ASDF：自动语音识别系统的差分测试框架

Daniel Hao Xian Yuen,* Andrew Yong Chen Pang,* Zhou Yang,† Chun Yong Chong,* Mei Kuan Lim,* David Lo†

*马来西亚莫纳什大学信息技术学院{dyue0003, apan0027,
chong.chunyong, lim.meikuan}@monash.edu
†新加坡管理大学计算与信息系统学院{zyang, davidlo}@smu.edu.sg

arXiv:2302.05582v1 [eess.AS] 11 Feb 2023

摘要——**近年来，自动语音识别（ASR）技术已在各个领域得到越来越广泛的应用。因此，评估和提升自动语音识别系统的质量至关重要。本文提出 ASDF，这是一种用于测试自动语音识别系统的自动语音识别差分测试框架。ASDF 扩展了现有的 ASR 测试工具 CrossASR++，该工具可从文本语料库合成测试用例。然而，CrossASR++未能有效利用文本语料库，且在如何通过失败测试用例改进自动语音识别系统方面提供的信息有限。为解决这些局限性，我们的工具加入了两项新功能：(1)** 文本转换模块，**用于增加生成的测试用例数量并发现自动语音识别系统中的更多错误；(2)** 语音分析模块，**用于确定自动语音识别系统容易在哪些音素上出错。ASDF 通过对失败测试用例中的文本应用各种文本转换方法（例如，改变时态），生成更多高质量的测试用例。这样，ASDF 就能利用小型文本语料库生成大量音频测试用例，而 CrossASR++无法做到这一点。此外，ASDF 还实现了更多指标，从多个角度评估自动语音识别系统的性能。ASDF 对识别出的失败测试用例进行语音分析，以确定自动语音识别系统容易转写错误的音素，为开发者改进自动语音识别系统提供有用的信息。该工具的演示视频已发布在 https://www.youtube.com/watch?v=DzVwfc3h9As。实现代码可在 https://github.com/danielyuenhx/asdf-differential-testing 获取。**

I. 引言

自动语音识别（ASR）系统在现代社会中的应用日益广泛 [1], [2], [3] ，这使得对自动语音识别系统进行充分测试的需求愈发迫切。研究人员从不同角度（如公平性 [8],、鲁棒性 [9]）提出了一系列方法，用于测试各种人工智能系统（如图像分类 [4], [5],、自动驾驶 [6], [7], 等）。近年来，自动音频测试用例生成 [10], [11], [12], [13], [14] 的兴起显著减少了自动语音识别测试过程中的人工参与。例如，CrossASR++ [11]是一种利用文本转语音（TTS）服务将文本自动生成为音频文件，并用这些音频文件测试自动语音识别系统的工具。直观来说，自动语音识别系统生成的转录文本应与用于生成音频的文本一致。否则，就发现了一个自动语音识别系统的失败测试用例。

尽管 CrossASR++ 展现了成功发现失败测试用例的能力，但它仍存在一些局限性。首先，它未能高效利用文本语料库；为了找到足够多的失败测试用例，它需要以大量文本作为输入。生成的测试用例的多样性和数量完全取决于所提供文本语料库的质量和数量。CrossASR++采用的这种穷举式筛选耗时且低效。其次，CrossASR++ 只报告发现的失败测试用例数量，而开发人员还需要更细粒度的信息来改进自动语音识别系统。

为了解决上述局限性，我们提出了 ASDF——一种用于测试自动语音识别系统的新型自动语音识别差分测试框架。我们的工具具有以下特点：它采用文本转换模块，并利用自动语音识别系统中已知的错误，从单条文本合成多个音频测试用例。在收集到一小批初始失败测试用例后，ASDF 会使用各种文本转换方法转换失败文本，从而进一步生成更多测试用例。例如，改变句子的时态，或将致错词语替换为音素相似的其他词语。实验结果表明，使用文本转换模块可使失败文本数量平均增加 22.3%。实验结果详情请见我们的 GitHub 仓库 [15]。ASDF 还会进行语音分析，以识别自动语音识别系统较难转写的音素。这些音素信息可为开发者提供参考，帮助他们进一步改进自动语音识别系统及其鲁棒性 [16]。

II. 工具设计

ASDF 扩展了 CrossASR++ [17]中使用的差分测试工作流。图 1 展示了 ASDF 用于生成测试用例的流程。该流程包含两次迭代。在第一次迭代中，语料库中的每段文本都会按如下方式处理：使用文本转语音（TTS）服务根据文本生成音频，再将音频输入自动语音识别系统，以获取各系统的转录文本。随后，我们通过比较不同自动语音识别系统的输出进行交叉参照。如果至少有一个系统能够正确转录音频，而某个特定的自动语音识别系统未能做到，我们就称该系统存在一个尚未发现的失败测试用例 。

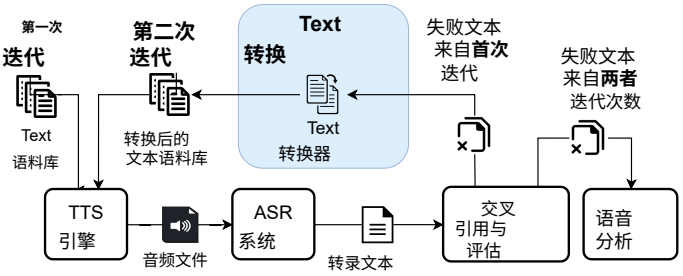


图 1。ASDF 中生成测试用例的流程概览。

如需详细了解该流程，请参阅我们之前的论文 [11] 。ASDF 可以轻松配置为测试多个自动语音识别系统。ASDF 的安装、配置和使用说明已公开发布在我们的 GitHub 仓库 [15]中。

第一次迭代中发现的失败测试用例（即原始文本及其对应的转录文本）会在第二次迭代中进行收集和分析，其中包括一个文本转换模块。在这次迭代中，文本转换器会修改失败测试用例中的文本，从而生成一个或多个转换后的测试用例，并将其添加回语料库。这一阶段对于扩展原始语料库中可用的测试用例数量至关重要。

最后，会对两次迭代中的失败测试用例进行语音分析。研究人员将分析测试用例中发现的致错词语的语音特征，以确定失败测试用例中最常出现的音素。对致错词语中高频出现的音素进行分析，有助于识别自动语音识别系统在转录时更难处理的音素。这些易出错的音素可作为自动语音识别转录领域进一步改进或研究的目标。

III. 核心功能

ASDF 需要 5 项输入：1）文本语料库；2）结果输出目录；3）待处理文本的数量；4）待测的自动语音识别系统；5）文本转换方法。输入文本语料库应为 .txt 文件格式。文件中的每行文本都会由 gTTS [18] 转换为音频文件。用户可以指定在文本转换阶段开始前要处理的文本数量。ASDF 目前支持的自动语音识别系统包括 DeepSpeech [19], 、wav2letter [20], 和 wav2vec2 [21]。通过继承自动语音识别抽象类并遵循接口规则 [17]，即可轻松添加并测试其他自动语音识别系统。

在选定的自动语音识别系统之间进行交叉参照，以确保测试用例有效，且至少能被一个自动语音识别服务识别。如果指定了文本转换器，它会转换第一次迭代中发现的失败测试用例的原始文本。此步骤会生成新的测试用例，并将其追加到原始语料库中，用于第二次迭代中测试自动语音识别系统。最后，测试结果会输出到指定路径。

目前，ASDF 支持多种转换方法。同音词转换方法首先识别导致失败测试用例出错的词语的同音词。随后，通过 WordHoard Python 库 [22] 获取一个包含该同音词的新例句。该新句子将用作新的测试用例。有关其他可用文本转换方法的详细信息和示例，例如扩增、相邻删除、复数形式转换和时态转换，请参阅我们的复现包 [15]。

IV. 用法

使用十项指标评估自动语音识别系统的性能。此外，还会绘制柱状图，展示致错词语中音素及其出现频率。系统会生成一个 CSV 文件，可用于分析各个自动语音识别系统的性能。有关这些指标的详细信息，请参阅 GitHub 仓库 [15]。

失败文本是指至少被一个自动语音识别系统错误转写的音频文件。需要注意的是，使用交叉参照来过滤文本转换产生的低质量、不真实文本。转换后的文本输入必须至少被一个自动语音识别服务正确转写，才能被视为有效（或可判定）；否则，就视为不可判定并予以丢弃。

转换后失败文本的比例，是转换后失败文本输入数量与转换后文本输入总数之比。转换后失败文本的比例越高，说明转换方法的质量越好，因为这表明其转换后的文本对自动语音识别系统而言更难正确转写。失败测试用例是指自动语音识别服务输出的特定文本与其对应的输入文本不匹配。

转换后失败用例的比例，是自动语音识别系统的转换后输出中与对应转换后输入不匹配的数量，与自动语音识别系统转换后输出总数之比。转换后失败用例的比例越高，说明生成的转换后文本对不同自动语音识别系统的覆盖率越高。

V. 挑战与未来工作

ASDF使用文本转语音库生成音频测试套件。这一做法假设生成的音频可以替代人类的现实语音，但在实际情境中未必如此。人类语音与计算机生成的音频仍然存在差异，具体包括发音、口音和声调。因此，如果使用人类语音而非计算机生成的音频，这些差异可能会使我们的差分测试产生不同结果。此外，还可以在差分测试中采用更复杂的文本转换策略，以检验这些策略在揭示错误词语方面的有效性。这类技术包括改变语法、替换词语以及修改句子结构。

参考文献

[1] V. K˙epuska 和 G. Bohouta，“新一代虚拟个人助理（微软小娜、苹果Siri、亚马逊Alexa和Google Home）”，载于*2018*年*IEEE*第八届计算与通信研讨会暨会议*(CCWC)*，2018年，第99–103页。[2] N. Yoshimura、H. Yoshida、F. Matulic 和 T. Igarashi，“通过连续语音扩展离散语音命令，实现灵活的机器人控制”，载于*2019*年*CHI*计算系统人因会议扩展摘要，系列丛书 CHI EA ’19。美国纽约州纽约：美国计算机协会，2019年，第1–6页。 [在线]。网址：https://doi.org/ 10 1145 3290607 3312791 。/。A. B. Kocaballi、J. C. Quiroz、L. Laranjo、D. Rezazadegan、R. Kocielnik、L. Clark、Q. V. Liao、S. Y. Park、R. J. Moore 和 A. Miner，“面向健康与福祉的对话代理”，载于*2020*年*CHI*计算系统人因会议扩展摘要，系列丛书 CHI EA ’20。美国纽约州纽约：美国计算机协会，2020年，第1–8页。 [在线]。网址：https://doi.org/ 10 1145 3334480 3375154 。/。Z. Yang、J. Shi、M. Asyrofi 和 D. Lo，“重新审视神经元覆盖率度量及深度神经网络的质量”，载于*2022*年*IEEE*软件分析、演化与再工程国际会议*(SANER)*。美国加利福尼亚州洛斯阿拉米托斯：IEEE计算机学会，2022年3月，第408–419页。 [在线]。网址：https://doi.ieeecomputersociety.org/ 10 1109 53432 2022 00056 。/。SANER。。X. Xie、L. Ma、F. Juefei-Xu、M. Xue、H. Chen、Y. Liu、J. Zhao、B. Li、J. Yin 和 S. See，“Deephunter：面向深度神经网络的覆盖率引导模糊测试框架”，载于第*28*届*ACM SIGSOFT*软件测试与分析国际研讨会论文集，系列丛书 ISSTA 2019。美国纽约州纽约：美国计算机协会，2019年，第146–157页。 [在线]。网址：https://doi.org/ 10 1145 3293882 3330579 。/。C. Gong、Z. Yang、Y. Bai、J. Shi、A. Sinha、B. Xu、D. Lo、X. Hou 和 G. Fan，“好奇心驱动且受害者感知的对抗策略”，载于第*38*届年度计算机安全应用会议论文集，系列丛书 ACSAC ’22。美国纽约州纽约：美国计算机协会，2022年，第186–200页。 [在线]。网址：https://doi.org/ 10 1145 3564625 3564636 。/。M. Zhang、Y. Zhang、L. Zhang、C. Liu 和 S. Khurshid，“Deeproad：面向自动驾驶系统的基于GAN的蜕变测试与输入验证框架”，载于第*33*届*ACM/IEEE*自动化软件工程国际会议论文集，系列丛书 ASE ’18。美国纽约州纽约：美国计算机协会，2018年，第132–142页。 [在线]。网址：https://doi.org/ 10 1145 3238147 3238187 。/。M. H. Asyrofi、Z. Yang、I. N. B. Yusuf、H. J. Kang、F. Thung 和 D. Lo，“Biasfinder：通过蜕变测试生成揭示情感分析系统中的偏差”，*IEEE*软件工程汇刊，第48卷，第12期，第5087–5101页，2022年。

[9] Z. Yang、J. Shi、J. He 和 D. Lo，“针对预训练代码模型的自然攻击”，载于第44届国际软件工程会议论文集，丛书：ICSE ’22。美国纽约州纽约：美国计算机协会，2022年，第1482–1493页。 [在线]。可访问：https://doi.org/ 10 1145 3510003 3510146 。/。P. Ji、Y. Feng、J. Liu、Z. Zhao 和 Z. Chen，“Asrtest：针对深度神经网络驱动语音识别系统的自动化测试”，载于第31届ACM SIGSOFT国际软件测试与分析研讨会论文集，丛书：ISSTA 2022。美国纽约州纽约：美国计算机协会，2022年，第189–201页。 [在线]。可访问：https://doi.org/ 10 1145 3533767 3534391 。/。M. H. Asyrofi、Z. Yang 和 D. Lo，“Crossasr：用于自动语音识别的模块化差分测试框架”，载于第29届ACM欧洲软件工程会议与软件工程基础研讨会联合会议，2021年，第1575–1579页。[12] M. H. Asyrofi、Z. Yang、J. Shi、C. W. Quan 和 D. Lo，“差分测试能否改进自动语音识别系统？”，载于2021年IEEE国际软件维护与演化会议（ICSME），2021年，第674–678页。[13] Z. Yang、J. Shi、M. H. Asyrofi、B. Xu、X. Zhou、D. Han 和 D. Lo，“语音测试用例优先级排序”，2023年。 [在线]。可访问：https://arxiv.org/abs/ 2302 00330 。S. S. Rajan、S. Udeshi 和 S. Chattopadhyay，“Aequevox：语音识别系统的自动化公平性测试”，载于《软件工程基础方法》，E. B. Johnsen 和 M. Wimmer 编。施普林格国际出版公司，2022年，第245–267页。[15] （2023年1月）ASDF差分测试框架（asdf）。在线。GitHub。 [在线]。可访问：https://github.com/danielyuenhx/ asdf-differential-testing [16] A. Fang、S. Filice、N. Limsopatham 和 O. Rokhlenko，“利用音素表示构建对自动语音识别错误具有鲁棒性的预测模型”，载于第43届ACM国际信息检索研究与发展会议（SIGIR ’20 07 2020, 699 708 ），第–页。[17] （2021年12月）CrossASRplus。在线。GitHub。 [在线]。可访问：https: // github.com/soarsmu/CrossASRplus [18] （2022年12月）gTTS。在线。GitHub。 [在线]。可访问：https: //github.com/pndurette/gTTS [19] （2021年11月）DeepSpeech。在线。GitHub。 [在线]。可访问：https://github.com/mozilla/DeepSpeech [20] （2022年12月）wav2letter。在线。GitHub。 [在线]。可访问：https: //github.com/flashlight/wav2letter [21] wav2vec2。在线。HuggingFace。[在线]。可访问：https: // huggingface.co/docs/transformers/model doc/wav 2 2 vec J. Bumgarner。[22] （5月）Wordhoard。2022。在线。GitHub。 [在线]。可访问：https://github.com/johnbumgarner/wordhoard