分析和改进语音合成中的说话人相似性评估

Marc-André Carbonneau[1], *Benjamin van Niekerk*[1,2], *Hugo Seuté*[1], *Jean-Philippe Letendre*[1], *Herman Kamper*[2], *Julian Zaïdi*[1]

[1]Ubisoft La Forge, Canada

[2]Electrical and Electronic Engineering, Stellenbosch University, South Africa

marcandre.carbonneau@gmail.com,

Abstract

建模语音身份具有挑战性，因为其性质多方面。在生成式语音系统中，身份通常通过自动说话人验证（ASV）嵌入进行评估，这些嵌入是设计用于区分而非表征身份的。本文探讨了这种表示捕捉到的声音方面的哪些部分。我们发现，广泛使用的 ASV 嵌入主要集中在诸如音色和音域之类的静态特征上，而忽视了节奏等动态元素。我们还确定了一些混淆因素，这些因素会损害说话人相似度测量，并提出了缓解策略。为了解决这些问题，我们提出了一种评估说话人动态节奏模式的指标 U3D。这项工作对在不断改进的声音克隆系统背景下评估说话人身份一致性这一持续挑战做出了贡献。我们将我们的 代码公开发布。

Index Terms: 语音合成，计算副语言学

1. 介绍

生成虚拟人物的生成模型在过去几年里取得了巨大的进步。虚拟角色被期望具备独特且在时间上保持一致的身份特征。尽管最近有所进展，但在视觉 [1–4] 和文本基础系统 [5,6] 中，这仍然是一个挑战。本文探讨了如何在生成语音系统中衡量身份。难以定义声音身份是因为它涵盖了与解剖学相关的因素，如音域和音色，以及受说话者语言、口音和自然说话模式及节奏影响的行为因素 [7,8] 。然而，人类通常能在许多变化的条件下识别彼此，这表明存在一组可测量的标记。

从语音中表征身份主要被研究用于自动说话人验证（ASV）应用 [9–12]，其目标是区分说话人。这与建模说话人的身份 [13] 不同。对于说话人区分，模型编码了突出显示说话人间差异的一组标记最小的，而用于生成任务的表示应编码相关的身份标记所有。因此，ASV 模型可以决定忽略对说话人复制至关重要的细微语音身份因素（例如音调模式、节奏和口音）。

然而，判别性的 ASV 表示已被证明与人类的判断相关 [14]，成功用于在少样本语音合成系统中建模声音 [15–18]，并且可以用来对齐模型 [19]。此外，这些表示通常被用于衡量研究论文 [16, 18–25] 和社区挑战赛 [26, 27] 中合成系统的身份保持情况。

在这篇论文中，我们通过综合的视角来审视 ASV 表示，强调它们的优势、局限性和对语音生成的影响。我们首先回顾了语音中的身份标记，并确定用于衡量这些标记的指标。接下来，我们分析当前的 ASV 表示如何捕捉这些标记。我们的研究发现，ASV 嵌入主要编码频谱信息如音高范围和音色，而动态特征常常被忽略。

然后，我们确定了可能导致说话人相似性实验失败的混杂因素。具体来说，我们展示了诸如文件时长、信道噪声和均衡等因素可能会错误地导致来自同一说话人的语音被归因于不同的说话人。为了解决这个问题，我们提供了缓解策略和实验方案及结果解释指南。虽然我们的讨论限于语音合成，但我们的研究结果对说话人验证也有更广泛的影响。

为了解决 ASV 嵌入忽视身份动态方面的不足，我们提出了 U3D（单元持续时间分布距离），这是一种用于捕捉语音节奏模式的度量标准。通过基于音素内容并将其建模为分布而非统计矩来对节奏进行建模，我们表明 U3D 能够捕捉到比简单度量如语速更丰富的节奏信息。U3D 在自我监督的语音单元上建立持续时间模型。因此，它不需

要音素标注，使其成为语言无关，并且适用于许多场景。我们将一个用于声纹相似性测量的框架开源，实现了 U3D 和我们推荐的最佳实践 [1]。

2. 背景

2.1. 话语中的身份标志

基础语音研究 [7,8] 识别了人类相互辨识的两大类线索：解剖学因素和行为因素。

解剖因素主要与发声器官的物理结构相关。声腔的形态、形状和大小，以及肌肉的柔韧性，很大程度上定义了一个人的声音特征。虽然某些解剖特征相对固定（如鼻腔的尺寸），但其他部分——尤其是口腔和咽部——在语音生成过程中会持续改变以调节发音气流。这些物理特性决定了一个人的音高范围及其音色。音色构成了区分一个声音与其他声音的独特品质，即使词汇内容和音高保持不变 [28]。我们在第 3.2 节中的实验表明，在声纹嵌入中很好地表现了音色和音高，这表明它们在说话人区分中起着重要作用。

行为因素包括说话者习得的语音模式和习惯。这些因素除了受物理特性 [29] 影响外，主要与发声器官的使用方式有关。行为因素包括音素库存和发音习惯 [30]、自然节奏 [31,32]、词间独特的语流音变 [33]、个体化的发音模式 [34,35] 以及言语障碍。这些行为标志反映了年龄 [36]、时期 [37]、地区 [30,31]、社会因素 [38,39] 和个人特质等的影响，所有这些都是塑造说话人身份的因素。尽管在行为身份标志方面进行了大量研究，但仍不清楚如何为语音合成应用建模和测量这些标志。此外，对于不熟悉说话人的 [40,41] 人类评估者而言，这些复杂的身份标志极具挑战性，限制了听力测试在评估合成语音的音似度方面的有效性。这突显了需要包含解剖特征之外因素的自动化指标。

2.2. 测量合成研究中的说话人相似性

近年来，合成评估和实验协议的审查引起了越来越多的关注 [42–45]。尽管大多数努力集中在自然度和音质上，但仍有一些人提出了新的数据和工具用于相似性评估等更多方面 [46,47]。

在合成论文中报告说话人相似性有多种方法。一些作者使用预测模型来重现人类标注 [47–49]。但通常，说话人相似性是通过 ASV 嵌入之间的距离来测量的，这已经被证明与人类判断相关 [14]。一些作者直接报告真实语音和合成语音之间的平均相似度 [16,17,24,25]。然而，原始相似度数字难以解释，因为它们的范围在不同的 ASV 表示中变化，并且必须与同一说话人的真实示例之间的相似性进行比较。

这就是为什么许多研究报道了将合成语音与真实语音进行比较的声纹系统等错误率 (EER) [14]. 在实践中，报告 EER 相当于测量合成-真实语音对和真实-真实语音对之间的相似性。然后，设置一个决策阈值以产生相同的误接受率和误拒绝率。如果合成语音与真实语音有显著差异，它们的相似度得分将较低，从而允许设定一个能够完美分离且错误率为 0%的决策阈值。相反，如果合成语音与真实语音无法区分，则无法设定这样的阈值，使得决策类似于掷硬币，错误率为 50%。

3. 实验

虽然说话人验证嵌入在合成中广泛用于评估说话人相似性，但人们对它们实际编码的内容以及它们的稳健性知之甚少。我们探讨了一些广泛使用的说话人验证嵌入中包含哪些标记，并测量了混杂因素的影响。最后，我们提出了一种节奏指标并通过实验进行了验证。

3.1. 模型和数据集

我们的目标不是识别出最适合语音合成应用的 ASV 技术，而是对几种代表性方法进行观察。现代表示依赖于神经网络，这些网络将可变长度的发言总结为固定长度的嵌入向量。这些向量之间的余弦相似性表明了说话者之间的预测相似性。

我们将传统的但仍广泛使用的 ASV 方法，如 GE2E [12] 和 X-Vectors [9] 进行了比较。GE2E 使用对比损失来学习判别嵌入，而 X-Vectors 是基于

[1] https://github.com/ubisoft/ubisoft-laforge-spkrid

有限说话人集训练的分类器的内部表示。较新的 ResNet-TDNN [11] 和 ECAPA-TDNN [10] 模型在 X-Vectors 的基础上进行了架构改进，引入了残差连接、跳跃连接以及注意力机制用于池化。我们还包括依赖自监督学习进行特征提取的更近期的方法，因为它们越来越受欢迎，并且能够获得最先进的结果 [50]。

我们使用来自 SpeechBrain 的开源实现 [2] [51] 来进行 X-Vector、ECAPA-TDNN 和 ResNet-TDNN，以及 Resemble-ai 对 GE2E 的实现 [3]。最后，我们采用自监督方法，使用 WavLM [52] 作为特征提取器：WavLM+ECAPA-TDNN、WavLM Large+ECAPA-TDNN[4] 和 WavLM+X-Vector[5]。

我们在所有实验中使用了 ARCTIC [53] 和 L2-ARCTIC [54] 语音数据集。ARCTIC 包含 18 名英语母语者的发音记录，他们朗读了音素平衡的提示词。L2-ARCTIC 增加了 24 名非英语母语者，代表了各种口音，从而扩大了说话人的数量。我们选择这些数据集是因为它们具有多样性和高质量录音。在第 3.2 节的实验中，我们将数据集补充了来自 LibriSpeech [55] 的 train-clean-100 和 train-clean-360 子集中 1172 名说话人的数据，以增加多样性。

3.2. 由 ASV 嵌入捕获的标记器

我们首先探讨哪些身份标志被编码在 ASV 嵌入中。我们通过尝试从这些嵌入中预测捕捉不同身份方面的人工特征来实现这一点。表 1 汇总了这些特征，并提供了它们所测量内容的直觉理解。解剖特征，如音调和音色，由平均音调和谐噪比（HNR）以及 α-比率捕捉。更多的行为和动态语音模式通过说话速度、音调的标准差和响度，以及有声段和无声段长度来测量。我们使用 [59] 中的方法提取说话速度，因为它与真实的音节率高度相关。使用 REAPER[6]，我们仅在有声音段上计算平均值和标准偏差（std）的音调。其他特征是使用

[2] https://github.com/speechbrain/speechbrain

[3] https://github.com/resemble-ai/Resemblyzer

[4] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification

[5] https://huggingface.co/microsoft/wavlm-base-plus-sv

[6] https://github.com/google/REAPER

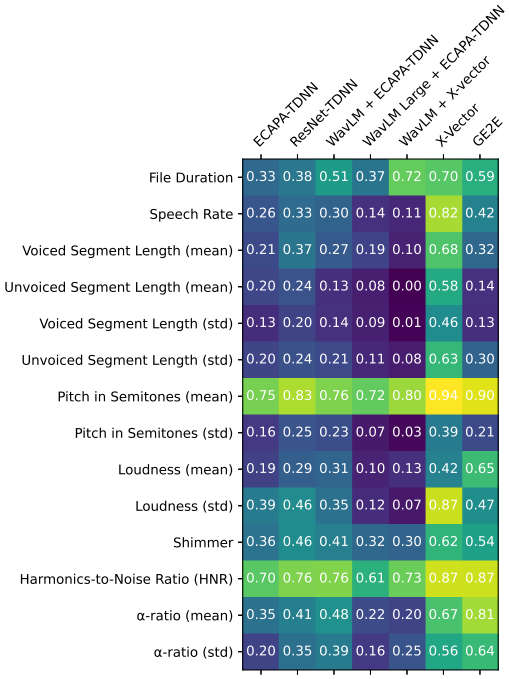

图 1: 决定系数 (r^2) 的特征预测。完美分数是 *1.0*，而不相关的预测得分为 *0.0*。

OpenSmile [60] 提取的。最后，我们还包括文件时长，这不应与说话人身份相关联。

我们提取了 LibriSpeech、ARCTIC 和 L2-ARCTIC 中每个语音的 ASV 嵌入。将这些嵌入作为输入，我们训练了一个 lasso 回归器来预测语音的特征。我们也尝试了一种非线性的随机森林回归器，结果与 lasso 相当；因此，我们仅报告了 lasso 的结果。我们使用四分位距 (IQR) 方法移除了异常值，并通过执行 5 折交叉验证优化了超参数，确保每位说话人只属于一个单独的折叠。在对所有数据重新训练回归器后，我们预测每个手工特征的值。最后，我们计算了特征值与其预测之间的决定系数 (r^2)。完全预测准确的特征得分为 1.0，而相关性为零的预测得分为 0.0。

图 1 显示说话人嵌入主要编码静态频谱信息（平均音高、HNR、颤抖度、α-比率）反映语音质量和频率范围。另一方面，嵌入未能捕捉动态行为身份标志如语速、浊音/清段时间长度、音高的变化和响度的变化。这一限制意味着嵌入仅表示了全面评估合成语音中身份所需的一组说话人特征的子集。

表 1: 提取的语音特征及其简短说明和它们所测量内容的直观理解。

标记物	描述	直觉
Duration	Length of the speech example.	This is independent from speaker identity.
Speech rate	Number of syllables per minute.	Indicates how fast a person speaks.
Voiced segment length	The average duration of a voiced segment.	Indicates how the person elongates vowel-like sounds.
Unvoiced segment length	The average duration of an unvoiced segment.	Indicates some pronunciation habits of the speaker.
Mean pitch	Average F_0 in semitones of voiced segments.	Indicates how high or low a person speaks.
Pitch std	Standard deviation of F_0.	An expressive speaker will have a high base pitch deviation.
Mean loudness	The average perceived energy in a speech signal.	Mostly indicates of the file normalization level.
Loudness std	Standard deviation of perceived energy.	Indicates if speaker is expressive.
Shimmer	Mean amplitude difference between consecutive F_0 periods.	Voice quality indicator; irregular vocal fold vibrations result in breathiness and can show poor phonation control [56].
Harmonics-to-Noise Ratio (HNR)	Energy ratio between harmonic and noise-like components.	Voice quality indicator; high HNR characterizes sonorant and harmonic voices, while low HNR denotes an asthenic voice and dysphonia [57].
α-ratio	Ratio of the energy from 50 – 1000 Hz and 1 – 5 kHz.	Reflects voice quality; bright voices have more energy in high frequencies as opposed to dark/mellow voices [58].

尽管我们的目标不是确定用于身份测量的最佳说话人嵌入，但我们发现 X-Vector 嵌入编码了更广泛的标记，包括更多的动态信息，这比最近的、在说话人验证任务中表现更好的嵌入要多。这与 [61] 的研究结果一致。这一略显意外的结果突出了说话人验证和语音合成之间的差异，并呼吁需要在一个更广泛的合成背景下对 ASV 嵌入进行比较。可能还需要开发更适合的身份表示。

我们的实验还表明，嵌入可以编码与身份无关的分散因素，如语音文件时长。我们现在更详细地探讨这一限制及其影响。

3.3. 自动说话人相似性评估的鲁棒性

在本节中，我们演示了基于 ASV 的比较如何可能因混杂变量而受到损害，特别是文件持续时间和通道特征。我们不想否定已建立的方法论，而是旨在指导研究人员避免在解释实验结果时可能出现的方法论错误。

这里我们在不同条件下测量了 ARCTIC 和 L2-ARCTIC 数据集上的等错误率（请参见第 2.2 节）。表 2 报告了所有说话人的平均等错误率及其标准差。作为参考，我们通过将每个说话人与其他说话人进行比较来报告等错误率，类似于语音验证任务。一个完美的 ASV 嵌入应获得 0%的等错误率。作为第二个参考实验，我们将每个说话人单独与他们自己进行比较：对于每个说话人，我们将语音文件随机分配到两组中，第一组充当真实，第二组充当合成的示例。由于来自同一个人，这两组应该是无法区分的，并且等错误率应为 50%。接下来，我们不是随机分配话语，而是按持续时间排序，将较短的话语分到一组，较长的话语分到另一组。然后我们测量区分较短和较长话语之间的等错误率。由于这些是同一说话人所说的内容相同的语句，因此等错误率应为 50%。

如预期的那样，使用同一说话人的控制实验得出所有方法的平均等错误率为 50%左右，并且在比较不同人时 EERs 在 0%到 5%之间。这验证了我们的协议并表明所有嵌入都能准确区分说话人。然而，当使用相同的语音片段但按持续时间排序时，所有方法的 EER 显著降低。这意味着这些 ASV 系统可以根据文件持续时间来区分语音片段，而这并不是身份的指标。这一点与图 1 的结果一致，表明 ASV 嵌入编码了持续时间。这种不希望的行为可能会使实验偏离轨道，如果合成示例的持续时间与真实示例相差很大：一个合成系统将表现不佳，并且系统之间的比较可能没有意义。幸运的是，通过仔细设计实验以确保语音片段的持续时间相似可以很容易地缓解这一点。在文本到语音评估的情况下，我们建议生成与真实语料库相同的文本语音片段，并采取通常的预防措施，即不使用训练期间使用的语音片段。

表 2: 不同实验中 *EER* 在 *ARCTIC* 和 *L2-ARCTIC* 说话人中的平均值和标准差。

实验	ECAPA-TDNN	ResNet-TDNN	WavLM +ECAPA-TDNN	WavLM Large +ECAPA-TDNN	WavLM +X-Vector	X-Vector	GE2E
One speaker 对阵 rest	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.05 ± 0.03	0.03 ± 0.01	0.02 ± 0.01
Same speaker	0.50 ± 0.03	0.50 ± 0.02	0.51 ± 0.02	0.50 ± 0.02	0.50 ± 0.03	0.50 ± 0.03	0.50 ± 0.03
Same speaker, short 对战 long	0.34 ± 0.04	0.35 ± 0.03	0.32 ± 0.03	0.33 ± 0.03	0.39 ± 0.02	0.30 ± 0.03	0.34 ± 0.03
SNR 40	0.47 ± 0.03	0.47 ± 0.03	0.47 ± 0.03	0.48 ± 0.03	0.46 ± 0.04	0.45 ± 0.05	0.42 ± 0.04
SNR 20	0.34 ± 0.06	0.32 ± 0.05	0.33 ± 0.06	0.38 ± 0.05	0.36 ± 0.07	0.21 ± 0.07	0.15 ± 0.06
SNR 0	0.05 ± 0.04	0.04 ± 0.04	0.05 ± 0.03	0.10 ± 0.05	0.06 ± 0.04	0.01 ± 0.01	0.01 ± 0.01
+emphasis	0.47 ± 0.04	0.48 ± 0.03	0.45 ± 0.02	0.43 ± 0.03	0.48 ± 0.03	0.47 ± 0.04	0.07 ± 0.03
+de-emphasis	0.38 ± 0.05	0.41 ± 0.05	0.37 ± 0.07	0.32 ± 0.06	0.44 ± 0.06	0.28 ± 0.07	0.01 ± 0.01
+emphasis +re-equalization	0.50 ± 0.02	0.49 ± 0.02	0.49 ± 0.02	0.44 ± 0.04	0.50 ± 0.03	0.50 ± 0.03	0.50 ± 0.02
+de-emphasis +re-equalization	0.50 ± 0.03	0.50 ± 0.02	0.49 ± 0.03	0.45 ± 0.05	0.50 ± 0.02	0.49 ± 0.02	0.50 ± 0.02

我们接下来研究与信道相关的混淆因素，特别是噪声和均衡效果。这对于合成评估非常重要，因为声音质量会因方法不同而有所差异。虽然干净、无噪声的语音是理想的，但质量上的差异不应影响说话人相似度测量。有些系统会产生降低真实语音音质的伪影，而其他系统则生成比注册录音更清晰的样本。这种后一种情况经常出现在声音克隆应用中，在这些应用中用户使用消费级设备录制注册样本。2023 年 Blizzard 挑战赛展示了这一现象，有几个参赛者产生的样本优于训练材料 [40]。为了评估噪声效果，我们逐步向每个说话人的语音样本引入白噪声，并在不同的信噪比（SNR）下比较相同发音时计算等错误率（EER）。表 2 的中间部分列出了各说话人平均的 EER。

类似地，我们通过应用由 $H(z) = 1 - \alpha z^{-1}$ 定义的流行强调和去强调滤波器（使用 $\alpha = 0.97$ 进行强调，或其逆运算进行去强调）来测量文件均衡的效果。这些滤波器修改了信号的频谱平衡，但没有改变人类感知到的身份。结果在表 2 的底部部分给出。

我们在信道噪声上的实验（表 2 中间部分）揭示了扰动后 EER 存在显著变化。在低 SNR 下，所有模型都将带有噪声的样本错误地分类为与它们对应的干净样本不同的身份。虽然 0 dB SNR 表示极严重的退化，但结果表明噪声对 ASV 系统测量的相似性分数有显著影响。这在比较输出质量不同的系统时具有重要意义，例如，在使用未受控环境收集的数据评估高性能语音克隆系统时。**我们建议研究人员对真实语音和合成语音都进行信噪比评估**以确保正确解释结果。

通过应用均衡（表 2 下部分）来改变录制的频谱平衡显著改变了大多数方法的等错误率。对于我们使用的 GE2E 实现，这甚至会导致灾难性的失败。实际合成系统由于各种原因可能会使语音样本变色，导致评估变得复杂。使用**我们建议重新均衡来自评估系统的样本**以获得与真实样本相似的频谱平衡。例如，在我们的实验中，通过使用简单的均衡匹配算法来减轻音调不平衡。我们首先估计真实音频样本 $S_{rr}(f)$ 以及强调/去强调音频样本 $S_{xx}(f)$ 的功率谱密度。我们计算校正均衡的频率响应为 $|G(f)|^2 = S_{rr}(f)/S_{xx}(f)$。然后，我们获得一个 16 带 FIR 图形均衡器滤波器 [62]，以匹配这种频率响应。此滤波器应用于强调/去强调样本进行重新均衡，确保其采用真实示例的音调特性。在表 2 的最后一行中，我们看到这将等错误率带回到预期的 50%。

3.4. 测量节奏

我们在第 3.2 节中已经表明，基于 ASV 的相似性评估仅捕捉了一部分身份标记。它们特别不擅长测量动态特征。为了更好地描述行为身份，我们引入了 U3D（单元持续时间分布距离），这是一种用于建模节奏和音素时长的新颖指标。首先，我们认为，基于说话速率或平均带声/非带声音段长度的粗糙方法不足以区分说话人之间的动态语音模式。为了解释这一点，图 2 显示了 L2-ARCTIC 数据集（包括来自不同背景的说话者）的真实音节

率和平均带声/非带声音段长度分布。在许多情况下，相邻的说话者对之间在音节速率上没有显示出统计学上的显著差异（顶部）。此外，具有相似音节速率的说话者可能会表现出明显不同的带声音段长度（底部），这表明自然节奏比单纯的语音速率要复杂得多。这些发现强调了需要一种更复杂的评估方法来考虑内容，并且相比第一或第二统计矩来说更具描述性。

我们提出比较不同音素类型和静默的持续时间分布以捕捉节奏的更细微方面。虽然音素持续时间分析需要强制对齐器——一种并非在所有语言中都普遍可用的工具——我们的 U3D 指标提供了一个更具可访问性的替代方案。通过计算无监督语音单元组（从 SSL 模型获得）的持续时间分布，U3D 提供了对节奏的更精细描述，建模不同声音和音素类型的持续时间变化。该方法为研究人员提供了一种强大且与语言无关的技术来量化说话人时间语音模式中的微妙变化。

U3D 的计算分为三个步骤：

发现类似音素的语音单元组： 根据 [59]，我们在从 HuBERT 模型中提取的语音单元 [23] 上使用凝聚层次聚类 [7]。结果树状图中的不同分支对应着广泛的音位样组群。例如，[59] 显示三个主要簇分别映射到响音、擦音和静默。参见他们的论文以获取可视化。

分割语音并计算持续时间分布： 首先，我们从所有语音发声中提取语音单元。然后，我们使用来自 [59] 的评分函数将它们划分成连续段的序列。最后，我们记录每个音素类群的段持续时间以获得分布。

测量分布之间的距离： 对于每个组，我们计算合成语音和真实语音的持续时间分布之间的 Wasserstein 距离。这些距离可以单独报告以进行细粒度分析，或者平均化以提供一个用于比较的单一数值。

为了验证我们的方法，我们首先使用通过强制对齐获得的真实音素和静音段来测量持续时间分布 [8]。我们将音素分为五类：元音、近音音素、鼻音、摩擦音和塞音。然后，我们在三种情况下计算 ARCTIC 和 L2-ARCTIC 说话者之间的平均距离：

1. **相同：** 比较单个说话者随机分割的两个子集之间的距离。这反映了我们度量的下界，我们预期在这里获得最小的距离。
2. **最近的：** 计算每个说话者与其最近的对应者的音节率之间的距离。此设置测试我们的指标是否能区分出用语速无法区分的说话者。
3. **随机：** 报告随机说话人对之间的距离。

[7] https://github.com/bshall/urhythmic

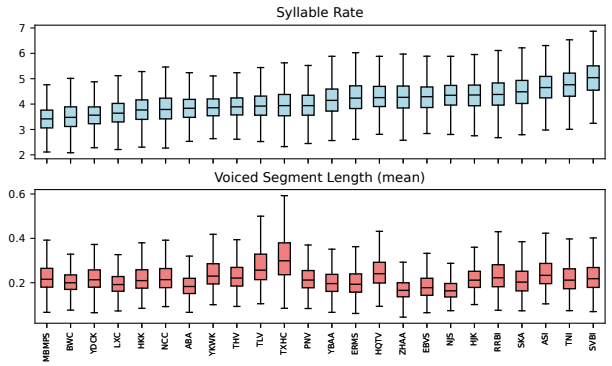

图 2: *所有 L2-ARCTIC 发言人的真实音节率和发音段长度。许多发言人的音节率相似，且发音段长度（底部）与语速（顶部）没有强烈相关性。*

表 3: *基于强制对齐和非监督语音单元（U3D）的* **相同***说话者之间的平均 Wasserstein 距离，以及根据音节率计算说话者与其* **最近的***邻居之间，以及* **随机***对说话者之间的距离。*

	approx.	fric.	nasal	stop	vowel	sil.	Avg.
强制对齐							
Same	1.7	1.5	1.7	1.3	1.4	7.3	2.48
Nearest	8.6	8.5	7.4	7.4	7.6	70.7	18.37
Random	11.7	13.9	12.9	11.5	15.0	81.6	24.43
无监督 (U3D)							
Same	2.3	1.0	1.5	1.2	1.0	5.9	2.15
Nearest	11.2	4.2	5.9	12.0	5.2	71.9	18.40
Random	15.4	7.0	10.7	15.6	9.1	71.4	21.53

表 3 中的顶部结果显示，不同说话者之间的节奏距离显著更大，即使那些语速相似的人也是如此。因此，我们的方法捕捉到了不能归因于语速的

[8] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

说话者之间有意义的区别。随后（底部部分），我们使用无监督语音单元组重复了实验，证实我们的方法不必依赖强制对齐。这提供了一种灵活的语言无关技术来量化说话者在时间上的细微变化。

4. 结论

在本文中，我们探讨了 ASV 嵌入在评估语音合成中的说话人身份方面的局限性。我们展示了 ASV 嵌入主要编码与解剖学相关的语音标识符（如音域和音色），而无法捕捉时间依赖的行为标识符。接下来，我们研究了嵌入对诸如样本时长、噪声水平和均衡等干扰因素的鲁棒性。我们提供了实验方案设计和有意义结果解释的建议。最后，为了解决基于 ASV 比较的不足，我们提出了 U3D 指标，通过测量节奏来更好地表征身份的动态方面。未来的工作包括针对行为标识符的附加指标、对说话人嵌入进行广泛的比较、专门用于合成评估的新表示方法，以及将讨论范围扩展到中性语音之外的内容。

5. References

[1] R. Gal et al., "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *ICLR*, 2022.

[2] N. Ruiz et al., "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *CVPR*, 2023.

[3] J. Huang et al., "ConsistentID: Portrait generation with multimodal fine-grained identity preserving," *Preprint arXiv:2404.16771*, 2024.

[4] G. Wang et al., "CharacterFactory: Sampling consistent characters with gans for diffusion models," *Preprint arXiv:2404.15677*, 2024.

[5] G. Lopez Latouche, L. Marcotte, and B. Swanson, "Generating video game scripts with style," in *ACL NLP4ConvAI Workshop*, 2023.

[6] S. Han et al., "Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances," in *NACL*, 2022.

[7] P. Ladefoged and D. Broadbent, "Information conveyed by vowels." *Journal of the Acoustical Society of America*, vol. 29, 1957.

[8] K. Johnson and M. Azara, "The perception of personal identity in speech: Evidence from the perception of twins' speech," 2000.

[9] D. Snyder et al., "X-Vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*, 2018.

[10] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, 2020.

[11] J. Villalba et al., "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations," *Computer Speech and Language*, vol. 60, 2020.

[12] L. Wan et al., "Generalized end-to-end loss for speaker verification," in *ICASSP*, 2018.

[13] I. Ulgen et al., "We need variations in speech synthesis: Sub-center modelling for speaker embeddings," *Preprint arXiv:2407.04291*, 2024.

[14] R. K. Das et al., "Predictions of subjective ratings and spoofing assessments of Voice Conversion Challenge 2020 submissions," in *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020.

[15] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *NeurIPS*, 2018.

[16] E. Cooper et al., "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP*, 2020.

[17] E. Casanova et al., "YourTTS: towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *ICML*, 2022.

[18] C-M. Chien et al, "Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech," in *ICASSP M2VoC Challenge*, 2021.

[19] S. Hussain et al., "Koel-TTS: Enhancing LLM based speech generation with preference alignment and classifier free guidance," *Preprint arXiv:2502.05236*, 2025.

[20] Q. Xie et al., "The Multi-Speaker Multi-Style Voice Cloning Challenge 2021," in *ICASSP*, 2021.

[21] W.Lin et al., "VoxGenesis: Unsupervised discovery of latent speaker manifold for speech synthesis," *Preprint arXiv:2403.00529*, 2024.

[22] S. Wang and É. Székely, "Evaluating text-to-speech synthesis from a large discrete token-based speech language model," in *LREC-COLING 2024*, 2024.

[23] B. van Niekerk et al., "A comparison of discrete and soft speech units for improved voice conversion," in *ICASSP*, 2022.

[24] M. Le et al., "Voicebox: Text-guided multilingual universal speech generation at scale," in *NeuRIPS*, 2023.

[25] S. Chen et al., "Neural codec language models are zero-shot text to speech synthesizers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, 2025.

[26] W.-C. Huang et al., "The Singing Voice Conversion Challenge 2023," in *ASRU*, 2023.

[27] J.-S. Bae et al., "Generative Data Augmentation Challenge: Zero-shot speech synthesis for personalized speech enhancement," in *ICASSP GenDA Workshop*, 2025.

[28] T. F. Cleveland, "Acoustic properties of voice timbre types and their influence on voice classification," *The Journal of the Acoustical Society of America*, vol. 61, no. 6, Jun. 1977.

[29] K. Johnson, "Speech production patterns in producing linguistic contrasts are partly determined by individual differences in anatomy," *UC Berkeley PhonLab Annual Report*, vol. 14, no. 1, 2018.

[30] M. C. Resnick, *Phonological Variants and Dialect Identification in Latin American Spanish*. De Gruyter Mouton, 1980.

[31] E. Torgersen and A. Szakay, "A study of rhythm in London: Is syllable-timing a feature of Multicultural London English?" *University of Pennsylvania Working Papers in Linguistics*, vol. 17, no. 2, 2011.

[32] D. Deterding, "The measurement of rhythm: a comparison of Singapore and British English," *Journal of Phonetics*, vol. 29, no. 2, 2001.

[33] M. Avanzi et al., "Sociophonetics of phonotactic phenomena in french," in *International Congress of Phonetic Sciences*, 2015.

[34] K. Johnson, P. Ladefoged, and M. Lindau, "Individual differences in vowel production," *The Journal of the Acoustical Society of America*, vol. 94, no. 2, 1993.

[35] J. S. Allen and J. L. Miller, "Individual differences in speech production: Voice-onset-time," *The Journal of the Acoustical Society of America*, vol. 108, no. 5, 2000.

[36] P. Foulkes and G. Docherty, "The social life of phonetics and phonology," *Journal of phonetics*, vol. 34, no. 4, 2006.

[37] E. R. Thomas and P. M. Carter, "Prosodic rhythm and African American english," *English World-Wide*, vol. 27, no. 3, 2006.

[38] P. Encrevé, "La liaison sans enchaînement," *Actes de la recherche en sciences sociales*, vol. 46, no. 1, 1983.

[39] J. Stuart-Smith, E. Lawson, and J. M. Scobbie, *Derhoticisation in Scottish English: A sociophonetic journey.* John Benjamins Publishing Company, 2014.

[40] O. Perrotin et al., "The Blizzard Challenge 2023," in *18th Blizzard Challenge Workshop*, 2023.

[41] ——, "Refining the evaluation of speech synthesis: A summary of the blizzard challenge 2023," *Computer Speech and Language*, vol. 90, 2025.

[42] E. Cooper, S. L. Maguer, E. Klabbers, and J. Yamagishi, "Good practices for evaluation of synthesized speech," *Preprint arxiv.2503.03250*, 2025.

[43] A. Kirkland et al., "Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation," in *Interspeech SSW Workshop*, 2023.

[44] P. S. Varadhan et al., "Rethinking MUSHRA: Addressing modern challenges in text-to-speech evaluation," *Preprint arXiv.2411.12719*, 2024.

[45] J. Camp, T. Kenter, L. Finkelstein, and R. Clark, "MOS vs. AB: evaluating text-to-speech systems reliably using clustered standard errors," in *Interspeech*, 2023.

[46] J. Ahn et al., "VoxSim: A perceptual voice similarity dataset," in *Interspeech*, 2024.

[47] G. Maimon, A. Roth, and Y. Adi, "Salmon: A suite for acoustic language model evaluation," in *ICASSP*, 2025.

[48] K. Deja et al., "Automatic evaluation of speaker similarity," in *Interspeech*, 2022.

[49] CH. Hu et al., "SVSNet: An end-to-end speaker voice similarity assessment model," *IEEE Signal Processing Letters*, 2022.

[50] J. Huh et al., "The VoxCeleb Speaker Recognition Challenge: a retrospective," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[51] M. R. et al., "SpeechBrain: A General-Purpose Speech Toolkit," *Preprint arXiv:2106.04624*, 2021.

[52] S. Chen et al., "WavLM: large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, 2022.

[53] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Interspeech SSW Workshop*, 2004.

[54] G. Zhao et al, "L2-ARCTIC: a non-native english speech corpus," in *Interspeech*, 2018.

[55] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[56] H. F. Wertzner, S. Schreiber, and L. Amaro, "Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders," *Revista Brasileira de Otorrinolaringologia*, vol. 71, 2005.

[57] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis – jitter, shimmer and HNR parameters," *Procedia Technology*, vol. 9, 2013.

[58] M.-J. Marsano-Cornejo and Ángel Roco-Videla, "Variation of the acoustic parameters: f0, jitter, shimmer and alpha ratio in relation with different background noise levels," *Acta Otorrinolaringologica (English Edition)*, vol. 74, no. 4, 2023.

[59] B. van Niekerk, M.-A. Carbonneau, and H. Kamper, "Rhythm modeling for voice conversion," *IEEE Signal Processing Letters*, 2023.

[60] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *ACM-MM*, 2010.

[61] R. Fu et al., "ASRRL-TTS: agile speaker representation reinforcement learning for text-to-speech speaker adaptation," *Preprint arXiv:2407.05421*, 2024.

[62] V. Välimäki and J. D. Reiss, "All about audio equalization: Solutions and frontiers," *Applied Sciences*, vol. 6, no. 5, 2016.