更多数据，更少变音符号：扩展阿拉伯语文本转语音

Ahmed Musleh、Yifan Zhang、Kareem
DarwishQCRI、HBKU、卡塔尔
{ahmusleh,yzhang,kadarwish}@hbku.edu.qa

arXiv:2603.01622v1 [cs.CL] 2 Mar 2026

摘要

阿拉伯语文本转语音（TTS）研究一直受到公开训练数据和准确的阿拉伯语自动变音符号标注模型可用性的限制。本文通过探索使用大规模自动标注数据训练阿拉伯语文本转语音模型，来应对这一局限。具体而言，我们构建了一套稳健的处理流程，用于收集阿拉伯语录音，并通过语音活动检测、语音识别、自动变音符号标注和噪声过滤对其进行自动处理，最终获得约 4,000 小时的阿拉伯语文本转语音训练数据。随后，我们使用不同规模的数据（即 100、1,000 和 4,000 小时），分别在添加和未添加变音符号的条件下，训练了多个稳健的语音克隆文本转语音模型。我们发现，尽管使用添加了变音符号的数据训练的模型总体表现更好，但增加训练数据量在很大程度上弥补了缺少变音符号的不足。我们计划发布一个无需添加变音符号即可运行的公开阿拉伯语文本转语音模型。

1 引言

阿拉伯语文本转语音（TTS）合成对于实现语音接口至关重要——语音接口是人机交互中最自然、最直观的方式。因此，实现清晰易懂且听起来自然的阿拉伯语文本转语音，对于教育工具和语音虚拟助手等广泛应用至关重要。尽管许多语言的文本转语音技术取得了长足进展，尤其是英语和中文，但阿拉伯语仍然落后，并面临着独特挑战：高质量大规模训练数据集有限，且需要准确地进行文本变音符号标注。附加符号是与字母相关联的语音标记，但在标准书面文本中通常会被省略，导致自动化计算系统面临大量歧义。本文提出了一种全面的方法，旨在推进最先进的阿拉伯语文本转语音，重点关注扩大数据规模并尽量减少对添加附加符号的需求。

我们的首项主要贡献是开发了一套稳健的自动化处理流程，用于整理准确的阿拉伯语音频和文本（含附加符号和不含附加符号）。该处理流程利用不断改进的技术，包括语音活动检测（VAD）、自动语音识别（ASR）、自动添加附加符号和噪声过滤，从而高效地整理海量数据。借助我们提出的处理流程，我们成功构建了一个规模庞大的阿拉伯语文本转语音训练语料库，约含 4,000 小时的数据。尽管这些技术并不完美，但我们表明，随着收集的数据不断增加，它们造成的错误会变得越来越无关紧要，从而提升文本转语音性能。

我们的第二项贡献是训练并评估了多个神经文本转语音模型，这些模型无论是否使用附加符号，都能够以极少的注册音频实现高质量的语音克隆（实验证明，注册音频仅需 3 秒也能有效实现）。我们系统地研究了训练数据量以及是否显式添加附加符号所带来的影响，并使用不同规模的数据集训练模型，分别为 100、1,000 和 4,000 小时。

第三，我们深入分析了阿拉伯语文本转语音的数据需求。尽管在带有显式音标标注的文本输入上训练的模型，其合成质量通常优于未进行音标标注的对应模型，但我们的实验清楚地表明，随着训练数据总量的增加，这一性能差距会逐渐缩小。这表明，大规模未标注音标的数据在很大程度上可以弥补缺乏显式附加符号信息的不足。最先进的音标标注器使用超过500万词的音标标注语料库进行训练，这大致相当于500至625小时音频中的语音量（假设每小时音频包含8千至1万个词）。因此，从理论上讲，在超过该数据量上训练的文本转语音模型有可能

隐式地学会有效的音标标注。

最后，基于这些发现，为推动阿拉伯语文本转语音领域的进一步研究与应用开发，我们发布了一个高质量、公开可用的阿拉伯语文本转语音模型。该模型的一项核心特性是能够直接根据未经变音符号标注的原始阿拉伯语文本输入，有效合成语音，从而简化实际部署。我们的工作为阿拉伯语语音社区提供了宝贵的资源、实验见解和性能出色的公开模型。

2 相关工作

2.1 文本转语音建模

文本转语音（TTS）合成领域正在迅速发展，重点在于开发高效的非自回归生成模型，以实现高保真度的零样本性能。Eskimez 等人 (2024) 提出了 E2-TTS，这是一种“简单得令人难以置信”的范式，采用非自回归流匹配架构，通过简化处理流程实现了最先进的效果。F5-TTS 在此框架基础上发展而来，具体通过集成 ConvNeXt 架构，解决了原始模型在训练稳定性和鲁棒性方面的挑战（Chen 等人，2024）。这一方法与利用扩散 Transformer（DiT）进行语音合成的大趋势相一致（Popov 等人，2021）；TTS Arena 上表现最好的 5 个非专有模型（mrfakename 等人，2024），例如 Kokoro[1]Fish Speech（Liao 等人，2024）和 MetaVoice [2]，均基于 DiT（Liao 等人，2024；Li 等人，2023）。这些模型体现了一个共同趋势：转向更简单的端到端系统，以消除复杂的中间表示。本文使用 F5-TTS 训练阿拉伯语文本转语音系统。

2.2 阿拉伯语文本转语音数据

早期的拼接式（Hunt 和 Black，1996）和基于隐马尔可夫模型（Zen 等人，2007）的语音合成模型需要对经过音素切分和对齐的数据集进行整理，这需要耗费大量时间和精力（Yuan 等人，2013）。公开可用且经过此类切分和对齐的阿拉伯语文本转语音数据集包括：一套由Halabi 等人（2016）开发的、包含 4 小时音频的数据集，以及另一套由 Almosallam 等人（2013）创建的、包含 7 小时音频的数据集。上述 DiT 模型无需此类数据集，因而为使用更多音频进行训练打开了大门，这些音频配有相应转录文本，并按句子级别对齐。Abdelali 等人（2022）使用一位女性配音演员录制的大约 7 小时音频和一位男性配音演员录制的大约 10 小时音频，训练了两个基于 TACOTRON 2 的阿拉伯语文本转语音模型（Wang 等人，2017）。他们的数据集并未公开。ClArTTS（Kulkarni 等人，2023）创建了一个 12 小时的单说话人语料库，涵盖古典阿拉伯语和现代标准阿拉伯语。最近，Olamide Toyin 等人（2025）推出了一套规模大得多的 83 小时混合数据集，其中包括 10 小时的语音录音和 73 小时的合成数据。迄今为止，这仍是规模最大的公开可用数据集。这些数据集中的所有文本都标注了完整的变音符号，以匹配录制的语音。Liao 等人（2024）声称使用 2 万小时的阿拉伯语音频训练了 Fish Speech 文本转语音模型，但他们既未发布该数据集，也未说明数据收集方式。

3 数据收集

我们的数据收集目标是构建一个高质量的阿拉伯语语音数据集，其中包含来自不同说话人的音频片段，每段时长为 3 至 15 秒，并附有带变音符号和不带变音符号的对应转录文本。为收集数据，我们利用了一个热门社交媒体平台上的频道和群组，用户会在其中分享长篇音频内容。收集到约 2 万小时的原始音频后，我们便停止了收集。实际上，我们本可以收集到多得多的数据。我们按以下步骤处理收集到的音频：

语音活动检测（VAD）： 我们使用 Silero 开源语音活动检测（团队，2024年）来识别包含人类语音的片段。Silero 使用数百种语言的音频进行训练，在保持轻量级的同时，提供了最先进的语音活动检测效果。

自动语音识别（ASR）： 我们使用了 Fanar ASR 模型，该模型基于 Conformer 架构，使用约 1.5 万小时的英语和阿拉伯语音频训练，其中包括现代标准阿拉伯语（MSA）和多种阿拉伯语方言（Team 等人，2025年）。据称，该模型在阿拉伯语自动语音识别方面达到了最先进的水平，并可通过公共 API 使用[3]。

添加附加符号： 我们重新实现了Elmallah 等人 (2024年) 所描述的附加符号添加器，使用

[1]https://huggingface.co/hexgrad/Kokoro-82M[2] https://github.com/metavoiceio/metavoice-src

[3]https://api.fanar.qa/docs

一个由 400 万个词组成的内部现代标准阿拉伯语语料库。该模型采用形态学感知字符级循环神经网络模型。我们的实现在维基新闻语料库上的词错误率（WER）为 5.5%（Darwish 等人，2017）。

过滤：为了过滤掉噪声较大的片段，我们测量了通过 VAD 提取的音频片段周围音频的噪声水平。如果音频片段前后 1 秒内的噪声水平高于 -30 db，我们就将该片段过滤掉。我们还过滤掉了所有短于 3 秒和长于 15 秒的片段。

处理并过滤所有原始音频后，我们得到约 4,000 小时的干净音频，并将其切分为片段，每个片段都对应带有和不带添加附加符号的转录文本。我们从完整数据集中随机抽取了 100 小时和 1,000 小时的数据，以测试数据规模对文本转语音的影响。我们创建了两个 100 小时的子集，其中一个尽可能提高说话人多样性，另一个则尽可能降低说话人多样性。总共，我们创建了 8 个训练集：4,000 小时的带有和不带有变音符号的数据、1,000 小时的带有和不带有变音符号的数据，以及 100 小时的说话人多样性最高和最低的带有和不带有变音符号的数据。

为了进行评估，我们构建了一个测试集，其中包含来自 59 位不同说话人的音频片段。每位说话人提供 11 段语音：其中一段用作语音克隆的音频提示，其余十段则用于生成样本，并通过基于参考音频的指标评估文本转语音的质量。为了获取这些说话人的音频样本，我们使用 Pyannote 说话人分离模型，对收集到的数据中随机选取的一个子集进行了说话人分离[4]（Bredin，2023；Plaquet 和 Bredin，2023）。随后，我们找出那些在单个音频文件中包含超过 11 个片段且使用现代标准阿拉伯语说话的说话人。为避免数据污染，我们将提取测试说话人的音频文件从训练数据集中排除。

4 个文本转语音模型

为了训练我们的文本转语音模型，我们选择了 F5-TTS （ Chen 等人，2024 年），这是一种集成了 ConvNeXt V2 的扩散变换器模型。ConvNeXt V2 是一种具有全局响应归一化的全卷积掩码自编码器网络。F5-TTS 凭借真实实时因子 0.15，展现出卓越的性能和更高的效率。只要训练数据足够庞大且多样，模型仅凭几秒钟的某人语音就能克隆其声音。训练完成后，向模型提供以下提示：1）待克隆声音的一段音频；2）该音频片段的转录文本；3）需要合成语音的文本。

我们训练模型时采用逐帧批量大小 38,400、初始学习率 7.5e-5 和 20,000 个预热步数。我们的 DiT 由 22 层组成，维度为 1024，并采用 14 个自注意力头。文本输入经过投影并填充至 512 维空间。我们使用 Vocos（Siuzdak，2023）梅尔声码器，根据预测的梅尔频谱特征合成音频输出。对于我们的 8 个训练集，我们分别训练了 8 个不同的模型，每个模型训练 600k 步。

5 评估指标

我们使用了两种基于参考音频的评估指标，分别是：

词错误率（WER）：我们使用上述 Fanar ASR 引擎转写合成语音，然后计算转写文本与文本转语音模型输入文本之间的 WER。我们使用 jiwer Python 库计算 WER[5]。该指标更关注生成音频的正确性。

SpeechBERTScore：该指标的设计受到 BERTScore 的启发。我们使用语音自监督学习模型（如 Wav2Vec）对参考音频和生成音频进行编码，得分则为编码后的参考音频与生成音频之间的余弦相似度（Saeki 等人，2024年）。因此，该指标更关注参考音频与生成音频之间的语音相似性。研究表明，SpeechBERTScore 与人工评定的平均意见分（MOS）具有良好的相关性（Saeki 等人，2024年）。

6 实验结果

表1列出了所有模型按词错误率排序后的结果。此外，我们还在此匿名页面中提供了一些样例：`https://arabicnlp-tts-paper-samples.great-site.net/`。结果得出了几个值得注意的结论，具体如下：

• 词错误率和SpeechBERTScore对不同文本转语音模型的排序略有不同。

[4] https://huggingface.co/pyannote/speaker-diarization-3.0

[5] https://github.com/jitsi/jiwer

设置			结果	
音标符号	Div.	Size	WER	SpeechBERT
Yes	-	4,000	1.67	0.739
Yes	-	1,000	2.55	0.729
No	-	1,000	3.12	0.726
No	-	4,000	3.18	0.732
Yes	Max	100	3.48	0.719
Yes	Min	100	3.62	0.712
No	Max	100	4.06	0.704
No	Min	100	5.20	0.706

表 1：按词错误率升序排列的结果

最显著的差异在于不带变音符号的 4,000 小时模型：词错误率给它的排名低于两个 1,000 小时模型，而 SpeechBERTScore 给它的排名则高于两者。这表明，不带变音符号的 4,000 小时模型比两个 1,000 小时模型具有更高的语音相似度（语音克隆效果更好）。

• 与不使用变音符号相比，添加附加符号能显著提升结果。这一趋势在所有数据规模下都一致。

• 不出所料，增加训练时长可以提升文本转语音模型的效果。唯一的例外是未添加变音符号的 4,000 小时和 1,000 小时模型，后者的表现略好。对此还需要进一步调查。

•增大训练集规模可以提升语音相似度，这一点从各项 SpeechBERTScore 结果中均可看出。这表明说话人语音（以及环境）的克隆效果有所改善。在提供的样例中，这一点显而易见：与 1,000 小时和 4,000 小时模型相比，100 小时模型较难捕捉参考音频中的语音和环境特征。

• 对于 100 小时模型，最大多样性比最小多样性带来了更好的结果。

7 结论

阿拉伯语文本转语音的发展长期受到两大核心挑战的阻碍：大规模公开数据集稀缺，以及未标注附加符号的文本固有的歧义性。在本研究中，我们通过开发一套稳健的自动化处理流程，整理出一个时长达4,000小时的阿拉伯语语音语料库，从而克服了这些局限。我们的处理流程有效利用了现有技术（如语音活动检测、自动语音识别和自动添加附加符号），将原始、未经标注的音频处理为高质量的训练资源。

尽管采用显式音标标注训练的模型始终优于未添加变音符号的模型，但我们证明，随着训练数据量增加，两者之间的性能差距会缩小。例如，未添加变音符号的 1,000 小时模型明显优于添加变音符号的 100 小时模型。这一趋势在词错误率和 SpeechBERTScore 指标上均有所体现，表明在数据充足的情况下，模型能够根据上下文隐式学习正确的发音，从而有效弥补缺少显式变音符号的问题。这一发现具有深远意义，说明只要专注于大规模数据收集，而非追求完美的附加符号标注，就能大幅降低构建高质量阿拉伯语文本转语音系统的门槛。我们还将发布一个高质量的预训练阿拉伯语文本转语音模型，能够直接根据未标注附加符号的文本生成自然语音。

8 局限性与未来工作

尽管我们的工作在大规模阿拉伯语文本转语音方面迈出了重要一步，但我们承认存在若干局限，这些局限为未来研究提供了方向。

自动化处理流程的质量上限：数据集质量受各个组件性能的限制，例如自动语音识别系统和自动注音器。这些错误会传播到训练数据中。尽管我们的结果表明，这些错误的影响在超大规模下会减弱，但它们仍然是一种可以降低的噪声来源。

对客观指标的依赖：词错误率和 SpeechBERTScore 可能无法充分反映合成语音的感知质量，例如自然度、韵律、情感语调以及变音符号标注错误。此外，使用同一自动语音识别模型系列来生成训练数据并进行词错误率评估，可能会引入有利偏差。人工评估对于验证真实世界性能和用户偏好至关重要。

领域与方言范围：我们的研究并未分析最终数据集的方言分布，也未评估文本转语音模型呈现特定方言的能力。我们仅在现代标准阿拉伯语上进行了测试。未来研究可以聚焦于整理特定体裁语料库和特定方言语料库，以训练更加专业化且可控的阿拉伯语文本转语音系统。

参考文献

Ahmed Abdelali、Nadir Durrani、Cenk Demiroglu、Fahim Dalvi、Hamdy Mubarak 和 Kareem Darwish。2022。Natiq：面向阿拉伯语的端到端文本转语音系统。*arXiv*预印本 *arXiv:2206.07373*。

Ibrahim Almosallam、Atheer AlKhalifa、Mansour Al-Ghamdi、Mohamed I Alkanhal 和 Ashraf Alkhairy。2013年。Sassc：标准阿拉伯语单说话人语料库。载于*SSW*，第 249–253 页。

Hervé Bredin。2023年。pyannote.audio 2.1 说话人分离处理流程：原理、基准测试与实践指南。载于 *INTERSPEECH2023* 论文集。

Yushen Chen、Zhikang Niu、Ziyang Ma、Keqi Deng、Chunhui Wang、Jian Zhao、Kai Yu 和 Xie Chen。2024年。F5-tts：一位故事讲述者，借助流匹配生成流畅且忠实的语音。*arXiv* 预印本 *arXiv:2410.06885*。

Kareem Darwish、Hamdy Mubarak 和 Ahmed Abde- lali。2017年。阿拉伯语变音符号标注：统计方法、规则与技巧。载于第三届阿拉伯语自然语言处理研讨会论文集，第 9–17 页。

Muhammad Morsy Elmallah, Mahmoud Reda, Kareem Darwish, Abdelrahman El-Sheikh, Ashraf Hatim El- neima, Murtadha Aljubran, Nouf Alsaeed, Reem Mo- hammed 和 Mohamed Al-Badrashiny。2024年。使用形态学信息增强的字符级模型为阿拉伯语添加附加符号。载于*2024*年计算语言学、语言资源与评估联合国际会议（*LREC-COLING 2024*）论文集，第 1446–1454 页。Sefik Emre Eskimez、Xiaofei Wang、Manthan Thakker、Canrun Li、Chung-Hsien Tsai、Zhen Xiao、Hemin Yang、Zirun Zhu、Min Tang、Xu Tan 等 1 人。2024年。E2 TTS：极其简单的全非自回归零样本文本转语音。载于*2024 IEEE*口语技术研讨会（*SLT*），第 682–689 页。IEEE。Nawar Halabi 等 1 人。2016年。阿拉伯语语音语料库。牛津文本档案核心馆藏。Andrew J Hunt 和 Alan W Black。1996年。在大型语音数据库中使用单元选择的拼接式语音合成系统。载于*1996 IEEE*国际声学、语音与信号处理会议论文集，第 1 卷，第 373–376 页。IEEE。Ajinkya Kulkarni、Atharva Kulkarni、Sara Abedal- monem Mohammad Shatnawi 和 Hanan Aldarmaki。2023年。ClArTTS：一个开源的古典阿拉伯语文本转语音语料库。*arXiv*预印本 *arXiv:2303.00069*。Yinghao Aaron Li、Cong Han、Vinay Raghavan、Gavin Mischler 和 Nima Mesgarani。2023年。StyleTTS 2：通过风格扩散、结合大规模语音语言模型的对抗训练，实现媲美人类水平的文本转语音。神经信息处理系统进展，36:19594–19621。

Shijia Liao、Yuxuan Wang、Tianyu Li、Yifan Cheng、Ruoyi Zhang、Rongzhi Zhou 和 Yijin Xing。2024年。Fish-speech：利用大语言模型实现先进的多语言文本转语音合成。预印本，arXiv:2411.01156。

mrfakename、VaibhavSrivastav、ClémentineFourrier、Lucain Pouget、Yoach Lacombe、main 和 San- chit Gandhi。2024年。文本转语音竞技场。https://huggingface.co/spaces/TTS-AGI/TTS-Arena。

Hawau Olamide Toyin、Rufael Marew、Humaid Al-blooshi、Samar M Magdy 和 Hanan Aldarmaki。2025年。Arvoice：用于阿拉伯语语音合成的多说话人数据集。*arXiv* 电子预印本，页码 arXiv–2505。

Alexis Plaquet 和 Hervé Bredin。2023年。用于神经说话人分离的幂集多类交叉熵损失。载于*INTERSPEECH 2023* 论文集。

Vadim Popov、Ivan Vovk、Vladimir Gogoryan、Tasnima Sadekova 和 Mikhail Kudinov。2021年。Grad-tts：一种用于文本转语音的扩散概率模型。载于国际机器学习会议，页码 8599–8608。PMLR。

Takaaki Saeki、Soumi Maiti、Shinnosuke Takamichi、Shinji Watanabe 和 Hiroshi Saruwatari。2024年。SpeechBERTScore：利用自然语言处理评估指标进行具备参考音频感知能力的语音生成自动评估。载于 *Interspeech 2024*，第 4943–4947 页。

Hubert Siuzdak。2023年。Vocos：缩小时域与基于傅里叶变换的神经声码器之间的差距，实现高质量音频合成。*arXiv* 预印本 *arXiv:2306.00814*。Fanar Team、Ummar Abbas、Mohammad Shahmeer Ahmad、Firoj Alam、Enes Altinisik、Ehsannedin Asgari、Yazan Boshmaf、Sabri Boughorbel、Sanjay Chawla、Shammur Chowdhury 及另外 1 位作者。2025年。Fanar：以阿拉伯语为中心的多模态生成式人工智能平台。*arXiv* 预印本 *arXiv:2501.13944*。Silero Team。2024年。Silero 语音活动检测：预训练的企业级语音活动检测器、数字检测器和语言分类器。https://github.com/snakers4/silero-vad。Yuxuan Wang、RJ Skerry-Ryan、Daisy Stanton、Yonghui Wu、Ron J Weiss、Navdeep Jaitly、Zongheng Yang、Ying Xiao、Zhifeng Chen、Samy Bengio 及另外 1 位作者。2017年。Tacotron：迈向端到端语音合成。*arXiv* 预印本 *arXiv:1703.10135*。Jiahong Yuan、Neville Ryant、Mark Liberman、Andreas Stolcke、Vikramjit Mitra 和 Wen Wang。2013年。利用边界模型进行自动音素切分。载于 *Interspeech*，第 2306–2310 页。Heiga Zen、Takashi Nose、Junichi Yamagishi、Shinji Sako、Takashi Masuko、Alan W Black 和 Keiichi Tokuda。2007年。基于隐马尔可夫模型的语音合成系统（HTS）2.0 版。*SSW*，6：294–299。