separable 1-dimensional convolution".

[c] Li et al., "Towards Lightweight Applications: Asymmetric Enroll-Verify Structure for Speaker Verification".



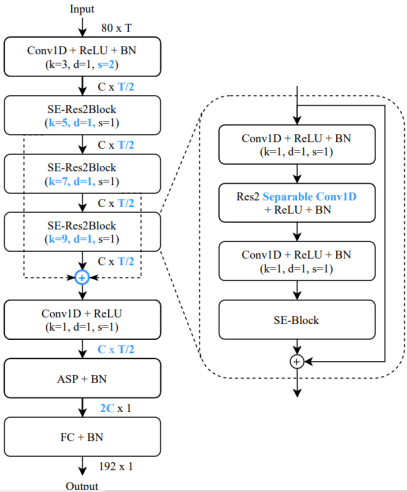


图: 非对称结构的训练过程。帧级输入特征分别馈

表: 压缩/量化 ResNet34 和其他全精度紧凑架构的实验结果。

模型	大小 (MB)	位宽 (bit)	Vox1-O EER(%)
KMQAT-ResNet34[8]	3.45	4	**0.957**
PoT-ResNet34[9]	3.45	4	1.09
TWN-ResNet34[10](our impl.)	1.80	2	1.473
b-vector(adaptive)[11]	0.97	**1**	**1.72**
ResNet34(binary)[12]	**0.66**	**1**	5.355
CS-CTCSConv1d	0.96	32	2.62
ECAPA-TDNNLite	1.2	32	3.07

[8]Wang et al., "Adaptive Neural Network Quantization For Lightweight Speaker Verification".

[9]Li et al., "Model Compression for DNN-based Speaker Verification Using Weight Quantization".

[10]Li, Zhang, and Liu, "Ternary weight networks".

[11]Liu, Wang, and Qian, "Extremely Low Bit Quantization for Mobile Speaker Verification Systems Under 1MB Memory".

[12]Zhu, Qin, and Li, "Binary Neural Network for Speaker Verification".

可逆神经网络[a] (RevNets) 缓解了在反向传播期间在内存中存储激活的需要。因此，RevNets 需要几乎恒定的内存成本，随着网络深度增加。

- 部分可逆网络
- 完全可逆网络

[a] Liu and Qian, "Reversible Neural Networks for Memory-Efficient Speaker Verification".

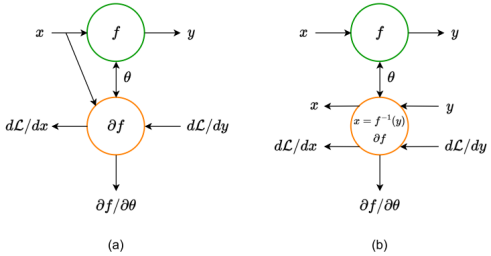
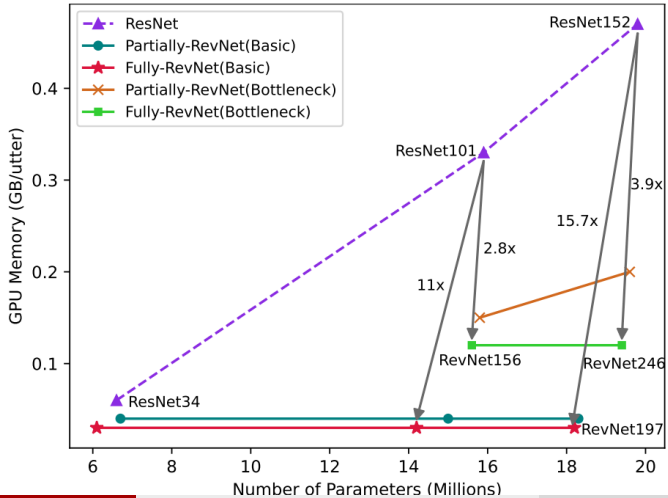


图: 非可逆算子 (a) 和可逆算子 (b) 的比较


ResNet
Partially-RevNet(Basic)
Fully-RevNet(Basic)
Partially-RevNet(Bottleneck)
Fully-RevNet(Bottleneck)
ResNet152
ResNet101
ResNet34
RevNet156
RevNet246
RevNet197
3.9x
15.7x
2.8x
11x
GPU Memory (GB/utter)
Number of Parameters (Millions)
0.1
0.2
0.3
0.4
6
8
10
12
14
16
18
20

模型效率的其他工作

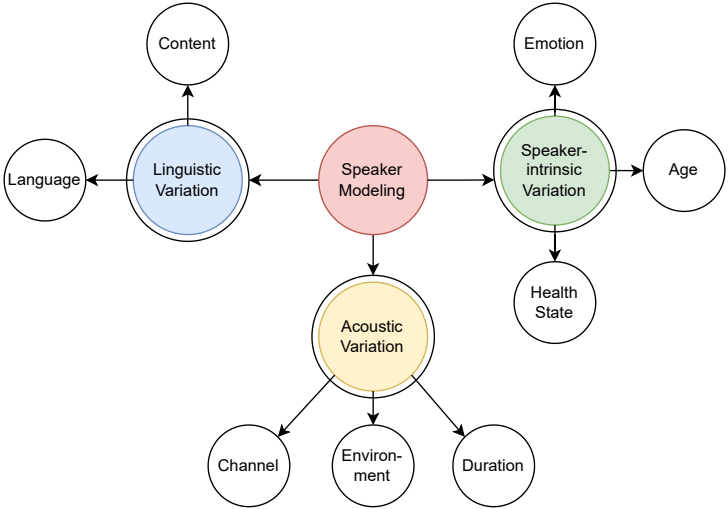
- Thin-ResNet[13]
- Fast-ResNet[14]
- ADMM[15]
- Small Footprint Text-Independent Speaker Verification[16]

[13] Cai, Chen, and Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system".

[14] Chung et al., "In Defence of Metric Learning for Speaker Recognition".

[15] Xu et al., "Mixed Precision Low-Bit Quantization of Neural Network Language Models for Speech Recognition".

[16] Balian et al., "Small footprint text-independent speaker verification for embedded systems".


Content
Emotion
Language
Linguistic Variation
Speaker Modeling
Speaker-intrinsic Variation
Age
Health State
Acoustic Variation
Channel
Environ-ment
Duration

录音环境也会在说话人身份建模中引入变异性，受录音设备和麦克风距离等因素影响。为了增强跨不同设备的模型鲁棒性，在说话人识别中应用了各种域自适应方法，包括

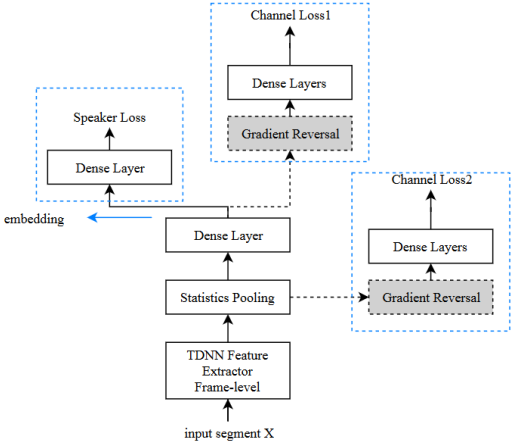
- 基于差异的对齐
- 对抗学习
- 域特定适配器

基于差异的对齐旨在最小化潜在特征空间中的域差异并促进学习域不变表征。为了实现这一目标，选择适当的散度度量是这些方法的核心。广泛使用的度量包括 MMD[17]、相关对齐 (CORAL)[18]等。

$$\mathcal{L}_{\text{mmd}} \triangleq \sup_{\phi \in \Phi} \left(\mathrm{E}_S \left[\phi \left(S \right) \right] - \mathrm{E}_T \left[\phi \left(T \right) \right] \right) \tag{2}$$

[17]Li, Han, and Song, "CDMA: Cross-Domain Distance Metric Adaptation for Speaker Verification".

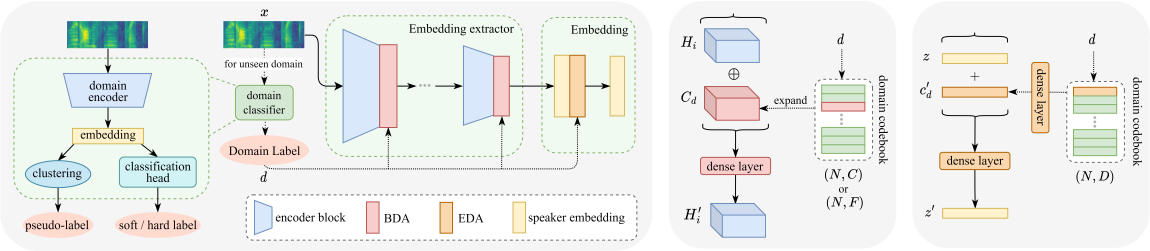
[18]Li, Zhang, and Chen, "The coral++ algorithm for unsupervised domain adaptation of speaker recognition".



对抗学习使用域分类器从特征中消除判别性域信息。域对抗训练中的最小-最大优化最小化域间隙并强制域不变特征提取[a]。

[a] Chen et al., "Channel invariant speaker embedding learning with joint multi-task and adversarial training".

不是直接用差异度量对齐域，而是结合域特定适配器等额外模块有助于捕获和缓解域方差，产生域不变嵌入。



(a) Model architecture (b) BDA (c) EDA

图：带域特定适配器的框架[19]

[19]Huang et al., "Enhancing Cross-Domain Speaker Verification through Multi-Level Domain Adapters".

观察：在现实场景中，说话人验证系统在一种语言上训练并在另一种语言上测试时可能会退化。

超过 40% 的世界人口是双语的，当注册和测试使用的语言不同时会发生这种不匹配。

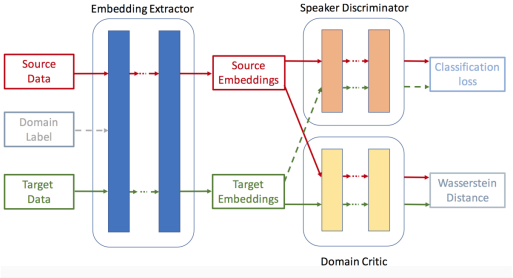


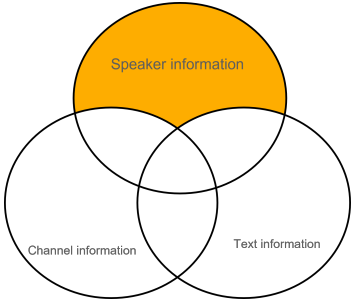
图: 语言不匹配对抗学习结构

对抗学习使用语言分类器从特征中消除判别性语言信息。域对抗训练中的最小-最大优化最小化语言间隙并强制语言不变特征提取[ab]。

[a] Rohdin et al., "Speaker verification using end-to-end adversarial language adaptation".

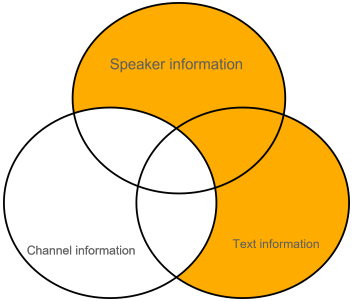
[b] Xia, Huang, and Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation".

除了说话人信息，文本或内容是语音传达的最关键信息。



对于文本无关的说话人任务，我们只需要说话人信息

注册：Hey Siri；测试：随便说什么



对于文本相关的说话人任务，我们还需要内容信息

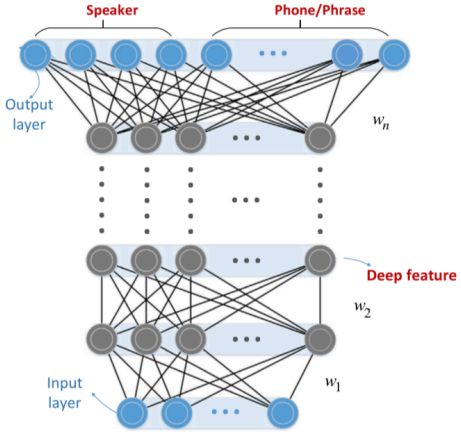
注册：Hey Siri；测试：Hey Siri

内容信息的表征

- 音素索引
- ASR 预测的音素后验
- ASR 模型的隐藏层输出
- 短语编号 (固定短语数据集)
- 归一化音素分布

文本鲁棒性

d-vector 框架中的多任务学习[20]



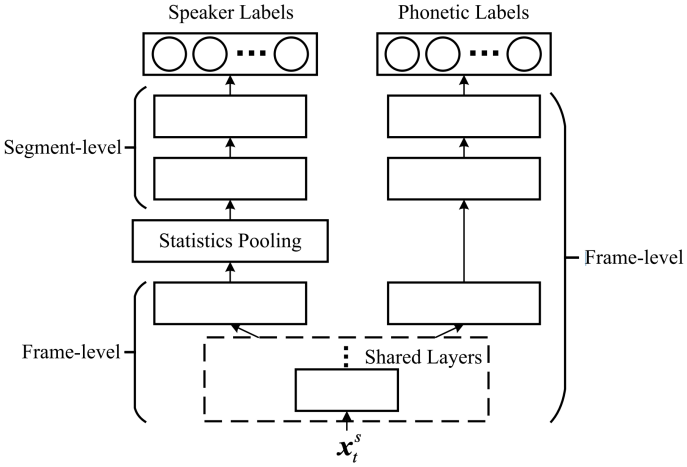
- 文本相关任务
- 在帧级多任务
- 性能提升

显式建模音素信息有助于文本相关说话人验证任务，这是直观的。

[20] Liu, Yuan, et al. "Deep feature for text-dependent speaker verification." Speech Communication 73 (2015): 1-13.

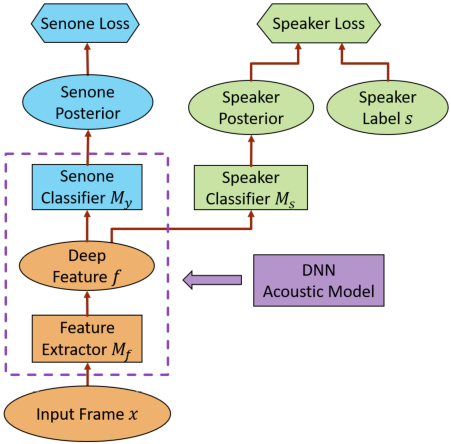
文本鲁棒性

x-vector 框架中的多任务学习[21]



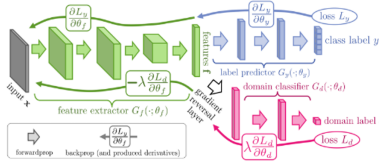
- 文本无关任务
- 在帧级多任务
- 性能提升！

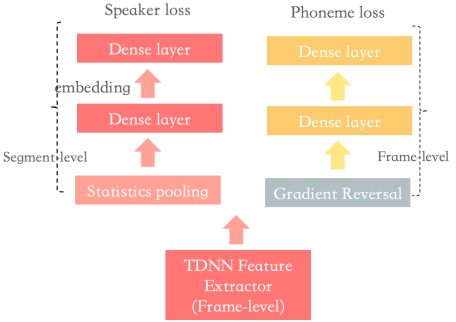
[21] Liu, Yi, et al. "Speaker Embedding Extraction with Phonetic Information." Proc. Interspeech 2018 (2018): 2247-2251.



- 声学建模
- 抑制说话人效应的对抗训练
- 性能提升

[22] Meng, Zhong, et al. "Speaker-invariant training via adversarial learning." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.



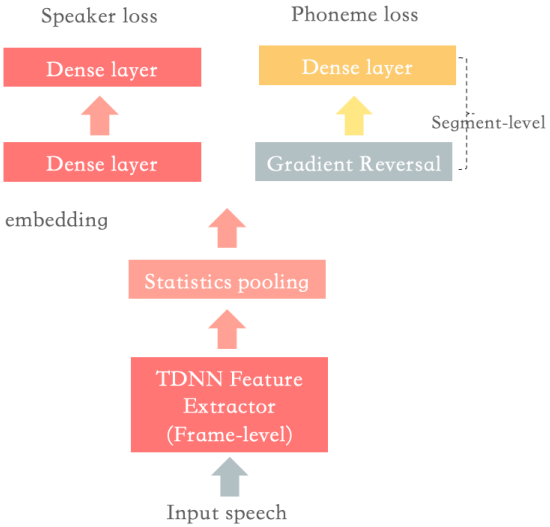


$$\mathcal{L}_s = \mathsf{CE}(M_s(M_f(\mathbf{X})), \mathbf{y}^s)$$

$$\mathcal{L}_p = \frac{1}{N}\sum_{i=1}^{N} \mathsf{CE}(M_p(M_f(\mathbf{x}_i)), \mathbf{y}_i^p)$$

$$\mathcal{L}_{total} = \mathcal{L}_s + \mathcal{L}_p$$

Systems	voxceleb1_O	voxceleb1_E	voxceleb1_H
x-vector baseline	2.361	2.470	4.260
FRM-MT	2.165	2.198	3.911
FRM-ADV	3.143	3.214	5.419

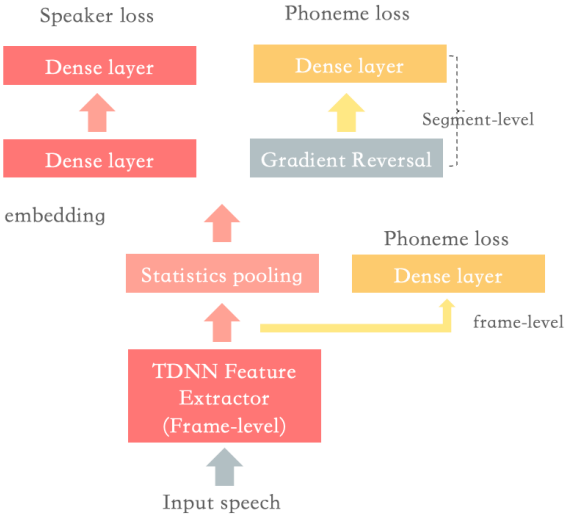


$$\mathcal{L}_s = \mathsf{CE}(M_s(M_f(\mathbf{X})), \mathbf{y}^s)$$
$$\mathcal{L}_p = \mathsf{CE}(M_p(M_f(\mathbf{x}_i)), \mathbf{y}^p)$$
$$\mathcal{L}_{total} = \mathcal{L}_s + \mathcal{L}_p$$

对于具有 N 帧的给定段 $\mathbf{x}$，段级音素标签 $\mathbf{y}^p$ 是

$$\mathbf{y}^p = \{y_1, y_2, \cdots, y_C\}$$
$$y_c = \frac{N_c}{N}$$

其中 C 是所选音素集的大小。N_c 表示第 c 个音素在 $\mathbf{x}$ 中的出现次数

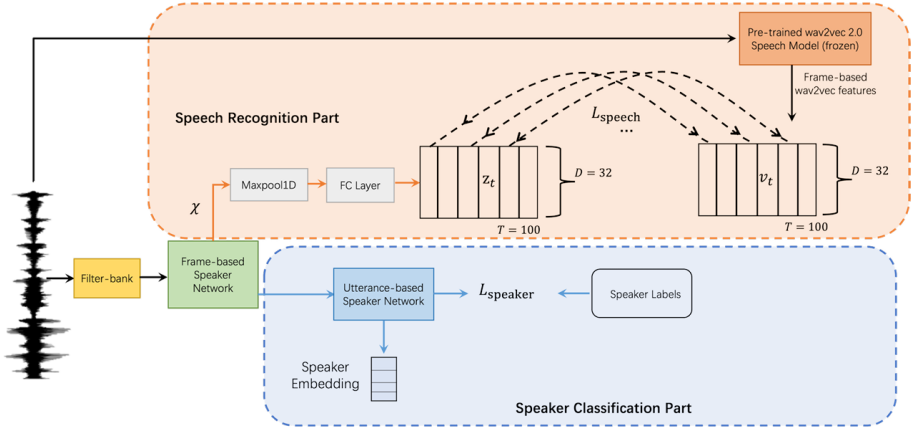


Systems	voxceleb1_O	voxceleb1_E	voxceleb1_H
x-vector baseline	2.361	2.470	4.260
SEG-MT	2.175	2.330	4.059
SEG-ADV	2.154	2.198	3.923

图：段级多任务/对抗训练

Systems	voxceleb1_O	voxceleb1_E	voxceleb1_H
x-vector baseline	2.361	2.470	4.260
FRM-MT	2.165	2.198	3.911
SEG-ADV	2.154	2.198	3.923
COMBINE	2.013	2.030	3.819

图：帧级多任务 + 段级对抗训练



[23] Jin, Tu, and Mak, "Phonetic-aware speaker embedding for far-field speaker verification".

从预训练 ASR 模型中提取音素瓶颈 (PBN) 并将其与滤波器组结合[24]

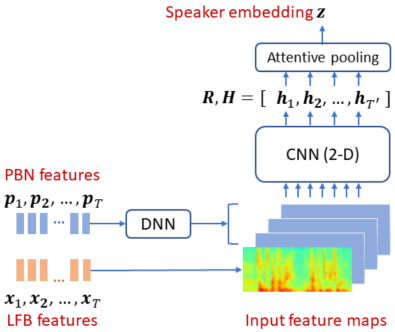


Fig. 1: *Implicit phonetic attention by combining LFB and PBN features at the input layer (LFB: log filter bank; PBN: phonetic bottleneck).*

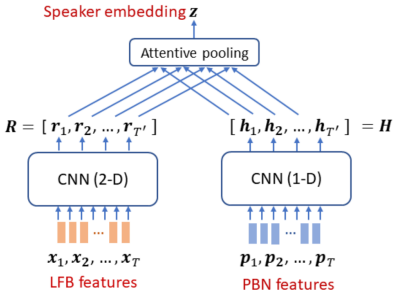


Fig. 2: *Explicit phonetic attention by routing LFB and PBN features through separate networks (LFB: log filter bank; PBN: phonetic bottleneck).*

[24] Zhou T, Zhao Y, Li J, et al. CNN with phonetic attention for text-independent speaker verification, ASRU 2019

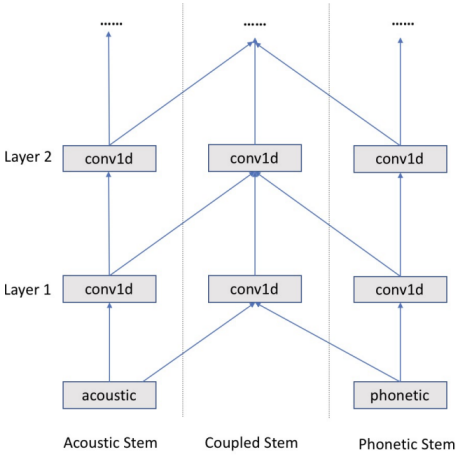


Table 1: *Network configurations of PacNet*

Layer 7	Linear	In=1024 Out=1000		
Layer 6	Pooling	In=1024 Out=1024		
Layer 5	Conv1d	In=2048 Out=1024		
Layer 4	Conv1d kernel=5	Out=512 In=512	Out=1024 In=2048	Out=512 In=512
Layer 3	Conv1d kernel=5	Out=512 In=512	Out=1024 In=2048	Out=512 In=512
Layer 2	Conv1d kernel=5	Out=512 In=512	Out=1024 In=2048	Out=512 In=512
Layer 1	Conv1d kernel=5	Out=512 In=40	Out=1024 In=140	Out=512 In=100
Stem		Acoustic	Coupled	Phonetic

- 使用 Triplet 损失而不是 softmax 损失

25

[25] Zheng, Lei, and Suo, "Phonetically-Aware Coupled Network For Short Duration Text-Independent Speaker Verification."

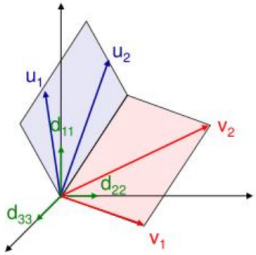
一般思想：分解和重构。应用远不止说话人建模

应用

- 说话人表征学习
- 语音转换
- 语音合成/语音克隆

说话人表征学习的联合因子分析[26]

$$\mathbf{M} = \mathbf{M}^{\text{UBM}} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z} + \mathbf{U}\mathbf{x}$$

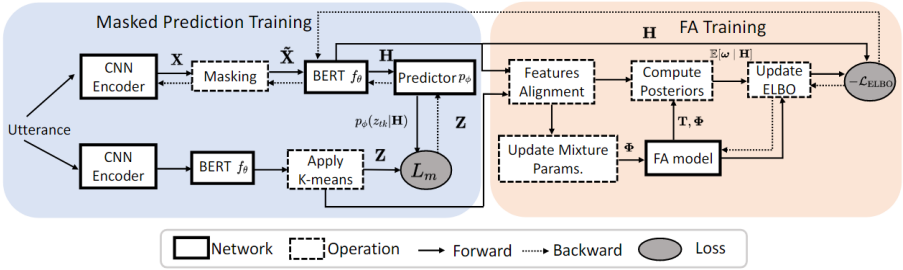


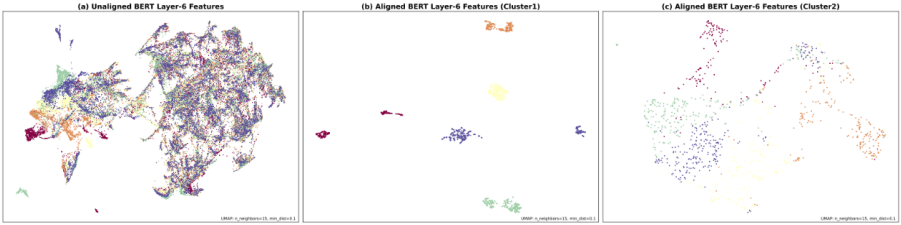
- 假设因子 $\mathbf{y}$、$\mathbf{z}$、$\mathbf{x}$ 的高斯先验
- 使用 EM 算法估计 $\mathbf{M}^{\text{UBM}}$、$\mathbf{V}$、$\mathbf{D}$、$\mathbf{U}$
- $\mathbf{V}$ 捕获主要说话人变异性 (特征语音)
- $\mathbf{D}$ 捕获信道变异性
- $\mathbf{U}$ 捕获残差变异性

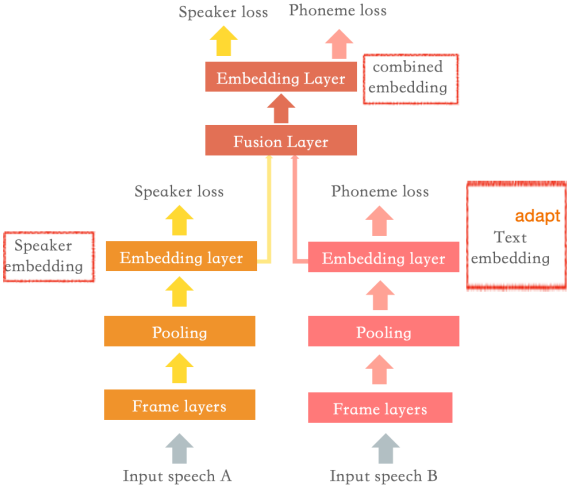
[26] Kenny, Patrick, et al. "Joint factor analysis versus eigenchannels in speaker recognition." TASLP 2007

语音表征解耦

神经因子分析：通过额外对齐忽略音素变异性 [27]







- 段级重构
- 解耦说话人和文本信息
- 对于文本无关任务，我们忽略文本信息
- 对于文本相关任务，我们使用组合嵌入
- 文本自适应任务：修改嵌入中的文本信息同时保持说话人身份。(更改注册关键词)

[28]Yang Y*, Wang S*, Gong X, et al. Text adaptation for speaker verification with speaker-text factorized embeddings. ICASSP 2020

语音表征解耦

音素信息的解耦和重组[29]

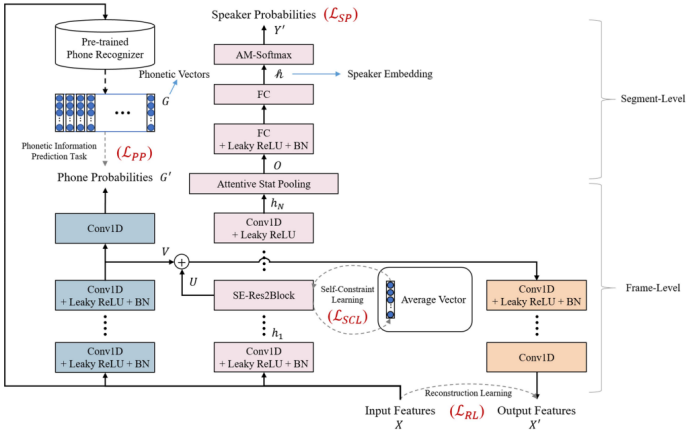
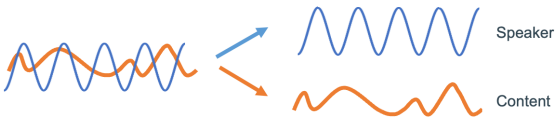


Fig. 3. The architecture of the proposed DROP-TDNN x-vector system. DROP-TDNN consists of three training procedures, including phonetic information prediction, reconstruction and speaker recognition.

- 帧级重构
- 将帧级说话人表征向其均值中心化
- 粗粒度音素类别 (元音、半元音、塞擦音、...)

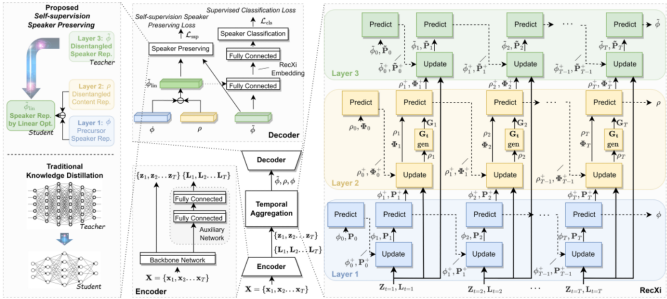
[29] Hong Q B, Wu C H, Wang H M. Decomposition and Reorganization of Phonetic Information for Speaker Embedding Learning. TASLP 2023

Disentangle **Static** and **Dynamic** components in Speech



by three **Gaussian inference layers**

and a novel **speaker preserving self-supervision**



[30] Liu T C, Disentangling Voice and Content with Self-Supervision for Speaker Recognition

确保第一层表征包含内容相关信息，后续残差层将自然地用剩余细节填补空白——具体来说，建模副语言信息。[31]

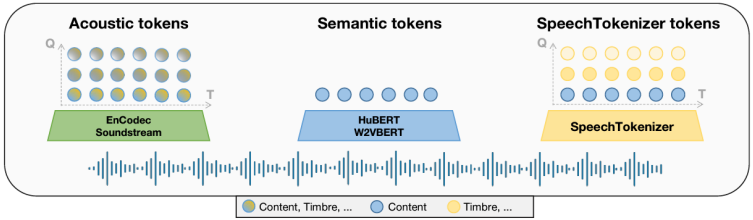


Figure 1: Illustration of information composition of different discrete speech representations. Speech tokens are represented as colored circles and different colors represent different information.

[31] Zhang X, Zhang D, Li S, et al. SpeechTokenizer: Unified Speech Tokenizer for Speech Large Language Models[J]. arXiv preprint arXiv:2308.16692, 2023.

语义蒸馏实现解耦

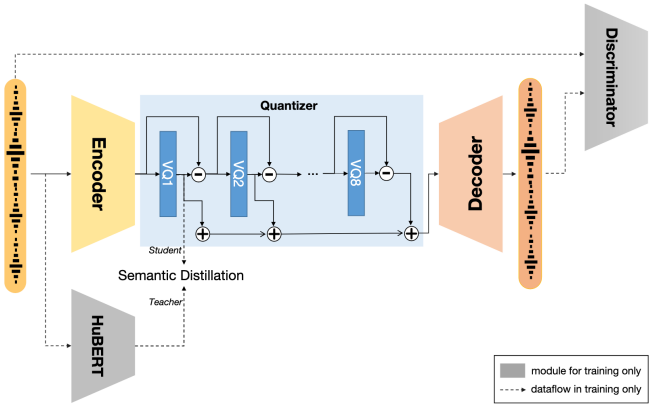


Figure 2: Illustration of SpeechTokenizer framework.

- 连续蒸馏 HuBERT 第 9 层输出/所有层平均
  $\mathcal{L}_{\text{distll}} = \frac{1}{T}\sum_{t=1}^{T}\log\sigma\left(\cos\left(Aq_1^t, s^t\right)\right)$
- 离散蒸馏伪标签预测
  $\mathcal{L}_{\text{distll}} = -\frac{1}{T}\sum_{t=1}^{T} u^t \log\left(\text{Softmax}\left(Aq_1^t\right)\right)$

假设 hubert 是完美的语义编码器

目录

假设：如果某个属性在说话人表征中编码，预测该属性的分类器的准确率取决于其嵌入程度。[32],[33],[34],[35]

- 说话人相关属性：身份、性别和语速。
- 文本相关因素：口语词汇、词序和话语长度。
- 信道相关元素包括手机 ID 和噪声类型。

[32] Wang, Qian, and Yu, "What does the speaker embedding encode?"

[33] Belinkov and Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems".

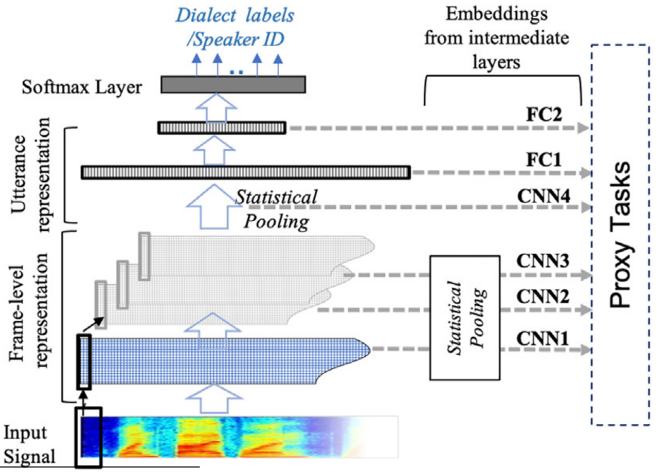
[34] Raj et al., "Probing the information encoded in x-vectors".

[35] Zhao et al., "Probing Deep Speaker Embeddings for Speaker-related Tasks".

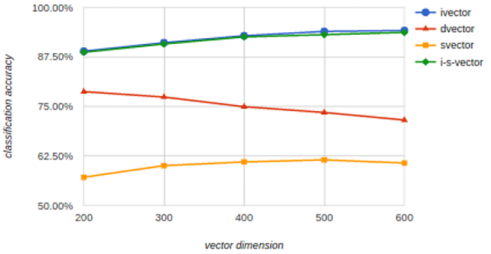
可解释性：通过探测任务探索模型能力

分析编码的信息

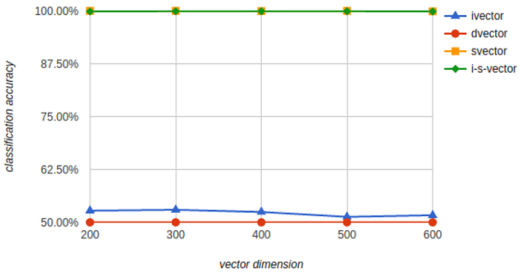
范式说明：用代理任务探测预训练嵌入[36]



说话人身份任务和词序任务的例子[37]



Speaker identity task



Word order task

[37]Wang, Qian, and Yu, "What does the speaker embedding encode?"

在说话人建模的背景下，f 是说话人分类器，c 表示类别，θ 表示可训练模型参数。

$$y^c = f_c(x;\theta)$$

对于第 k 个激活图 A^k(例如 k 表示第 k 个通道)，每个条目 w_{ij}^{kc} 定义为

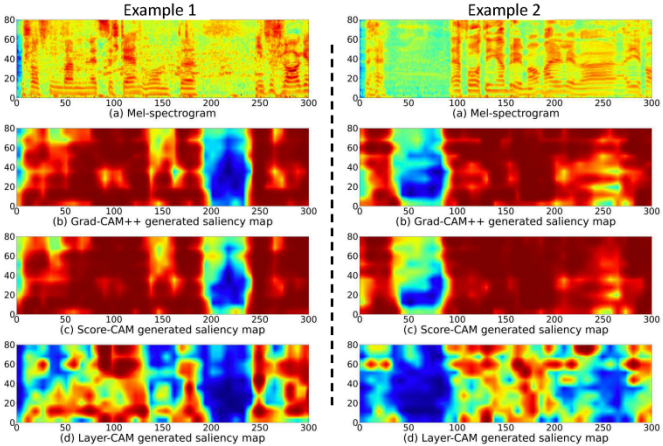
$$w_{ij}^{kc} = \mathrm{ReLU}\left(\frac{\partial y^c}{\partial A_{ij}^k}\right)$$

显著性图定义为线性组合

$$S_{ij}^c = \mathrm{ReLU}\left(\sum_k w_{ij}^{kc} \cdot A_{ij}^k\right)$$

可解释性：通过可视化测量重要性

说话人识别中的可视化[38,39]

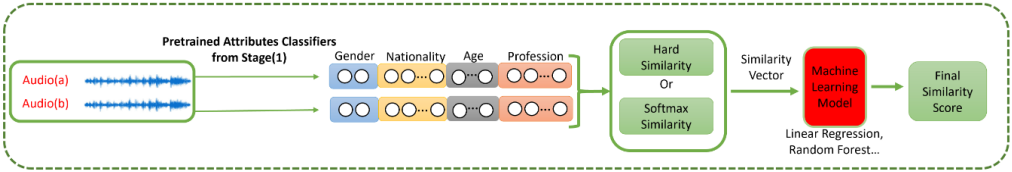


[38] Li et al., "Reliable visualization for deep speaker recognition".

[39] Li et al., "Visualizing data augmentation in deep speaker recognition".

基于属性的可解释 SV：

利用语音属性（性别、年龄、国籍、职业）实现更具有可解释性的说话人验证。

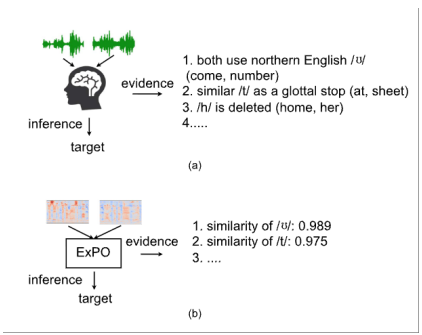


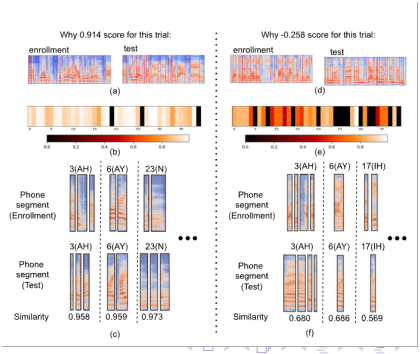
[40]Wu et al., "Explainable attribute-based speaker verification".

ExPO[a]

[a]Ma et al., "ExPO: Explainable Phonetic Trait-Oriented Network for Speaker Verification".

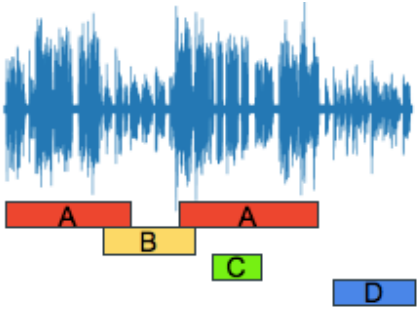
ExPO 不仅生成话语级说话人嵌入，还支持对语音特征进行细粒度分析与可视化。





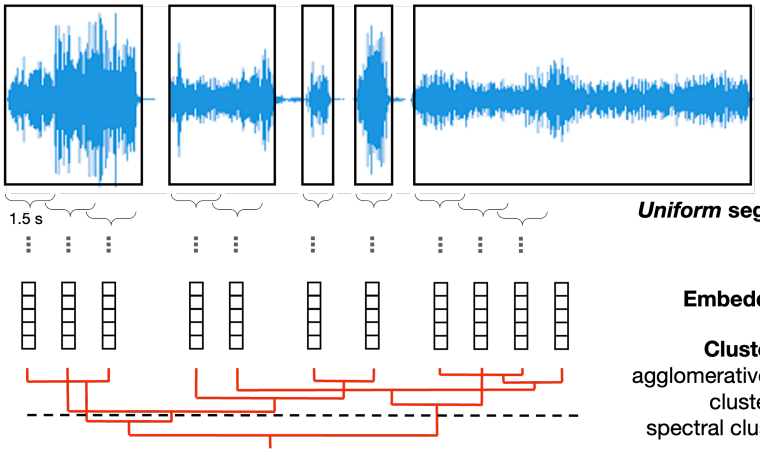
1. 预训练说话人嵌入作为额外输入
2. 联合训练学习任务特定嵌入
3. 隐式说话人建模

- 说话人日志（speaker diarization）又被称为说话人分割聚类
- 说话人日志要解决“什么时候谁在说话”的问题？
- 如下图所示，在一段音频中，A，B，C，D 四个人都说了话，任务的目标为：
  - 输出每个人每段音频的起始和结束位置
- 经过说话人分割聚类处理得到每个人说话的音频，更易于下游任务做进一步处理

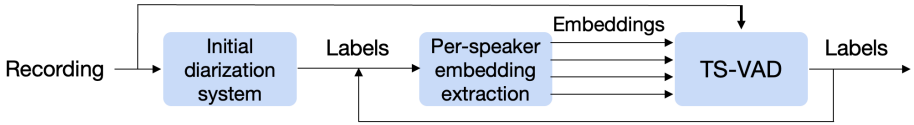


不同任务下的说话人建模

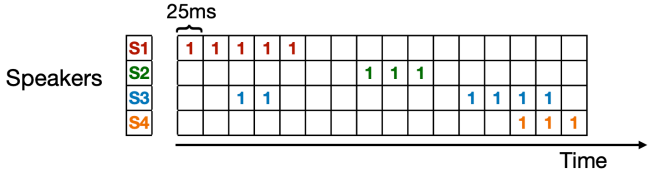
基于聚类的说话人分割聚类[41]



[41] Thanks to Mireia Diez Sánchez for the figure



Binary VAD decisions are made independently for each speaker per-frame with the same NN

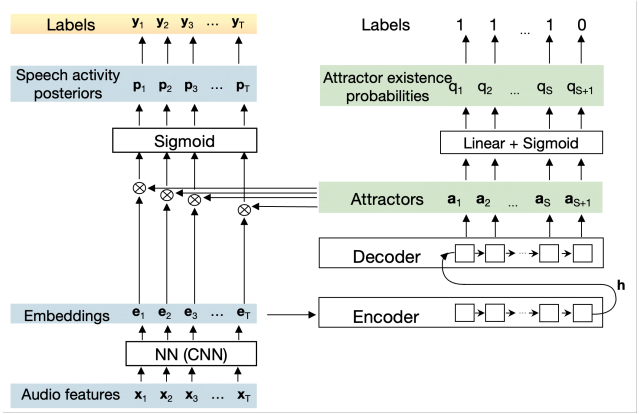


[42] Medennikov et al., "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario".

[43] Thanks to Mireia Diez Sánchez for the figure

不同任务下的说话人建模

基于 EEND-EDA 的说话人分割聚类[44,45,46]



[44] Fujita et al., "End-to-end neural speaker diarization with self-attention".

[45] Horiguchi et al., "Encoder-decoder based attractors for end-to-end neural diarization".

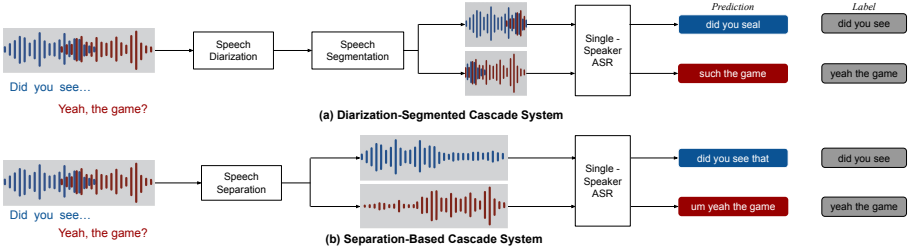
[46] Thanks to Mireia Diez Sánchez for the figure

- 多说话人 ASR：在同一段音频中，识别所有说话人的内容并进行归属（who says what）。
- 典型场景：会议、群体讨论、电话录音、课堂互动、播客访谈等。
- 与“鸡尾酒会效应”相关：在重叠语音中提取目标信息。

挑战

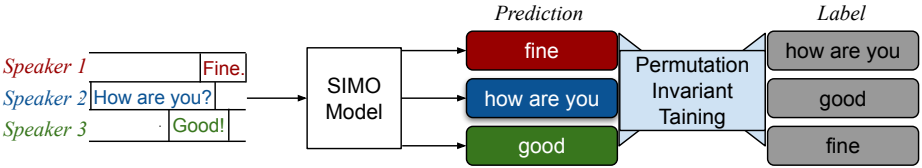
- 重叠语音：多个说话人同时说话导致的混叠。
- 说话人区分：不仅识别“说了什么”，还要准确地“谁说的”。
- 数据稀缺：大规模、细粒度标注的多说话人数据极少。
- 多子任务耦合：识别、分离、分段、归属、重叠/转写边界检测等。

[47] He and Whitehill, “Survey of End-to-End Multi-Speaker Automatic Speech Recognition for Monaural Audio”.

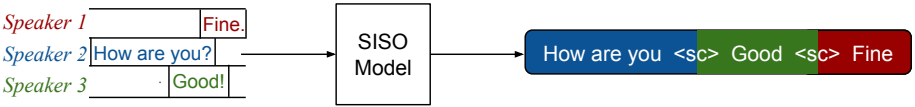


- Diarization-Segmented：先“说话人分段”，再单说话人 ASR。
- Separation-based：先“语音分离”，再单说话人 ASR。

[48]He and Whitehill, “Survey of End-to-End Multi-Speaker Automatic Speech Recognition for Monaural Audio”.



(a) Single-Input Multiple-Output Framework

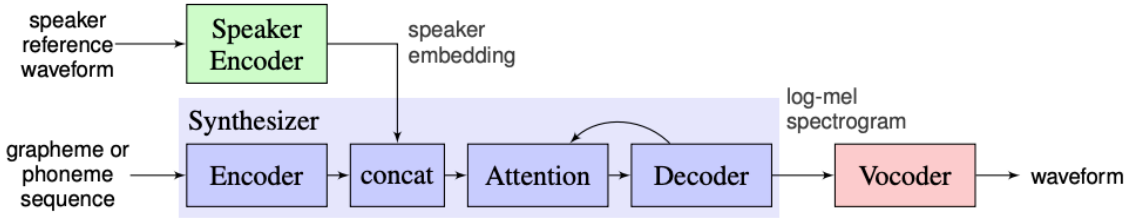


(b) Single-Input Single-Output Framework

- 直接从混合音频到说话人归属的转写，避免显式级联与误差累积。
- 支持联合优化“谁在说”与“说了什么”。
- 两大架构范式：**SIMO**（单输入多输出）、**SISO**（单输入单输出）。

不同任务下的说话人建模

例子：零样本 TTS 的显式说话人建模[49,50,51]



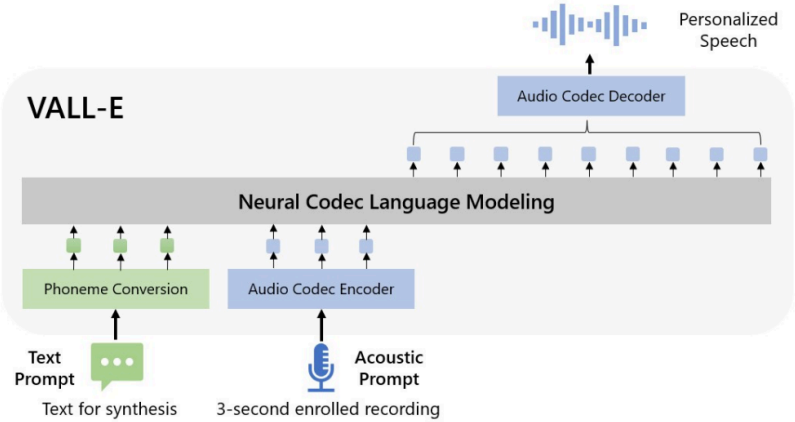
[49] Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis".

[50] Casanova et al., "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone".

[51] Wu et al., "Adaspeech 4: Adaptive text to speech in zero-shot scenarios".

不同任务下的说话人建模

例子：零样本 TTS 的隐式说话人建模[52,53,54]



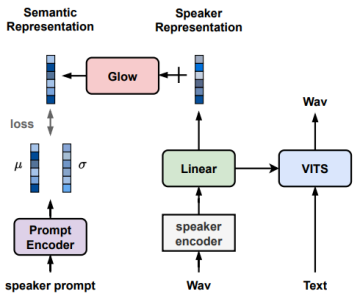
[52] Wang et al., "Neural codec language models are zero-shot text to speech synthesizers".

[53] Du et al., "UniCATS: A Unified Context-Aware Text-to-Speech Framework with Contextual VQ-Diffusion and Vocoding".

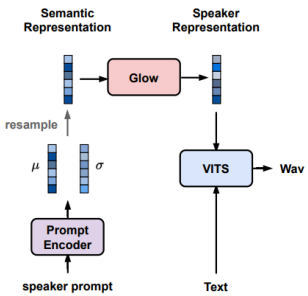
[54] Le et al., "Voicebox: Text-guided multilingual universal speech generation at scale".

不同任务下的说话人建模

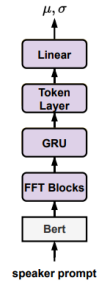
例子：走向可控性和新语音生成[55,56,57,58]



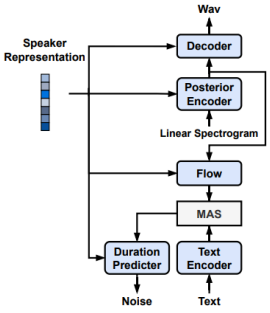
(a) training process



(b) inference process



(c) prompt encoder



(d) zero-shot VITS

[55] Zhang et al., "PromptSpeaker: Speaker Generation Based on Text Descriptions".

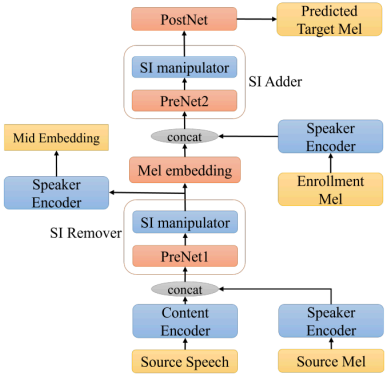
[56] Stanton et al., "Speaker generation".

[57] Shimizu et al., "PromptTTS++: Controlling Speaker Identity in Prompt-Based Text-to-Speech Using Natural Language Descriptions".

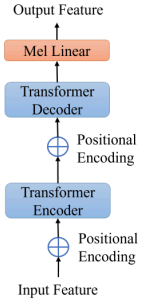
[58] Bilinski et al., "Creating new voices using normalizing flows".

不同任务下的说话人建模

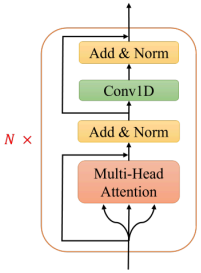
例子：零样本语音转换的显式说话人建模[59][60][61]



(a) Proposed System



(b) SI Manipulator



(c) Encoder and decoder architecture

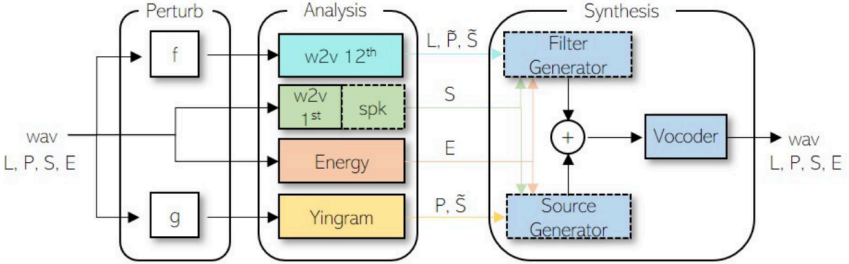
[59] Zhang et al., "SIG-VC: A Speaker Information Guided Zero-Shot Voice Conversion System for Both Human Beings and Machines".

[60] Chen and Duan, "ControlVC: Zero-Shot Voice Conversion with Time-Varying Controls on Pitch and Rhythm".

[61] Hussain et al., "ACE-VC: Adaptive and Controllable Voice Conversion Using Explicitly Disentangled Self-Supervised Speech Representations".

不同任务下的说话人建模

例子：零样本语音转换的隐式说话人建模[62,63,64]



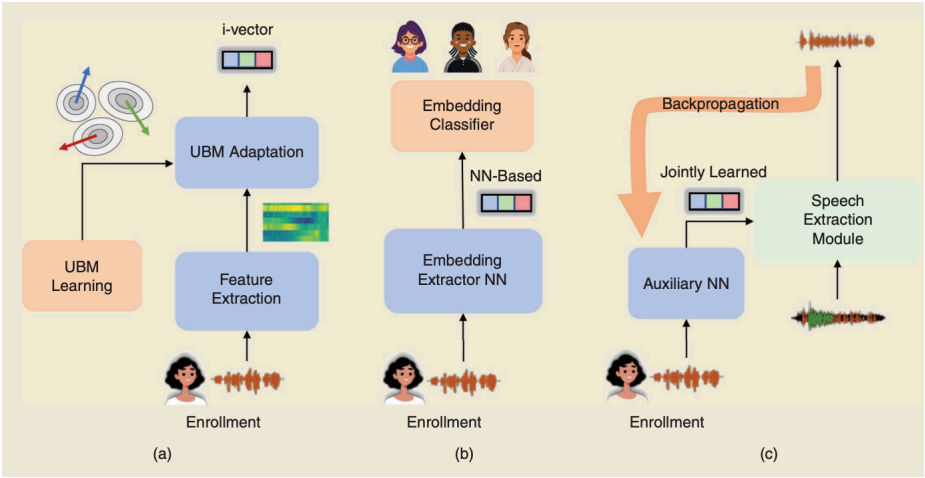
[62] Choi et al., "Neural analysis and synthesis: Reconstructing speech from self-supervised representations".

[63] Wu and Lee, "One-shot voice conversion by vector quantization".

[64] Wu, Chen, and Lee, "Vqvc+: One-shot voice conversion by vector quantization and u-net architecture".

不同任务下的说话人建模

例子：目标说话人提取的显式说话人建模[65][66][67]



[65] Zmolikova et al., "Neural Target Speech Extraction: An overview".

[66] Delcroix et al., "Single channel target speaker extraction and recognition with speaker beam".

[67] Ge et al., "Spex+: A complete time domain speaker extraction network".

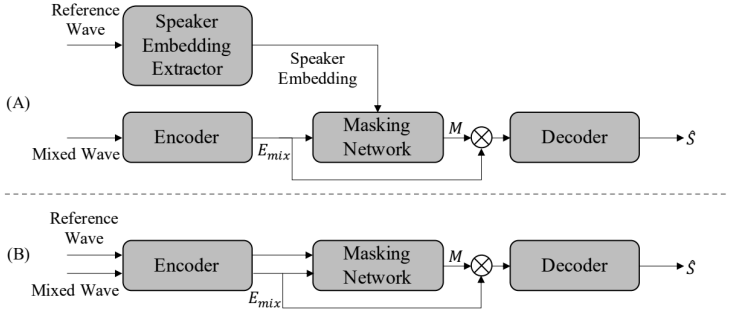


Figure 1: *(A) is the diagram of a typical time-domain target speaker extraction method. (B) is the diagram of our proposed method. $\otimes$ is an operation for element-wise product.*

[68] Zeng et al., "SEF-Net: Speaker Embedding Free Target Spekaer Extraction Network".

[69] Yang et al., "Target Speaker Extraction with Ultra-Short Reference Speech by VE-VE Framework".

不同任务下的说话人建模

"一个嵌入适用所有任务"的误区

普遍存在的假设

为识别任务优化的说话人嵌入在所有任务中都有效

现状：

- x-vector、ECAPA-TDNN、ResNet 被广泛采用
- 假设具有通用适用性
- 直接迁移到其他任务

问题所在

说话人识别 $\neq$ 语音合成 $\neq$ 语音转换 $\neq$ 目标说话人处理

关键问题

什么构成了"理想"的说话人表征？

不同任务下的说话人建模

核心困境

说话人识别

- 最大化判别性
- 最小化说话人内部方差
- 对噪声/信道的鲁棒性
- 紧凑的表征

生成任务

- 捕获细微细节
- 保持韵律/情感
- 实现自然合成
- 丰富、富有表现力的表征

目标说话人处理

- 相对判别性
- 最大化与混合语音中目标语音的相关性
- 紧凑的表征

冲突所在

判别性优化 **vs.** 生成性丰富
绝对判别性 **vs.** 相对判别性

- 说话人识别/验证：抑制说话人内部变异性
- 语音合成/转换：保持并利用这种变异性
- 目标说话人提取、目标说话人 ASR、说话人分离：在小集合内的相对判别性

挑战

生成听起来像目标说话人的自然、富有表现力的语音

关键要求：

- 音色准确性
- 韵律自然性
- 情感表现力
- 说话风格保持

证据：

- 定制编码器优于说话人识别嵌入
- 基于提示的方法主导零样本语音合成

GAP

说话人识别嵌入缺乏：

- 动态特征
- 韵律细节
- 情感线索
- 风格变化

Adriana STAN 等人 (2023) 的主要发现

1. 嵌入选择不影响学习过程
   - 网络会适应说话人条件，无论嵌入选择如何
   - 产生高质量的合成结果
2. 说话人信息泄露不可避免
   - 核心模块在标准训练中包含说话人信息
   - 简单条件化无法确保完全解耦
3. 零条件输出不一致
   - 核心网络学习相似表征
   - 零条件下说话人身份不稳定

[70]Stan and O'Mahony, "An analysis on the effects of speaker embedding choice in non auto-regressive TTS".

根本问题

$$\text{语音} = \text{说话人身份} + \text{内容} + \text{韵律} + \text{口音} + \text{情感} + \cdots$$

当前方法：

- 对抗训练
- 梯度反转层
- 多编码器架构
- 自监督表征

挑战：

- 完美解耦不可能
- 残留说话人信息
- 内容-说话人纠缠
- 韵律控制复杂性

开放问题

我们能否实现完美解耦？

范式转变

从预计算嵌入到自适应、上下文感知建模

传统方法：

- 预训练说话人嵌入
- 固定表征
- 有限的上下文感知
- 性能瓶颈

最新方法：

- USEF-TSE（无嵌入）
- 注意力机制
- 多级表征
- 基于 SSL 的特征

关键洞察

直接声学匹配可以超越抽象嵌入

便利性陷阱

易于使用，但往往次优

原因：

- 预训练模型的可用性
- 在说话人识别中的初步成功
- 便利性因素

代价：

- 次优性能
- 信息瓶颈
- 有限创新
- 任务不匹配

Evidence

USEF-TSE 优于基于嵌入的方法

YourTTS 定制编码器 > 说话人识别嵌入

"平均化"问题

说话人识别嵌入将多样化的声学表现压缩为单点

丢失的内容：

- 情感变化
- 韵律模式
- 语速变化
- 风格细微差别
- 上下文相关特征

对应用的影响：

- 单调的语音合成输出
- 有限的语音转换表现力
- 情感控制差
- 不自然的韵律

关键问题

如何在保持判别性的同时保留说话人内部变异性？

当前说话人相关任务的局限性

根本复杂性

语音因子本质上是相互交织的，而非独立编码

纠缠关系：

- 基频轮廓：情感 + 语言结构
- 频谱特征：音色 + 语音内容
- 韵律：说话人 + 情感 + 内容
- 无明确边界

当前解决方案：

- 对抗训练
- 互信息最小化
- 多编码器架构
- 专门损失函数

现实

完美解耦仍然是一个开放的研究问题

音频 **LLMs (ALLMs)**:

- 受文本 LLMs (GPT, Qwen) 启发
- 在多样化音频任务上表现强劲:
  - 自动语音识别
  - 音频字幕
  - 音乐问答
- 卓越的泛化能力
- 能识别说话人属性 (性别、年龄、口音)

关键问题

ALLMs 能执行说话人验证吗？

观察

当前基于 ALLM 的系统在对话中仍然对说话人身份基本不敏感

[71] Ren et al., "Can Audio Large Language Models Verify Speaker Identity?"

关键思想：将说话人验证转换为基于音频的 QA 任务

四种提示策略：

1. 分离：两个音频段作为独立输入
   - 提示："音频 1：[audio1]，音频 2：[audio2]。它们来自同一个说话人吗？"
2. 连接：连接两个段
   - 提示："这个音频中有多少个说话人？"
3. 连接 + 静音：用 1 秒静音连接
   - 提示：与连接相同
   - 假设：静音有助于区分说话人
4. 混合：叠加两个段
   - 提示："这个音频包含两个混合轨道。同一个说话人？"

模型	性别	语言	年龄	设备	时长 <2s	时长 >6s
Kimi (C+S)	70.20	68.40	63.40	52.67	53.70	**73.60**
Qwen2 (C+S)	59.40	58.60	53.87	52.20	50.60	59.10
Step (C+S)	64.20	60.40	56.80	57.47	54.60	71.80

观察：

- 长时长话语上约 70% 准确率
- 在挑战条件下性能显著下降：
  - 跨设备：约 52–57%
  - 短时长：约 50–55%
- 选择 **Kimi-Audio with Concat + Silence** 进行微调

结论

零样本 **ALLMs** 的 **SV** 能力有限 → 激励微调方法

模型	性别	语言	年龄	设备	对话	时长 <2s	时长 >6s
Kimi (零样本)	70.20	68.40	63.40	52.67	55.00	53.70	73.60
Kimi (微调)	**95.07**	**97.00**	**92.40**	**88.20**	**89.00**	**80.90**	**89.50**
Kimi (随机采样)	94.80	93.07	92.27	85.60	80.53	77.00	89.00
ECAPA-TDNN	99.33	99.27	94.13	94.67	93.00	78.80	95.60

关键发现：

1. 巨大改进：性别维度上 70% → 95%
2. 困难对采样关键：相比随机采样持续提升
3. **ALLM** 在短时长上超越 **ECAPA-TDNN**！(80.90% vs 78.80%)
4. 在简单条件下仍有差距(例如，性别：95% vs 99%)

洞察

ALLMs 在挑战场景中表现出更强的鲁棒性，暗示在现实世界噪声环境中的潜力。

联合验证表述：

- 注册：[audio1]，测试：[audio2]，目标文本：“Hello world”
- 问题：“测试与注册是同一个说话人吗？测试匹配目标文本吗？”
- 答案：“说话人：是/否，内容：是/否”

在 **LibriSpeech** 上的评估：

模型	说话人准确率 (%)	文本准确率 (%)	总体准确率 (%)
Kimi (零样本)	62.09	89.61	52.31
Kimi (微调)	98.92	**99.95**	**98.87**
Whisper + ECAPA	**99.08**	99.75	98.83

Toolkit	Speaker-specific	SSL	Pre-trained Models	Deployment
Kaldi	No	No	No	No
VoxCeleb_Trainer	Yes	No	No	No
ASV-Subtools	Yes	No	No	Yes
SpeechBrain	No	No	No	No
NeMo	No	No	No	Yes
Espnet	No	No	Yes	No
3D-Speaker	Yes	Yes	No	No
Wespeaker	Yes	Yes	Yes	Yes

表: 常见的说话人建模的开源工具包

Dataset	Year	Speakers	Utterances	Duration
VoxCeleb1	2017	1,251	153,516	351h
VoxCeleb2	2018	6,112	1,128,246	2,442h
CN-Celeb1	2020	1,000	130,109	274h
CN-Celeb2	2020	2,000	529,485	1,090h
3D-Speaker	2023	10,000	579,013	1,124h
VoxBlink	2023	38,065	1,455,190	2,135h

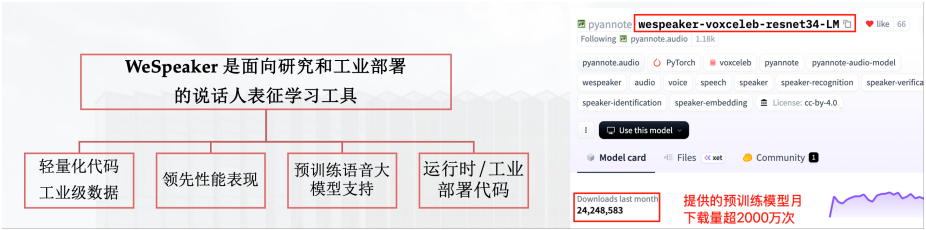
表: Representative Speaker Recognition Datasets

Performance Trends:

- VoxCeleb performance approaching saturation
- Need for more challenging scenarios
- Cross-genre and far-field datasets
- Large-scale unlabeled datasets for SSL

Wespeaker 是一个为研究和生产目的设计的说话人嵌入学习工具包，其特点是

- 轻量级代码库
- SOTA 性能
- 判别式和 SSL 范式
- 运行时/部署支持
- 被公司和学术机构的研究组采用：



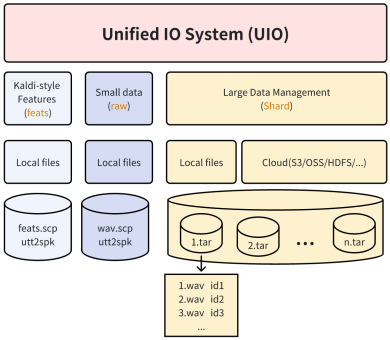


图: 统一 I/O 系统

统一 **I/O** 系统

- 也在 wenet ASR 工具包中采用
- 受webdataset和tfrecord启发

思想

- Raw: 从磁盘加载 wav 和标签文件 (小数据)
- Shard:
  - 将一组小文件打包成更大的分片
  - 实时读取和解压缩分片文件
- Feat: 兼容 kaldi 风格的特征文件
- 有效加载大规模数据集

步骤 1：下载并准备元数据

```
if [ ${stage} -le 1 ] && [ ${stop_stage} -ge 1 ]; then
  echo "Prepare datasets ..."
  ./local/prepare_data.sh --stage 2 --stop_stage 4 --data ${data}
fi
```

步骤 2：转换训练和测试数据

```
if [ ${stage} -le 2 ] && [ ${stop_stage} -ge 2 ]; then
  echo "Covert train and test data to ${data_type}..."
  for dset in vox2_dev vox1; do
    if [ $data_type == "shard" ]; then
      python tools/make_shard_list.py --num_utts_per_shard 1000 \
          --num_threads 16 \
          --prefix shards \
          --shuffle \
          ${data}/$dset/wav.scp ${data}/$dset/utt2spk \
          ${data}/$dset/shards ${data}/$dset/shard.list
    else
      python tools/make_raw_list.py ${data}/$dset/wav.scp \
          ${data}/$dset/utt2spk ${data}/$dset/raw.list
    fi
  done
fi
```

步骤 3：开始训练

```
if [ ${stage} -le 3 ] && [ ${stop_stage} -ge 3 ]; then
  echo "Start training ..."
  num_gpus=$(echo $gpus | awk -F ',' '{print NF}')
  torchrun --standalone --nnodes=1 --nproc_per_node=
        $num_gpus \
    wespeaker/bin/train.py --config $config \
      --exp_dir ${exp_dir} \
      --gpus $gpus \
      --num_avg ${num_avg} \
      --data_type "${data_type}" \
      --train_data ${data}/vox2_dev/${data_type}.list \
      --train_label ${data}/vox2_dev/utt2spk \
      --reverb_data ${data}/rirs/lmdb \
      --noise_data ${data}/musan/lmdb \
      ${checkpoint:+--checkpoint $checkpoint}
fi
```

数据集配置：

```
dataset_args:
  speed_perturb: True
  num_frms: 200
  aug_prob: 0.6
  # prob to add reverb & noise
        aug per sample
  fbank_args:
    num_mel_bins: 80
    frame_shift: 10
    frame_length: 25
    dither: 1.0
  spec_aug: False
  spec_aug_args:
    num_t_mask: 1
    num_f_mask: 1
    max_t: 10
    max_f: 8
    prob: 0.6
```

数据增强：

```
# add noise
dataset = Processor(dataset,
      processor.
      add_reverb_noise,
      reverb_data, noise_data,
       resample_rate, aug_prob
      )
# speed perturb
dataset = Processor(dataset,
      processor.speed_perturb,
       len(spk2id_dict))
# specaug
dataset = Processor(dataset,
      processor.spec_aug, **
      configs['spec_aug_args'
      ])
```

模型架构：

- ResNet 系列
- TDNN
- ECAPA-TDNN
- RepVGG
- CAM++
- ReDimNet
- Pretrained Frontend (e.g. WavLM)

池化方法：

- TSTP
- ASTP
- MQMHASTP

损失函数：

- add_margin
- arc_margin
- sphere
- sphereface2
- intertopk
- subcenter

模型配置：

```
model: ResNet34
    # ECAPA, CAMPPlus, REPVGG,
        ResNet152
model_args:
  feat_dim: 80
  embed_dim: 256
  pooling_func: "TSTP" # TSTP,
        ASTP, MQMHASTP
  two_emb_layer: False
projection_args:
  project_type: "arc_margin"
  # add_margin, arc_margin,
        sphere, sphereface2,
        softmax, aam_intertopk
  scale: 32.0
```

后端支持：

- Cosine
- LDA
- PLDA
- PSDA
- Adapt-PLDA

其他：

- 分数归一化
- 基于 QMF 的校准

评分：

```
if [ ${stage} -le 5 ] && [ ${stop_stage} -ge 5 ]; then
  echo "Score ..."
  local/score.sh \
    --stage 1 --stop-stage 2 \
    --data ${data} \
    --exp_dir $exp_dir \
    --trials "$trials"
fi

if [ ${stage} -le 6 ] && [ ${stop_stage} -ge 6 ]; then
  echo "Score norm ..."
  local/score_norm.sh \
    --stage 1 --stop-stage 3 \
    --score_norm_method $score_norm_method \
    --cohort_set vox2_dev \
    --top_n $top_n \
    --data ${data} \
    --exp_dir $exp_dir \
    --trials "$trials"
fi
```

Model	Params	vox1-O-clean	vox1-E-clean	vox1-H-clean
ReDimNetB0	1.0M	1.128	1.181	2.008
ReDimNetB3	3.2M	0.537	0.790	1.433
XVEC	4.61M	1.590	1.641	2.726
Res2Net34_Base	4.68M	1.234	1.232	2.162
ECAPA_TDNN_GLOB_c512	6.19M	0.782	1.005	1.824
RepVGG_TINY_A0	6.26M	0.824	0.953	1.709
Gemini_DFResNet114	6.53M	0.638	0.839	1.427
ResNet34	6.63M	0.659	0.821	1.437
ERes2Net34_Base	7.88M	0.744	0.896	1.603
CAM++	7.18M	0.659	0.803	1.569
ECAPA_TDNN_GLOB_c1024	14.6M	0.707	0.894	1.615
ResNet221	23.8M	0.505	0.676	1.213
SimAM_ResNet34 (VoxBlink2 Pretrain)	25.2M	0.372	0.559	0.997
ResNet293	28.6M	0.425	0.641	1.146
SimAM_ResNet100 (VoxBlink2 Pretrain)	50.2M	0.202	0.421	0.795
WavLM+EcapaTDNN		0.415	0.551	1.118

图: 支持模型列表

导出 Jit:

```
if [ ${stage} -le 7 ] && [ ${stop_stage} -ge 7 ]; then
  echo "Export the best model ..."
  python wespeaker/bin/export_jit.py \
    --config $exp_dir/config.yaml \
    --checkpoint $exp_dir/models/avg_model.pt \
    --output_file $exp_dir/models/final.zip
fi
```

导出 Onnx:

```
exp=exp  # Change it to your experiment dir
onnx_dir=onnx
python wespeaker/bin/export_onnx.py \
  --config $exp/config.yaml \
  --checkpoint $exp/avg_model.pt \
  --output_model $onnx_dir/final.onnx
```

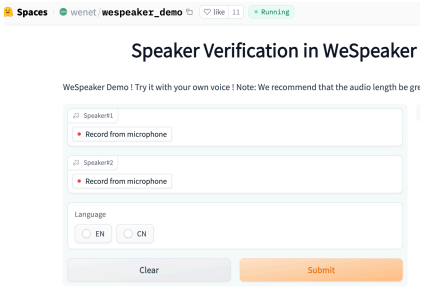


图: Wespeaker 演示页面

命令行使用：

```
wespeaker --task embedding --audio_file audio.wav --output_file
    embedding.txt -g 0
wespeaker --task embedding_kaldi --wav_scp wav.scp --output_file /path/
    to/embedding -g 0
wespeaker --task similarity --audio_file audio.wav --audio_file2 audio2.
    wav --g 0
```

Python 编程使用：

```
import wespeaker

model = wespeaker.load_model('chinese')
# set_gpu to enable the cuda inference, number < 0 means using CPU
model.set_gpu(0)
embedding = model.extract_embedding('audio.wav')
utt_names, embeddings = model.extract_embedding_list('wav.scp')
similarity = model.compute_similarity('audio1.wav', 'audio2.wav')
diar_result = model.diarize('audio.wav')
```

第一个面向目标语音提取的开源工具包.[72]

主要贡献

- 提出了WeSep开源工具包，专注于目标说话人提取
- 设计了多功能说话人建模能力
- 实现了在线数据模拟和可扩展性
- 提供了完整的训练和部署支持

技术特色

- 与 WeSpeaker 无缝集成
- 统一 I/O 数据管理机制
- 动态说话人混合策略
- 多种融合方法支持

[72] Wang et al., "Wesep: A scalable and flexible toolkit towards generalizable target speaker extraction".

UIO 框架

- 有效处理实验数据和生产级数据集
- 支持数万小时语音数据
- 处理大量小文件碎片
- 已在 WeNet 和 WeSpeaker 中集成

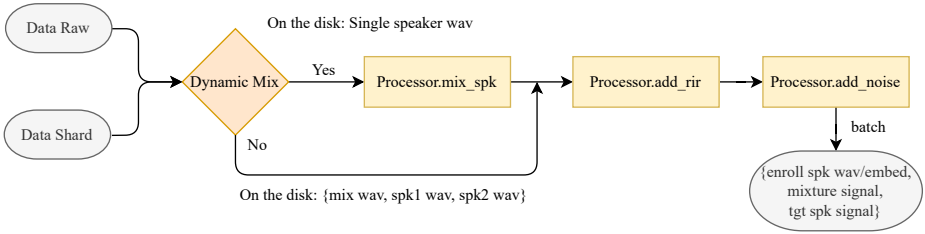


图: WeSep 在线数据准备管道（2 说话人示例）

在线数据模拟

传统方法问题：

- 预处理数据存储
- 硬盘空间消耗大
- 数据多样性有限

在线模拟优势：

- 节省存储资源
- 创建多样化训练数据
- 提高模型鲁棒性
- 灵活的数据生成

支持功能

- 在线噪声添加
- 混响生成（标准 RIR 采样 + 快速随机近似）
- 动态说话人混合策略

```
pip install git+https://github.com/wenet-e2e/wespeaker.git
```

```python
# psudo-codes for integrating wespeaker models
from wespeaker import get_speaker_model
# TDNN/ECAPA/ResNet/CAM++/WavLM...
s = get_speaker_model(spk_model_name)(**spk_args)
m = BSRNN(**sep_args) # Or other backbones
m.speaker_model = s
if use_pretrain_spk_encoder:
    m.spk_model.load_state_dict(pretrain_path)
    m.speaker_model.freeze()
```

```yaml
spk_fuse_type: 'multiply'
use_spk_transform: False
multi_fuse: False
joint_training: True
####### ResNet
spk_model: ResNet34
spk_model_init: False
#./wespeaker_models/model.pt
```

支持的骨干网络

1. **ConvTasNet**
   - 时域操作的卷积神经网络
   - 学习并估计分离掩码
   - 支持 Spex+ 变体
2. **BSRNN**
   - 频带分割循环神经网络
   - 显式分割频谱图到不同频带
   - 细粒度建模
3. **DPCCN**
   - 密集连接金字塔复卷积网络
   - 结合 DenseUNet、TCN 和 DenseNet 特征
   - 提升分离性能
4. **TF-GridNet**
   - T-F 域操作
   - 堆叠多路径块
   - 利用局部和全局频谱-时间信息

给定说话人嵌入 $\mathbf{e}_s$ 和中间输出 $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_T\}$：

1. **Concat**：直接复制 $\mathbf{e}_s$ 并拼接
2. **Add**：投影后逐元素相加
3. **Multiply**：投影后逐元素相乘
4. **FiLM**：特征级线性调制

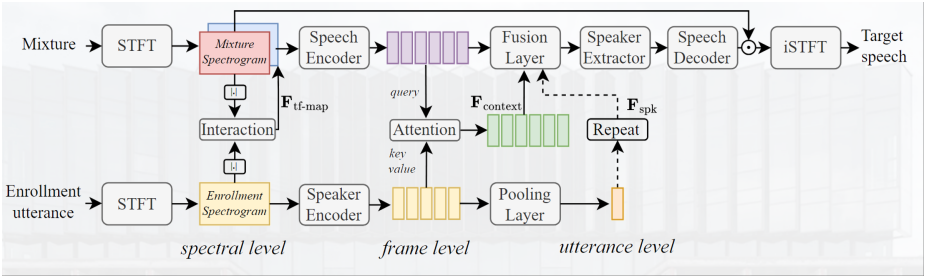
$$\mathbf{h}'_t = \gamma(\mathbf{e}_s) \odot \mathbf{h}_t + \beta(\mathbf{e}_s) \tag{3}$$

FiLM 优势

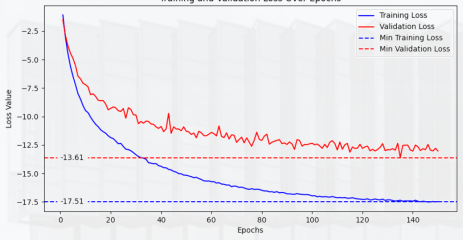
- γ 和 β 是说话人嵌入 $\mathbf{e}_s$ 的函数
- $\odot$ 表示逐元素乘法
- 学习到的仿射变换

除了基于嵌入的说话人引导，添加更细粒度的 Context 引导[a]

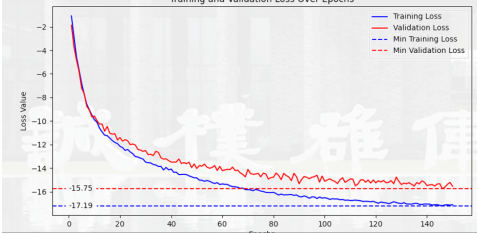
[a]Zhang et al., "Multi-level speaker representation for target speaker extraction".



Speaker embedding (Baseline)



TF-Map (Proposed)



Model	Speaker Model	Speaker Model	SI-SDRi on Libri2mix	Accuracy / %	Pub.
TD-SpeakerBeam	ResNet	Joint	13.03	95.2	ICASSP, 2020
SpEx+*	ResNet	Joint	13.41	-	Interspeech, 2020
sDPCCN	ConvNet	Joint	11.61	-	ICASSP, 2022
Target-Confusion*	ResNet	Joint	13.88	-	Interspeech, 2022
MC-SpEx*	ResNet	Joint	14.61	-	Interspeech, 2023
X-T-TasNet	d-vector	Pretrained	13.48	95.3	Interspeech, 2024
SSL-TD-SpeakerBeam	ResNet + WavLM	Pretrained	14.01	96.1	ICASSP, 2024
		Pretrained + Fine-tuning	14.65	97.0	
BSRNN	Campplus + SHuBERT	Pretrained	15.39	-	SPL, 2024
BSRNN	Ecapa-TDNN	Pretrained	**15.91**	**97.0**	Proposed
BSRNN			**17.99**	**98.6**	

- ✓ SOTA Performance, with simple but effective multi-level speaker modeling
- ✓ The generalization ability is largely enhanced (The gap between training and validation error)

- 技术演进：从传统 GMM-UBM 到深度学习的转变
- 多任务应用：说话人识别、分离、合成、转换等
- 技术挑战：鲁棒性、效率、可解释性、多模态融合
- 发展趋势：自监督学习、联合建模、工具化发展

- 超越识别：说话人建模不仅仅是说话人识别
- 超越嵌入：说话人建模不仅仅是嵌入学习
- 任务导向：需要根据具体任务定制建模方法
- 评估导向：需要针对性的评估指标和方法

- Baevski, Alexei et al. “wav2vec 2.0: A framework for self-supervised learning of speech representations”. In: Advances in neural information processing systems 33 (2020), pp. 12449–12460.
- Balian, Julien et al. “Small footprint text-independent speaker verification for embedded systems”. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2021, pp. 6179–6183.
- Belinkov, Yonatan and James Glass. “Analyzing hidden representations in end-to-end automatic speech recognition systems”. In: Advances in Neural Information Processing Systems 30 (2017).
- Bilinski, Piotr et al. “Creating new voices using normalizing flows”. In: (2022).

- Cai, Danwei, Weiqing Wang, and Ming Li. “An iterative framework for self-supervised deep speaker representation learning”. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2021, pp. 6728–6732.
- Cai, Danwei et al. “Pretraining Conformer with ASR for Speaker Verification”. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2023, pp. 1–5.
- Cai, Linjun et al. “CS-CTCSCONV1D: Small footprint speaker verification with channel split time-channel-time separable 1-dimensional convolution”. In: Proc. Interspeech 2022. 2022, pp. 326–330. DOI: 10.21437/Interspeech.2022-913.
- Cai, Weicheng, Jinkun Chen, and Ming Li. “Exploring the encoding layer and loss function in end-to-end speaker and language recognition system”. In: arXiv preprint arXiv:1804.05160 (2018).

- Caron, Mathilde et al. “Emerging properties in self-supervised vision transformers”. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, pp. 9650–9660.
- Casanova, Edresson et al. “Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone”. In: International Conference on Machine Learning. PMLR. 2022, pp. 2709–2720.
- Chen, Meiying and Zhiyao Duan. “ControlVC: Zero-Shot Voice Conversion with Time-Varying Controls on Pitch and Rhythm”. In: arXiv preprint arXiv:2209.11866 (2022).
- Chen, Sanyuan et al. “Unispeech-sat: Universal speech representation learning with speaker aware pre-training”. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2022, pp. 6152–6156.

- Chen, Sanyuan et al. “Wavlm: Large-scale self-supervised pre-training for full stack speech processing”. In: IEEE Journal of Selected Topics in Signal Processing 16.6 (2022), pp. 1505–1518.
- Chen, Ting et al. “A simple framework for contrastive learning of visual representations”. In: International conference on machine learning. PMLR. 2020, pp. 1597–1607.
- Chen, Zhengyang et al. “A comprehensive study on self-supervised distillation for speaker representation learning”. In: 2022 IEEE Spoken Language Technology Workshop (SLT). IEEE. 2023, pp. 599–604.
- Chen, Zhengyang et al. “Channel invariant speaker embedding learning with joint multi-task and adversarial training”. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2020, pp. 6574–6578.

- Chen, Zhengyang et al. “Large-scale self-supervised speech representation learning for automatic speaker verification”. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2022, pp. 6147–6151.
- Choi, Hyeong-Seok et al. “Neural analysis and synthesis: Reconstructing speech from self-supervised representations”. In: Advances in Neural Information Processing Systems 34 (2021), pp. 16251–16265.
- Chowdhury, Shammur Absar, Nadir Durrani, and Ahmed Ali. “What do end-to-end speech models learn about speaker, language and channel information? a layer-wise and neuron-level analysis”. In: Computer Speech Language 83 (2023), p. 101539.
- Chung, Joon Son et al. “In defence of metric learning for speaker recognition”. In: arXiv preprint arXiv:2003.11982 (2020).

- Chung, Joon Son et al. “In Defence of Metric Learning for Speaker Recognition”. In: Proc. Interspeech 2020. 2020, pp. 2977–2981. DOI: 10.21437/Interspeech.2020-1064.
- Dehak, Najim et al. “Front-end factor analysis for speaker verification”. In: IEEE Transactions on Audio, Speech, and Language Processing 19.4 (2010), pp. 788–798.
- Delcroix, Marc et al. “Single channel target speaker extraction and recognition with speaker beam”. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE. 2018, pp. 5554–5558.
- Desplanques, Brecht, Jenthe Thienpondt, and Kris Demuynck. “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification”. In: arXiv preprint arXiv:2005.07143 (2020).

- Du, Chenpeng et al. “UniCATS: A Unified Context-Aware Text-to-Speech Framework with Contextual VQ-Diffusion and Vocoding”. In: arXiv preprint arXiv:2306.07547 (2023).
- Fujita, Yusuke et al. “End-to-end neural speaker diarization with self-attention”. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE. 2019, pp. 296–303.
- Ge, Meng et al. “Spex+: A complete time domain speaker extraction network”. In: arXiv preprint arXiv:2005.04686 (2020).
- Han, Bing, Zhengyang Chen, and Yanmin Qian. “Self-Supervised Learning with Cluster-Aware-DINO for High-Performance Robust Speaker Verification”. In: arXiv preprint arXiv:2304.05754 (2023).
- — .“Self-supervised speaker verification using dynamic loss-gate and label correction”. In: arXiv preprint arXiv:2208.01928 (2022).

- He, Xinlu and Jacob Whitehill. “Survey of End-to-End Multi-Speaker Automatic Speech Recognition for Monaural Audio”. In: arXiv preprint arXiv:2505.10975 (2025).
- Horiguchi, Shota et al. “Encoder-decoder based attractors for end-to-end neural diarization”. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2022), pp. 1493–1507.
- Hsu, Wei-Ning et al. “Hubert: Self-supervised speech representation learning by masked prediction of hidden units”. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021), pp. 3451–3460.
- Huang, Wen et al. “Enhancing Cross-Domain Speaker Verification through Multi-Level Domain Adapters”. In: (2023).
- Huh, Jaesung et al. “Augmentation adversarial training for self-supervised speaker recognition”. In: arXiv preprint arXiv:2007.12085 (2020).

- Hussain, Shehzeen et al. “ACE-VC: Adaptive and Controllable Voice Conversion Using Explicitly Disentangled Self-Supervised Speech Representations”. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2023, pp. 1–5.
- Jia, Ye et al. “Transfer learning from speaker verification to multispeaker text-to-speech synthesis”. In: Advances in neural information processing systems 31 (2018).
- Jin, Yufeng et al. “Cross-modal distillation for speaker recognition”. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. 11. 2023, pp. 12977–12985.
- Jin, Zezhong, Youzhi Tu, and Man-Wai Mak. “Phonetic-aware speaker embedding for far-field speaker verification”. In: arXiv preprint arXiv:2311.15627 (2023).
- Le, Matthew et al. “Voicebox: Text-guided multilingual universal speech generation at scale”. In: arXiv preprint arXiv:2306.15687 (2023).

- Li, Fengfu, Bo Zhang, and Bin Liu. “Ternary weight networks”. In: arXiv preprint arXiv:1605.04711 (2016).
- Li, Jianchen, Jiqing Han, and Hongwei Song. “CDMA: Cross-Domain Distance Metric Adaptation for Speaker Verification”. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2022, pp. 7197–7201.
- Li, Jingyu et al. “Model Compression for DNN-based Speaker Verification Using Weight Quantization”. In: Proc. INTERSPEECH 2023. 2023, pp. 1988–1992. DOI: 10.21437/Interspeech.2023-1524.
- Li, Pengqi et al. “Reliable visualization for deep speaker recognition”. In: arXiv preprint arXiv:2204.03852 (2022).
- — .“Visualizing data augmentation in deep speaker recognition”. In: arXiv preprint arXiv:2305.16070 (2023).

- Li, Qingjian et al. “Towards Lightweight Applications: Asymmetric Enroll-Verify Structure for Speaker Verification”. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processi 2022, pp. 7067–7071. DOI: 10.1109/ICASSP43922.2022.9746247.
- Li, Rongjin, Weibin Zhang, and Dongpeng Chen. “The coral++ algorithm for unsupervised domain adaptation of speaker recognition”. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2022, pp. 7172–7176.
- Li, Ze, Ming Cheng, and Ming Li. “Enhancing Speaker Verification with w2v-BERT 2.0 and Knowledge Distillation guided Structured Pruning”. In: arXiv preprint arXiv:2510.04213 (2025).
- Liao, Dexin et al. “Towards a unified conformer structure: from asr to asv task”. In: ICASSP 2023. IEEE. 2023, pp. 1–5.

- Liu, Bei, Zhengyang Chen, and Yanmin Qian. “Depth-First Neural Architecture With Attentive Feature Fusion for Efficient Speaker Verification”. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023), pp. 1825–1838. DOI: 10.1109/TASLP.2023.3273417.
- Liu, Bei and Yanmin Qian. “Reversible Neural Networks for Memory-Efficient Speaker Verification”. In: Proc. INTERSPEECH 2023. 2023, pp. 3127–3131. DOI: 10.21437/Interspeech.2023-844.
- Liu, Bei, Haoyu Wang, and Yanmin Qian. “Extremely Low Bit Quantization for Mobile Speaker Verification Systems Under 1MB Memory”. In: Proc. INTERSPEECH 2023. 2023, pp. 1973–1977. DOI: 10.21437/Interspeech.2023-800.

- Liu, Bei et al. “Self-Knowledge Distillation via Feature Enhancement for Speaker Verification”. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processi 2022, pp. 7542–7546. DOI: 10.1109/ICASSP43922.2022.9746529.
- Ma, Yi et al. “ExPO: Explainable Phonetic Trait-Oriented Network for Speaker Verification”. In: IEEE Signal Processing Letters (2025).
- Medennikov, Ivan et al. “Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario”. In: arXiv preprint arXiv:2005.07272 (2020).
- Peng, Junyi et al. “Parameter-efficient transfer learning of pre-trained Transformer models for speaker verification using adapters”. In: ICASSP 2023. IEEE. 2023, pp. 1–5.

- Qian, Yanmin, Zhengyang Chen, and Shuai Wang. “Audio-visual deep neural network for robust person verification”. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021), pp. 1079–1092.
- Raj, Desh et al. “Probing the information encoded in x-vectors”. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE. 2019, pp. 726–733.
- Ren, Yiming et al. “Can Audio Large Language Models Verify Speaker Identity?” In: arXiv preprint arXiv:2509.19755 (2025).
- Reynolds, Douglas A et al. “Gaussian mixture models.”. In: Encyclopedia of biometrics 741.659-663 (2009).

- Rohdin, Johan et al. “Speaker verification using end-to-end adversarial language adaptation”. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2019, pp. 6006–6010.
- Shimizu, Reo et al. “PromptTTS++: Controlling Speaker Identity in Prompt-Based Text-to-Speech Using Natural Language Descriptions”. In: arXiv preprint arXiv:2309.08140 (2023).
- Snyder, David et al. “X-vectors: Robust dnn embeddings for speaker recognition”. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE. 2018, pp. 5329–5333.
- Stan, Adriana and Johannah O’Mahony. “An analysis on the effects of speaker embedding choice in non auto-regressive TTS”. In: arXiv preprint arXiv:2307.09898 (2023).

- Stanton, Daisy et al. “Speaker generation”. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2022, pp. 7897–7901.
- Tao, Ruijie et al. “Self-supervised speaker recognition with loss-gated learning”. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2022, pp. 6142–6146.
- Variani, Ehsan et al. “Deep neural networks for small footprint text-dependent speaker verification”. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE. 2014, pp. 4052–4056.
- Wan, Li et al. “Generalized end-to-end loss for speaker verification”. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2018, pp. 4879–4883.

- Wang, Chengyi et al. “Neural codec language models are zero-shot text to speech synthesizers”. In: arXiv preprint arXiv:2301.02111 (2023).
- Wang, Haoyu et al. “Adaptive Neural Network Quantization For Lightweight Speaker Verification”. In: Proc. INTERSPEECH 2023. 2023, pp. 5331–5335. DOI: 10.21437/Interspeech.2023-927.
- Wang, Shuai, Yanmin Qian, and Kai Yu. “What does the speaker embedding encode?” In: 2017.
- Wang, Shuai et al. “Discriminative neural embedding learning for short-duration text-independent speaker verification”. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 27.11 (2019), pp. 1686–1696.

- Wang, Shuai et al. “Knowledge Distillation for Small Foot-print Deep Speaker Embedding”. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processi 2019.
- Wang, Shuai et al. “Wesep: A scalable and flexible toolkit towards generalizable target speaker extraction”. In: arXiv preprint arXiv:2409.15799 (2024).
- Wu, Da-Yi, Yen-Hao Chen, and Hung-Yi Lee. “Vqvc+: One-shot voice conversion by vector quantization and u-net architecture”. In: arXiv preprint arXiv:2006.04154 (2020).
- Wu, Da-Yi and Hung-yi Lee. “One-shot voice conversion by vector quantization”. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2020, pp. 7734–7738.
- Wu, Xiaoliang et al. “Explainable attribute-based speaker verification”. In: arXiv preprint arXiv:2405.19796 (2024).

- Wu, Yihan et al. “Adaspeech 4: Adaptive text to speech in zero-shot scenarios”. In: arXiv preprint arXiv:2204.00436 (2022).
- Xia, Wei, Jing Huang, and John HL Hansen. “Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation”. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2019, pp. 5816–5820.
- Xu, J. et al. “Mixed Precision Low-Bit Quantization of Neural Network Language Models for Speech Recognition”. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021), pp. 3679–3693.

- Yang, Lei et al. “Target Speaker Extraction with Ultra-Short Reference Speech by VE-VE Framework”. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2023, pp. 1–5.
- Zeinali, Hossein et al. “But system description to voxceleb speaker recognition challenge 2019”. In: arXiv preprint arXiv:1910.12592 (2019).
- Zeng, Bang et al. “SEF-Net: Speaker Embedding Free Target Spekaer Extraction Network”. In: ().
- Zhang, Haoran, Yuexian Zou, and Helin Wang. “Contrastive self-supervised learning for text-independent speaker verification”. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2021, pp. 6713–6717.

- Zhang, Haozhe et al. “SIG-VC: A Speaker Information Guided Zero-Shot Voice Conversion System for Both Human Beings and Machines”. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2022, pp. 6567–65571.
- Zhang, Ke et al. “Multi-level speaker representation for target speaker extraction”. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processin IEEE. 2025, pp. 1–5.
- Zhang, Leying, Zhengyang Chen, and Yanmin Qian. “Knowledge Distillation from Multi-Modality to Single-Modality for Person Verification”. In: Proc. Interspeech 2021 (2021), pp. 1897–1901.
- — .“Knowledge Distillation from Multi-Modality to Single-Modality for Person Verification”. In: Proc. Interspeech 2021. 2021, pp. 1897–1901. DOI: 10.21437/Interspeech.2021-2119.

- Zhang, Yongmao et al. “PromptSpeaker: Speaker Generation Based on Text Descriptions”. In: arXiv preprint arXiv:2310.05001 (2023).
- Zhao, Zifeng et al. “Probing Deep Speaker Embeddings for Speaker-related Tasks”. In: arXiv preprint arXiv:2212.07068 (2022).
- Zheng, Rong, Shuwu Zhang, and Bo Xu. “Text-independent speaker identification using gmm-ubm and frame level likelihood normalization”. In: 2004 International Symposium on Chinese Spoken Language Processing. IEEE. 2004, pp. 289–292.
- Zheng, Siqi, Yun Lei, and Hongbin Suo. “Phonetically-Aware Coupled Network For Short Duration Text-Independent Speaker Verification.”. In: INTERSPEECH. 2020, pp. 926–930.
- Zhu, Tinglong, Xiaoyi Qin, and Ming Li. “Binary Neural Network for Speaker Verification”. In: Proc. Interspeech 2021. 2021, pp. 86–90. DOI: 10.21437/Interspeech.2021-600.

- Zhu, Tinglong, Xiaoyi Qin, and Ming Li. “Binary Neural Network for Speaker Verification”. In: Proc. Interspeech 2021. 2021, pp. 86–90. DOI: 10.21437/Interspeech.2021-600.
- Zmolikova, Katerina et al. “Neural Target Speech Extraction: An overview”. In: IEEE Signal Processing Magazine 40.3 (2023), pp. 8–29.