ClonEval：一个开放的语音克隆基准

Iwona Christop, Tomasz Kuczyski, Marek Kubis
Adam Mickiewicz University, Pozna
ul. Uniwersytetu Poznaskiego 4
61-614 Pozna, Poland
iwona.christop@amu.edu.pl, tomkuc2@st.amu.edu.pl, marek.kubis@amu.edu.pl

Abstract

我们提出了一种针对语音克隆文本转语音模型的新颖基准。该基准包括一个评估协议、一个用于评估语音克隆模型性能的开源库以及一个附带的排行榜。本文讨论了设计考虑因素，并详细描述了评估程序。软件库的使用进行了说明，以及排行榜上结果的组织方式。

1 介绍

随着能够克隆任意声音的语音合成模型（例如，Wang et al., 2023 ，Le et al., 2023 ）的出现，开发可靠且可重复的方法来评估语音克隆能力的需求不断增加。这对于促进模型之间的比较是必要的。为了解决这一需求，我们提出了 ClonEval，一种用于语音克隆的新基准。它包括：

1. 一个确定性的评估协议，为在语音克隆评估过程中使用的数据、指标和模型设定默认值。
2. 一个开源软件库，可用于以可复现的方式评估语音克隆模型。
3. 一个公共排行榜，可以对模型进行相互比较。

2 基准设计

为了开发用于语音克隆模型的基准，采用了以下设计原则：

1. 声音克隆模型应被视为不透明的，而不是假设任何拟议解决方案的特定架构。
2. 在评估过程中不应需要人工干预。
3. 评估程序应能够轻松适应新模型。
4. 评价结果应该是可再现的。

由于语音克隆技术快速发展，评估程序的设计应能够抵御未来文本到语音（TTS）模型中出现的技术变化。为了满足原则（1），对待评估模型施加的要求被降低到绝对最低限度。模型必须接受两个样本：一个要克隆的语音的音频样本和一个讲话文本样本。模型应生成单一的音频样本作为输出，代表语音克隆的结果。没有做出其他假设，例如提供置信度评分或 n -最佳列表的能力。确定平均意见分数（MOS）是对 TTS 质量进行主观评估的广泛认可的做法。然而，在开放基准的背景下采用这种做法会带来重大挑战。如果我们按照 (Chiang et al., 2024) 中提出的方法建立一个比较不同模型的场地，并要求人们投票，那么有必要精心挑选可供评估的模型列表。然而，这种方法带来了挑战，因为它阻止独立研究人员评估因商业或道德考虑而未公开的模型。或者，如果依赖模型作者报告的结果，可能会损害研究结果的可信度，因为验证报告结果的准确性需要重新进行人工评估，这是一个不可行的过程。因此，我们决定采用原则（2）。按照 Wang et al. (2023) 和 Le et al. (2023) 的方法，我们使用 WavLM (Chen et al., 2022) 来评估原始样本和克隆样本之间的相似性。WavLM 是一个预训练模型，利用其学习的通用语音表示在多种英语语音处理任务中实现有效的性能。为了实现原则（3）和（4），我们开发了 ClonEval，一个封装了整个评估过程的软件库。它可以被语音克隆模型的供应商用来独立测试他们的模型并报告其分数。如果供应商释放克隆样本，外部实体无需访问模型也可使用 ClonEval 库来重现报告的结果。

3 评估程序

图 1 中概述的评估程序包含两个阶段。首先，使用声音克隆模型生成样本，主要的克隆数据集是 LibriSpeech test-clean (Panayotov et al., 2015) 。该数据集目前是用于许多英语语音处理任务模型评估中最广泛使用的数据集之一。选择了另外四个流行的英语情感语音数据集来研究情感性对声音克隆质量的影响。选定的情感数据集是 CREMA-D (Cao et al., 2014) 、RAVDESS (Livingstone and Russo,

2018)、SAVEE (Haq and Jackson, 2010) 和 TESS (Pichora-Fuller and Dupuis, 2020)。如前所述，模型必须输入要克隆的声音样本和一个话语的文本样本。为了获取文本样本，对于每个声音样本，从给定数据集中所有可用的文本提示中抽取一个，排除与该声音样本对应的提示。

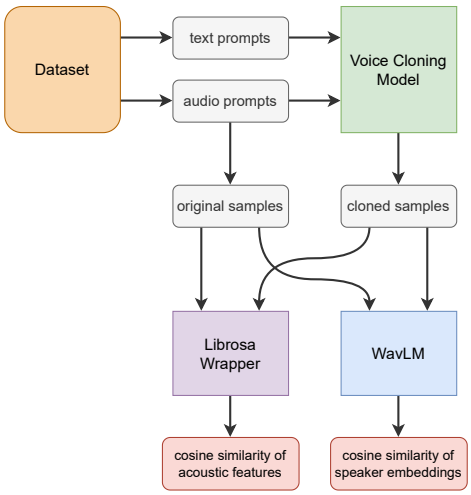

Figure 1: 评估过程概述。

通过语音克隆模型生成样本后，使用 WavLM 模型进行评估，每个样本都经过相同的处理。由于 WavLM 模型接受采样率为 16 kHz 的音频样本作为输入，因此需将音频重新采样为该值。随后使用 WavLM [1] 模型生成说话人嵌入。对于每对样本（参考和生成），计算它们从 WavLM 生成的说话人嵌入之间的余弦相似度。对于某个数据集中的所有样本，计算得到的相似度值进行平均，以获得最终评估结果。

为了进行细粒度错误分析，我们还使用 Librosa (McFee et al., 2015) 从每个样本中提取声学特征。提取的特征如下：

- 与 $f_{min} = 65$ Hz 和 $f_{max} = 2093$ Hz 的音高
- 声谱图
- 梅尔频谱图
- 带有 n_mfcc = 13 的梅尔频率倒谱系数（MFCCs），
- 均方根 (RMS)
- 谱质心
- 光谱带宽
- 谱对比
- 谱平坦度
- 频谱滚降
- 过零率
- 线性预测系数（LPCs）与 order = 2
- 节奏图
- 色谱图
- 伪常数 Q 变换
- 使用 IIR 滤波器的时频表示 (IIRT)
- 变 Q 变换
- 常数-Q 色谱图

。在适用的情况下，所用的提取参数如下：sampling_rate = 16 kHz、hop_length = 512、n_fft = 2048。如果未指定参数值，则使用 Librosa 包中的默认值。

4 软件库

用于评估过程的代码在 GitHub 仓库 [2] 中可用。要使用该代码，必须克隆该仓库，并安装必要的包。由于评估需要使用 WavLM-Large 模型，必须下载相关的检查点并将其放置在 `checkpoints` 目录中。

评估的输入文件必须组织成两个目录：一个包含参考样本，另一个包含生成的样本。每个目录都应包含用于正确比较的具有相同文件名的文件。如果评估程序要考虑情绪状态，每个文件名必须在 _ 后面包含相关情绪的名称，如 Listing 1 所示。

Listing 1: Example of proper organization of input sample directories.

```
original_samples/
    sample_1_anger.wav
    sample_2_neutral.wav
    sample_3_neutral.wav

cloned_samples/
    sample_1_anger.wav
    sample_2_neutral.wav
    sample_3_neutral.wav
```

一旦输入文件准备好，评估脚本应在指定相关目录作为参数的情况下执行，如 2 中所示。如果评估过程要考虑情感状态并按情感汇总结果，则应包含 `--evaluate_emotion_transfer` 标志。

[1] `microsoft/wavlm-base-plus-sv`

[2] https://github.com/amu-cai/cloneval

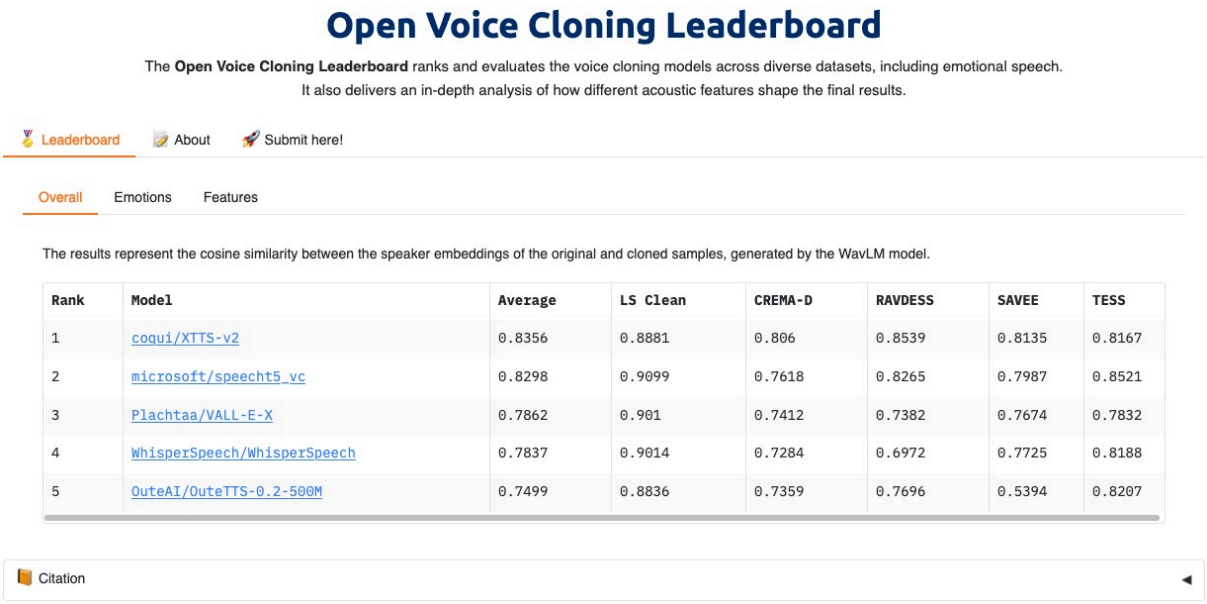

Figure 2: Hugging Face 的开放语音克隆排行榜概览。

Listing 2: Command used to run the evaluation script.

```
python eval.py \
    --original_dir <directory_A> \
    --cloned_dir <directory_B>
```

评估脚本在当前目录中生成两个输出文件。第一个文件 `results.csv` 包含每个文件对的详细指标。第二个文件 `aggregated_results.csv` 包含数据集的平均结果，如果适用，还按情感分类。

5 排行榜

使用 ClonEval 库进行的评估过程结果，在 Hugging Face 上的开放语音克隆排行榜 [3] 中展示（见图 2）。排行榜的主页包括模型获得的汇总结果，表示参考样本和生成样本的说话人嵌入之间的平均余弦相似度。这些结果是由 WavLM-Large 模型生成的。

除了总体结果外，排行榜还提供了两个选项卡：

- 情感：展示了每个情感语音数据集中由模型获得的说话者嵌入之间平均余弦相似度，分别按每个数据集和情感以及总计进行报告。
- 特征：展示每个数据集及整体由模型获得的选定声学特征值之间的平均余弦相似度。

已添加了一个名为“在此提交”的额外选项卡，以简化提交语音克隆模型供社区评估的过程。

[3] https://huggingface.co/spaces/amu-cai/Open_Voice_Cloning_Leaderboard

6 实验

我们进行了一系列实验，以评估利用 ClonEval 库的语音克隆模型。

6.1 整体上

表 1 中展示的结果显示了从参考样本中提取并通过每个模型生成的 WavLM 说话人嵌入之间的平均余弦相似度。XTTS-v2 (coqui, 2023) 在大多数数据集上取得了最高分，这也导致了整体最高分。在 LS test-clean 和 TESS 数据集上，SpeechT5 模型 (Ao et al., 2022) 显示出更好的性能。OuteTTS-0.2-500M (OuteAI, 2024) 、VALL-E X [4] (Zhang et al., 2023) 以及 WhisperSpeech (Collabora, 2024) 模型的得分略低。值得注意的是，大多数情况下余弦相似度值超过了 0.7，这表明所有模型都有效地从参考样本中克隆了声音，生成了类似的说话人嵌入。最好的结果是在 LS test-clean 数据集中获得的，表明模型在克隆非情感语音时表现更佳。

6.2 声学特征

表 2 显示了从参考样本中提取的声学特征与每个模型生成的特征之间的平均余弦相似度。OuteTTS 模型表现出最高的性能，尽管其他模型在大多数情况下获得的结果没有显著差异。

音高的相似性表明两个样本共享整体趋势，但在细节上存在明显差异，这表明它们可能由同一说话者发出，但文本内容可能不同，这一结论与评估程序的原则一致。此外，生成样本

[4] 使用了一个开源实现。可在 https://github.com/Plachtaa/VALL-E-X 获取

Dataset	OuteTTS	SpeechT5	VALL-E X	WhisperSpeech	XTTS-v2
LS test-clean	0.8836	0.9099	0.9010	0.9014	0.8881
CREMA-D	0.7359	0.7618	0.7412	0.7284	0.8060
RAVDESS	0.7696	0.8265	0.7382	0.6972	0.8539
SAVEE	0.5394	0.7987	0.7674	0.7725	0.8135
TESS	0.8207	0.8521	0.7832	0.8188	0.8167
Average	0.7499	0.8298	0.7862	0.7837	0.8356

Table 1: WavLM 模型生成的原始样本和克隆样本的说话人嵌入之间的余弦相似度。

Feature	OuteTTS	SpeechT5	VALL-E X	WhisperSpeech	XTTS-v2
pitch	0.6094	0.5278	0.5818	0.5863	0.5287
spectrogram	0.2609	0.2406	0.2451	0.2316	0.2298
mel spectrogram	0.9259	0.9109	0.9208	0.9091	0.8622
MFCCs	0.0486	0.0189	0.0494	0.0261	0.0352
RMS	0.6970	0.6040	0.6810	0.6400	0.6238
spectral centroid	0.7674	0.7855	0.7608	0.7485	0.7387
spectral bandwidth	0.9318	0.9379	0.9256	0.9238	0.9048
spectral contrast	0.9365	0.9423	0.9381	0.9359	0.9392
spectral flatness	0.3229	0.3091	0.3199	0.2655	0.2418
spectral roll-off	0.8198	0.8357	0.8134	0.8037	0.7907
zero-crossing rate	0.6115	0.6052	0.6057	0.5913	0.5785
LPCs	0.9720	0.9760	0.9747	0.9628	0.9808
tempogram	0.5159	0.5127	0.5030	0.5071	0.3658
chromagram	0.6036	0.6312	0.5966	0.5694	0.5775
pseudo-constant-Q transform	0.6707	0.6261	0.6499	0.6388	0.6242
IIRT	0.9550	0.9500	0.9559	0.9531	0.9520
variable-Q transform	0.9002	0.8910	0.8991	0.8881	0.8787
constant-Q chromagram	0.7167	0.7330	0.7146	0.6836	0.7138
WavLM	0.7499	0.8298	0.7862	0.7837	0.8356

Table 2: LS 测试-清洁数据集中原始样本和克隆样本所选声学特征值之间的余弦相似度。

中的失真可能会降低相似性的价值。频谱图的低相似性进一步支持了这一观点，因为音素的差异导致频率内容的不匹配。梅尔频谱图的高度相似性表明录音在感知尺度上高度对齐，暗示这些录音是由同一个讲话者产生的。MFCC 的相似性表明音色特征的重叠最小，指示信号在声学特性上有所不同，这可能是由于讲话内容或声学环境的变化所致。RMS 的相似值表明响度均匀，伴随有时机、重音和强度的变化，暗示信号可能来自同一讲话者，但内容不同。包括频谱重心、带宽、对比度、平坦度和截止频率在内的频谱特征的相似性值进一步指示说话者在频率分布上非常相似，噪音很小。零交叉率的中等相似性表明样本可能来自同一讲话者，但在不同条件下讲话。值得注意的是，LPC 几乎相同，表明它们的频谱包络高度一致，强烈表明这些信号来自同一讲话者。节奏图的相似性表明节奏结构部分对齐，暗示样本可能代表不同的内容。观察到色谱图和常量 Q 色谱图的高度相似性表明样本共享中等相似的谐波结构，暗示在共享谐波趋势下讲话内容有所变化。IIRT 的高度相似性表明信号在细粒度频率特性上几乎相同，暗示它们很可能来自同一讲话者，并且内容非常相似。伪常量 Q 变换和可变 Q 变换的相似值进一步暗示说话风格相似，但内容有一些变化。

如上所述，所有被考虑的声学特征的相似度值显示样本是由同一个人说的，但文本内容不同。这个发现与从 WavLM 获得的相似性一致。

6.3 情感

表 3 显示了从参考样本中提取的说话人嵌入与每个模型在给定情绪状态下生成的嵌入之间的平均余弦相似度。

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. *Preprint*, arXiv:2110.07205.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *Preprint*, arXiv:2403.04132.

Collabora. 2024. WhisperSpeech. `https://github.com/collabora/WhisperSpeech`.

coqui. 2023. XTTS-v2.

S. Haq and P. J. B. Jackson. 2010. *Multimodal Emotion Recognition*, pages 398–423. IGI Global, Hershey PA.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5):1–35.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25.

OuteAI. 2024. OuteTTS-0.2-500M.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

M. Kathleen Pichora-Fuller and Kate Dupuis. 2020. Toronto emotional speech set (TESS).

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *Preprint*, arXiv:2301.02111.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *Preprint*, arXiv:2303.03926.

Emotion	OuteTTS	SpeechT5	VALL-E X	WhisperSpeech	XTTS-v2
Anger	0.7197	0.7923	0.7623	0.7462	0.8098
Disgust	0.7034	0.8172	0.7600	0.7458	0.8325
Fear	0.6953	0.7996	0.7466	0.7601	0.7929
Happiness	0.7329	0.8068	0.7658	0.7462	0.8160
Neutral	0.7370	0.8322	0.7699	0.7714	0.8480
Sadness	0.7135	0.8099	0.7525	0.7516	0.8365
Average	0.7499	0.8298	0.7862	0.7837	0.8356

Table 3: 由 WavLM 模型为情绪状态生成的原始样本和克隆样本的说话人嵌入之间的余弦相似度。