# ACE-VC：使用显式解耦的自监督语音表征实现自适应、可控的语音转换

*Shehzeen Hussain[1] *Paarth Neekhara[1] Jocelyn Huang[2] Jason Li[2] Boris Ginsburg[2]

[1]美国加利福尼亚州圣迭戈，加利福尼亚大学圣迭戈分校[2]美国加利福尼亚州圣克拉拉，英伟达公司

arXiv:2302.08137v1 [cs.SD] 16 Feb 2023

**摘要**

在本研究中，我们提出了一种使用通过自监督学习训练的语音表征的零样本语音转换方法。首先，我们开发了一个多任务模型，将一段语音话语分解为语言内容、说话人特征和说话风格等特征。为了将内容表示与说话人表示解耦，我们提出了一种基于孪生网络的训练策略，促使原始音频与移调音频的内容表示保持相似。随后，我们开发了一个带有音高和时长预测器的合成模型，能够根据分解后的表征有效地重建语音信号。我们的框架支持可控且自适应于说话人的合成，可执行零样本任意到任意语音转换，并在评估说话人相似度、可懂度和自然度的指标上取得了最先进的结果。仅使用目标说话人的 10 秒数据，我们的框架即可执行语音交换；对于已见说话人和未见说话人，说话人验证 EER 分别达到 5.5% 和 8.4%。[1]

索引词— 语音转换、表征学习、自监督学习、语音合成

## 1. 引言

语音转换的任务是修改源说话人的一段话语，使其符合目标说话人的声音特质。传统语音转换系统 [1, 2]依赖平行训练数据，即由多个说话人说出同一句话；而近年来，使用非平行多说话人数据集 [3, 4, 5, 6]训练的语音转换系统迅速涌现。非平行语音转换系统背后的核心思想，是将语音解耦为描述语言内容和说话人特征的表征。利用这些解耦特征合成语音，可以将给定话语的说话人嵌入替换为目标说话人的说话人嵌入，从而实现语音转换。

为从语音信号中解耦内容和说话人信息以进行语音转换，一些先前的研究 [7, 8]使用了预训练的自动语音识别（ASR）和说话人验证（SV）模型。ASR 模型预测的文本或提取的音素后验图（PPG）被视为内容表征，而从 SV 模型中提取的嵌入被视为说话人表征。尽管这种方法在语音转换方面展现出潜力，但仍存在一些局限：1）此类系统容易受到 ASR 错误的影响，可能导致某些词语发音错误或转换不准确。2）文本和 PPG 无法捕捉所有语言特征，例如口音、表达方式和说话人无关风格，因此合成语音听起来较为中性。

最近，研究人员 [9, 10, 11, 12] 提出了利用自监督学习（SSL）训练的模型输出合成语音的方法。尽管 SSL 表征与音系信息高度相关，但它们不一定能将说话人信息与语言内容解耦，并且可同时用于语音识别 [13, 14]和说话人验证 [15]。因此，为了提取解耦的语言内容，通常会对 SSL 表征进行量化，以便从语音话语中获得 伪文本 。这些方法的主要局限在于，对 SSL 表征进行量化属于有损压缩，从而导致次优的重建质量。此外，由于这种设置并未明确强制解耦，因此无法保证量化后的表征中不含说话人信息。

为解决上述语音转换系统的局限性，我们开发了语音表征提取器（SRE），以便从给定语音话语的自监督学习表征中更有效地分离说话人信息和语言内容。我们提出的 SRE 依赖于 Conformer-SSL 模型 [14]，该模型以多任务方式在两项下游任务——自动语音识别和说话人验证——上进行训练。为在训练过程中分离内容表征和说话人表征，我们通过音高偏移变换对给定语音的音色进行合成修改，然后使用带有余弦相似度损失的孪生网络设置 [16]，约束原始音频与变换后音频的内容嵌入保持相似。与以往的研究 [3, 4, 5, 6, 12], 相比，我们提出的方法使用单一神经网络（Conformer-SSL）骨干网络来提取内容信息和说话人信息。我们通过特征分

* 贡献相同。作为 NVIDIA[1]Audio 的实习生完成的工作，代码：https://paarthneekhara.github.io/ace

离来实现解耦，而不压缩或量化内容表征。

最后，我们构建了一个合成网络，用于根据 SRE 表征重建语音。我们的合成网络包含可学习的中间模型，能够根据内容表征和说话人表征预测基频和词元时长。我们的合成框架支持生成语音的两种功能：使用真实音高和时长模仿源语音的韵律和节奏，或使用预测的音高和时长根据目标说话人嵌入进行调整。在评估合成语音的说话人相似度、可懂度和自然度的指标上，我们的语音转换框架在针对已见和未见说话人的零样本语音转换任务中均取得了最先进的结果。

## 2. 方法

我们的语音转换框架由三个主要组件构成，且分别进行训练：1）基于自监督学习的语音表征提取器（SRE）；2）上游梅尔频谱图合成器；3）HiFi-GAN 声码器 [17]。

### 2.1. 语音表征提取器（SRE）

SRE 的目标是从给定的音频波形中提取解耦的说话人和内容表征。为此，我们采用以自监督方式训练的 Conformer 模型 [14] 作为框架的骨干网络。Conformer 架构结合了卷积和多头自注意力模块，从而构成一个计算效率高、能够建模局部和长期依赖关系的网络。该网络以音频的梅尔频谱表征为输入，并通过对比损失训练，以预测掩码输入的量化值。给定由一系列帧构成的梅尔频谱图 $x = (x_1, \ldots x_T)$，Conformer $E$ 输出一系列向量$z = E(x) = (z_1 \ldots z_{T'})$。

为了从预训练的 Conformer SRE 骨干网络中提取语言内容和说话人表示，我们添加两个随机初始化的下游头，并执行多任务训练，以进行语音识别和说话人验证：

**1）语音识别：**对于语音识别，下游头由两个线性层组成。第一个线性层将 SRE 骨干网络的输出映射为内容嵌入序列 $z_c = f_{\theta_{c1}}(z) = z_{c1} \ldots z_{cT'}$。第二个线性层后接 softmax，将每个时间步的内容嵌入映射为语言标记上的概率分数，即 $p_c = f_{\theta_{c2}}(z_c) = p_{c1} \ldots p_{cT'}$。语音识别下游头使用音频-文本对 $(x, y_c)$ 通过 CTC 损失 [18] 进行训练。即 $L_{content} = CTCLoss(p_c, y_c)$**2）说话人验证（SV）：**说话人验证头也由两个线性层组成。由于整个话语应映射为单个说话人表征，因此我们将 Conformer 在第一个时间步的输出映射为说话人嵌入 $z_s = f_{\theta_{s1}}(z_1)$。最后，第二个线性层将说话人嵌入映射为数据集中各说话人的分数，即 $p_s = f_{\theta_{s2}}(z_s)$。SV 头使用角度 softmax 损失 [19]，在音频-说话人对 $(x, y_s)$ 上进行训练，即 $L_{SV} = AngularSoftMax(p_s, y_s)$

**多任务训练：**由于 VoxCeleb [20] 等大型说话人验证数据集没有文本转录，因此我们为每项任务使用独立的数据集和数据加载器。在每次训练迭代中，我们加载两项任务的小批次并计算任务特定损失。接下来，我们将两项损失合并，得到 $L_{multi} = L_{content} + \alpha L_{SV}$，并通过两个下游头和 Conformer 骨干网络进行反向传播，以更新参数。对于下游网络参数，我们使用学习率为 $10^{-4}$ 的 Adam优化器；对于 Conformer 骨干网络，学习率为 $10^{-5}$。我们为 Conformer 使用较低的学习率，以防止微调期间过拟合。

**内容与说话人表示的解耦：**在初步实验中（第 3.2 节），我们发现，如果仅按上述方式进行多任务训练，说话人信息可能会泄漏到语言内容嵌入 $z_c$ 中。这种信息泄漏并不令人意外，因为底层 SSL编码 $z$ 经过微调后，同时包含说话人信息和语言内容信息。若未进行适当解耦，合成网络在很大程度上会忽略说话人嵌入 $z_s$，因此无法有效实现语音转换。为应对这一问题，我们在训练过程中通过移调数据增强对给定话语 $x$ 的音高进行合成调整，从而获得移调音频 $x'$。接下来，我们使用孪生余弦相似度损失，促使原始音频（$z_c = f_{\theta_{c1}}(E(x))$）与移调音频（$z'_c = f_{\theta_{c1}}(E(x'))$）的内容嵌入保持相似，如下所示：

$$L_{disentangle} = 1 - S_{cos}(z_c, z'_c) \quad (1)$$

其中，$S_{cos}$ 是两个向量之间的余弦相似度。上述解耦损失被加入多任务训练损失中，得到 $L_{SRE} = L_{multi} + \beta L_{disentangle}$。上述损失项旨在使内容嵌入独立于声音或说话人，从而实现我们的解耦目标。图 1 概述了 SRE 训练。

### 2.2. 梅尔频谱图合成器

合成器的任务是根据 SRE 提供的表示重建真实梅尔频谱图。请注意，内容表示 $z_c$ 的时间特征和长度决定了给定话语的语速和总时长。因此，基于 SRE 原始输出训练的合成器只能确定

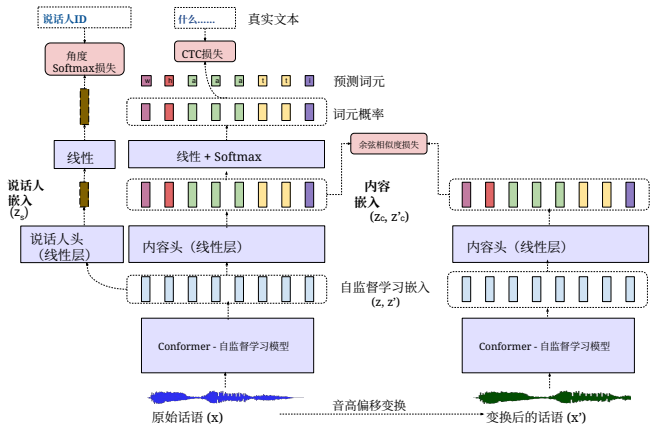


**图 1.** 语音表征提取器（SRE）的训练

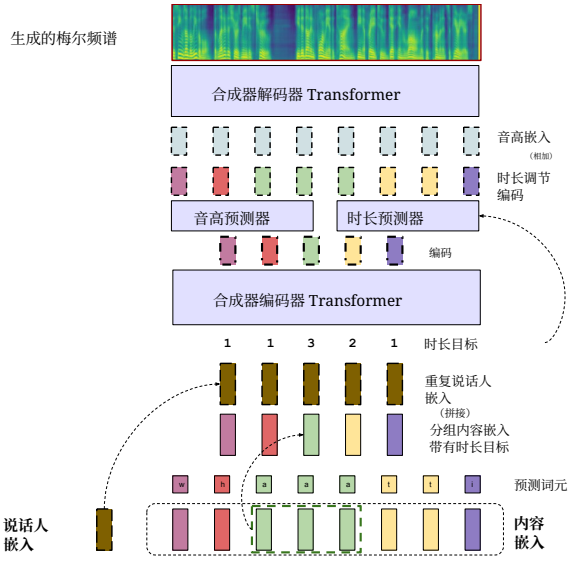


**图 2.** 从 SRE 表征合成梅尔频谱图

语速来自 $z_c$，而无法根据不同的目标说话人嵌入 $z'_s$调整语速。这一局限存在于语音转换领域的许多以往研究中 [10, 5]，其中转换后音频的时长和语速始终与源音频相同。

为了解决上述挑战，我们首先对原始内容表征 $z_c$ 进行处理：将预测词元相同的连续向量分组。考虑从时间步 $i$到 $j$、具有相同预测语言词元的连续内容向量 $z_{ci} \cdots z_{cj}$，即$argmax(p_{ci}) = \cdots = argmax(p_{cj})$。我们将这些连续向量分组，并沿时间维度取平均，得到时间步 $t$处的新向量 $g_{ct} = average(z_{ci} \cdots z_{cj})$。 $g_{ct}$ 的目标真实时长由分组后的向量数量决定，即 $d_{ct} = j - i + 1$。因此，对所有时间步执行此过程后，我们得到分组内容表征和目标时长，分别为 $g_c, d_c = G(z_c)$。

接下来，我们将音频的时长 $d_c$ 和基频 $p$ （基频轮廓）建模为经过处理的内容嵌入 $g_c$ 和说话人嵌入$z_s$的函数。为此，我们的合成器网络由两个前馈变换器 $F_e$ 和 $F_d$ 组成，与 Fast-Pitch [21], 类似，但处理的是输入 $g_c$ 和 $z_s$，而非文本。第一个变换器的隐藏表征用于预测时长和基频，即 $h = F_e(g_c, z_s)$。$\hat{d} = DurationPredictor(h), \hat{p} = PitchPredictor(h)$。接下来，将基频轮廓投影后，在隐藏表征 $h$ 的每个时间步上取平均，并将结果与 $h$ 相加，得到$k = h + PitchEmbedding(p)$。最后，按照真实时长 $d_c$ 对 $k$ 进行离散上采样，并将其输入第二个变换器，以得到预测的梅尔频谱图 $\hat{y} = F_d(DurationRegulation(k, d_c))$。图 2 展示了合成器模型。

合成器以无文本的方式在多说话人数据集上进行训练。真实基频轮廓 $p$ 由Yin算法 [22]得出。每个话语的基频轮廓均使用给定说话人的基频轮廓的均值和标准差进行归一化。这种针对每位说话人的归一化确保基频轮廓仅捕捉随时间变化的韵律变化，而不包含说话人的身份信息。模型通过优化三种损失进行训练——梅尔频谱重建误差、基频预测误差和时长预测误差。即，

$$L_{synth} = \|\hat{y} - y\|_2^2 + \lambda_1 \|\hat{p} - p\|_2^2 + \lambda_2 \|\hat{d} - d_c\|_2^2 \quad (2)$$

## 3. 实验与结果

### 3.1. 数据集与训练

**SRE：** 我们使用通过在整个 LibriSpeech 数据集上进行自监督学习训练的 Conformer 模型权重来初始化 SRE 的骨干网络，该数据集包含 960 小时的英语语音。为了在多任务设置下训练 SRE，我们使用 LibriSpeech 的 *train-clean-360* 子集进行语音识别（以字符作为语言标记），并使用 VoxCeleb-2 [20]数据集进行说话人验证。对于语音识别，我们使用时长在 4 秒至 16 秒之间的话语。对于说话人验证，在训练和推理时均使用 2 秒的话语切片。所有话语均重采样至 22050 Hz，并使用 80 个频带提取梅尔频谱图，快速傅里叶变换大小为=1024、窗口大小为=1024、跳跃长度为=256。采用这些超参数时，46 毫秒的语音对应一个内容向量。我们在 LibriSpeech 测试集上评估语音识别性能，并在 VoxCeleb-1 [2]的原始试验对上评估说话人验证性能。我们的多任务 SRE 在语音识别任务上的字符错误率（CER）为 2.9%，在说话人验证任务上的等错误率（EER）为 3.01%。

**合成器：** 梅尔频谱图合成器和 Hi-FiGAN 声码器在 *train-clean-360* 子集上进行训练，该子集来自

[2]https://www.robots.ox.ac.uk/ vgg/data/voxceleb/vox1.ht ml

| | 嵌入类型 | | |
|---|---|---|---|
| | 说话人嵌入 | 内容嵌入（不含 $L_{disentangle}$） | 内容嵌入（含 $L_{disentangle}$） |
| 说话人分类准确率 | 97.5% | 62.5% | **37.5%** |
| 自动语音识别字符错误率 | - | 2.9% | 2.9% |

**表1**。特征解耦评估。使用解耦损失训练的内容嵌入具有较低的说话人分类准确率，这表明说话人信息已从内容嵌入中有效去除。

LibriTTS，使用与 SRE 相同的短时傅里叶变换参数。有关实现和超参数的详细信息，请参阅首页链接的代码库。

### 3.2. 特征解耦

为了评估我们从内容表征中去除说话人信息的效果，我们仅使用内容表征训练说话人分类器。说话人分类准确率越低，表明解耦效果越好。我们通过训练三个说话人分类器进行消融研究，分别将以下三种不同的表征作为输入：1）说话人表示；2）未使用 $L_{disentangle}$ 训练的内容表示；3）使用 $L_{disentangle}$ 训练的内容表示。每个说话人分类器都是一个三层神经网络，每个隐藏层包含256个隐藏单元。分类器使用 LibriTTS 的 dev-clean子集中的40位说话人进行训练，并在每位说话人的5条未见过的语句上进行评估。表1所示的结果表明，使用由公式1给出的 $L_{disentangle}$训练的内容表征的分类器，说话人分类准确率最低，同时在文本识别任务中仍达到较低的字符错误率。这一结果表明，内容表征能够有效去除说话人信息，同时不损害其捕捉语言特征的能力。

### 3.3. 语音转换

进行语音转换时，我们使用合成模型，将任意给定源语句的内容嵌入与目标说话人的说话人嵌入相结合；两者均由 SRE 得出。目标说话人的嵌入根据该说话人 10 秒的语音估算。我们考虑两种语音转换场景：从一个已见说话人转换到另一个已见说话人（多对多），以及从一个未见说话人转换到另一个未见说话人（任意对任意）。对于已见说话人，我们使用 LibriTTS数据集 *train-clean-360* 子集中的留出语句；对于未见说话人，则使用*dev-clean* 子集。每种场景下，我们随机选择 20 位目标说话人（10 位男性和 10 位女性）。接着，我们选择 10 条源语句，每条分别来自 10 位不同的说话人。每种场景下共进行 200 次语音转换试验。

我们从三个方面评估合成语音：说话人相似度、可懂度和自然度。对于说话人相似度，我们使用独立的说话人验证模型计算合成话语和真实话语的说话人嵌入

| | 多对多 | | | 任意对任意 | | |
|---|---|---|---|---|---|---|
| 方法 | 说话人验证等错误率。字符错误率 | | 平均意见分 | 说话人验证等错误率。字符错误率 | | MOS |
| 真实数据 | 4.3% | - | $4.01 \pm 0.09$ | 4.2% | - | $4.04 \pm 0.09$ |
| AutoVC [4] | 24.3% | 20.4% | $2.87 \pm 0.12$ | 36.3% | 35.4% | $2.56 \pm 0.12$ |
| AdaIN-VC [3] | 17.4% | 27.4% | $2.60 \pm 0.12$ | 25.8% | 29.7% | $2.76 \pm 0.11$ |
| MediumVC [25] | 10.5% | 32.1% | $3.12 \pm 0.13$ | 21.4% | 35.4% | $3.09 \pm 0.12$ |
| FragmentVC [11] | 16.6% | 28.1% | $3.22 \pm 0.11$ | 23.5% | 39.7% | $3.05 \pm 0.12$ |
| S3PRL-VC [12] | 13.5% | 9.2% | $3.07 \pm 0.11$ | 21.3% | 8.9% | $3.17 \pm 0.13$ |
| YourTTS [6] | 9.5% | 5.8% | $3.59 \pm 0.09$ | 13。4% | 5.3% | $3.60 \pm 0.09$ |
| ACE-VC（本文方法） | **5.5%** | **2.7%** | $\mathbf{3.62 \pm 0.10}$ | **8.4%** | **2.8%** | $\mathbf{3.75 \pm 0.09}$ |

**表 2**。不同语音转换技术的比较。SV-EER 和字符错误率越低，分别表示说话人相似度和可懂度越高。MOS 越高（报告值附有 95% 置信区间），表示语音听起来自然度越高。

。随后，我们将合成话语与真实话语配对，为每个目标说话人构建数量相等的正样本对和负样本对（每种方法共 4000 对），以计算等错误率（SV-EER）。先前研究 [12, 25, 11] 普遍使用自动说话人验证指标来评估语音转换方法，并发现这些指标与人类感知到的说话人相似度高度相关。对于可懂度，我们使用预训练自动语音识别模型 [26] 转录生成话语和源话语，并计算两者之间的平均字符错误率。对于自然度，我们在亚马逊土耳其机器人上开展平均意见分（MOS）研究，请听众按 1 到 5 的量表为每条话语评分。每条话语由 2 位独立听众评分，因此每种方法共获得 400 次评价。

我们使用各方法官方提供的开源实现，在与我们相同的训练和测试基准上，将我们的框架与若干先前的语音转换方法进行比较。如表 2 所示，我们的模型在全部三项指标上均优于先前工作。从表 2 中较低的说话人验证等错误率可以看出，表现最佳的模型能够生成与目标说话人具有较高说话人相似度的转换后语音，同时提升自然度和可懂度。欢迎读者聆听我们在首页提供链接的音频示例。

## 4. 结论

在本研究中，我们提出了一种基于解耦语音表征的零样本语音转换框架。我们提出了一种语音表征提取器，它使用单个 Conformer-SSL 骨干网络，有效且明确地解耦内容和说话人信息，且无需对学习到的表征进行量化或压缩。我们的合成模型能够根据说话人嵌入和内容嵌入预测语速与基频轮廓，从而实现可控且自适应的语音转换模型。结果表明，在说话人相似度、自然度和生成语音的可懂度方面，我们的方法显著优于近期提出的语音转换模型及最先进的语音转换模型。

## 5. 参考文献


[1] D. Saito、K. Yamamoto、N. Minematsu 和 K. Hirose，“基于说话人空间张量表征的一对多语音转换”，发表于*Interspeech*，2011年。

[2] S. H. Mohammadi 和 A. Kain，《语音转换系统概述》，载于*Speech Communication*。Elsevier，2017年。

[3] J. Chou 和 H. Y. Lee，《通过实例归一化分离说话人和内容表示实现单次语音转换》，*Interspeech*，2019年。

[4]K. Qian、Y. Zhang、S. Chang、X. Yang 和 M. Hasegawa-Johnson，《仅使用自编码器损失的零样本语音风格迁移》，载于*ICML*。PMLR，2019年。

[5] H. S. Choi、J. Lee、W. Kim、J. Lee、H. Heo 和 K. Lee，《神经分析与合成：从自监督表示重建语音》，*NeurIPS*，2021年。

[6]E. Casanova、J. Weber、C.D. Shulby、A.C. Junior、E. G¨olge 和 M. A. Ponti，《Yourtts：迈向面向所有人的零样本多说话人文本转语音和零样本语音转换》，载于*ICML*。PMLR，2022年。

[7] L. Sun、K. Li、H. Wang、S. Kang 和 H. Meng，“无需平行数据训练的多对一语音转换声学后验图”，发表于*ICME*，2016年。

[8] X. Tian、J. Wang、H. Xu、E. S. Chng 和 H. Li，“使用非平行数据进行语音转换的平均建模方法”，发表于*Odyssey*，2018年。

[9]K. Lakhotia、E. Kharitonov、W.N. Hsu、Y. Adi、A. Polyak、B. Bolte、T. Nguyen、J. Copet、A. Baevski、A. Mohamed 等，“基于原始音频的生成式口语语言建模”，计算语言学协会汇刊，2021年。

[10] A. Polyak、Y. Adi、J. Copet、E. Kharitonov、K. Lakhotia、W.N. Hsu、A. Mohamed 和 E. Dupoux，“基于离散解耦自监督表示的语音重合成”，发表于*Interspeech*，2021年。

[11] Y. Lin、C. M. Chien、J. H. Lin、H. Lee 和 L. S. Lee，《Fragmentvc：通过端到端提取并融合细粒度语音片段，实现任意对任意语音转换》，发表于*ICASSP*。IEEE，2021年。

[12]W. C. Huang、S. W. Yang、T. Hayashi、H. Y. Lee、S. Watanabe 和 T. Toda，《S3prl-vc：采用自监督语音表征的开放式语音转换框架》，发表于*ICASSP*。IEEE，2022年。

[13]A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli，《wav2vec 2.0：用于语音表征自监督学习的框架》，*NeurIPS*，2020年。[14] A. Gulati、J. Qin、C. Chiu、N. Parmar、Y Zhang、J. Yu、W. Han、S. Wang、Z. Zhang、Y. Wu 等，《Conformer：用于语音识别的卷积增强型Transformer》，*Interspeech*，2020年。

[15] S. Hussain、V. Nguyen、S. Zhang 和 E. Visser，《采用自监督学习的多任务语音激活框架》，发表于*ICASSP*。IEEE，2022年。

[16] J. Bromley、I. Guyon、Y. LeCun、E. S¨ackinger 和 R. Shah，《使用“孪生”时延神经网络进行签名验证》，*NIPS*，1993年。

[17] J. Kong、J. Kim 和 J. Bae，《HiFi-GAN：用于高效、高保真语音合成的生成对抗网络》，*NeurIPS*，2020年。

[18] A. Graves、S. Fern´andez、F. Gomez 和 J. Schmidhuber，《连接主义时序分类：使用循环神经网络为未分段序列数据标注》，发表于*ICML*。ACM，2006年。

[19] W. Liu、Y. Wen、Z. Yu、M. Li、B. Raj 和 L. Song，《Sphereface：用于人脸识别的深度超球面嵌入》，发表于*CVPR*。IEEE，2017年。

[20] A. Nagrani、J.S. Chung、W. Xie 和 A. Zisserman，《Voxceleb：野外大规模说话人验证》，《计算机语音与语言》，2020年。

[21] A. Ła´ncucki，《Fastpitch：通过基频预测实现并行文本转语音》，发表于*ICASSP*。IEEE，2021年。

[22] A. De Cheveign´e 和 H. Kawahara，《Yin：一种用于语音和音乐的基频估计器》，发表于《美国声学学会杂志》，2002年。

[23] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，《Librispeech：基于公有领域有声读物的自动语音识别语料库》，发表于*ICASSP*。IEEE，2015年。

[24] N.R. Koluguri、J. Li、V. Lavrukhin 和 B. Ginsburg，“Speakernet：用于文本无关说话人识别与验证的 1D 深度可分离卷积网络”，*arXiv:2010.12653*，2020年。

[25] Y.Gu、Z.Zhang、X. Yi 和 X.Zhao，“Mediumvc：利用合成的特定说话人语音作为中间特征实现任意到任意语音转换”，*arXiv:2110.02500*，2021年。

[26]S. Kriman、S. Beliaev、B. Ginsburg、J. Huang、O. Kuchaiev、V. Lavrukhin、R. Leary、J. Li 和 Y. Zhang，“Quartznet：采用 1D 时间通道可分离卷积的深度自动语音识别”，发表于*ICASSP*。IEEE，2020年。