浙江大学

本科课程论文

课程名称：计算机科学思想史

姓　　名：胡若凡

学　　院：计算机学院

专　　业：计算机科学与技术

学　　号：3200102312

指导教师：巫英才等

2023 年 6 月 6 日

语音合成技术文献综述

胡若凡，3200102312

摘　要:　语音合成技术（Text to Speech），旨在让计算机将一段文本信息转换为声音信息，目前在人机交互领域有着重要作用。从技术层面而言，目前广泛采纳的深度学习语音合成技术，在生成的质量和速度上都超过了传统语音合成技术，并不断快速发展。从语音合成的深度学习原理出发进行综述，比对传统语音合成方法，分析从自回归模型到非自回归模型的模型发展历史，总结模型的特征与优缺点。针对目前还欠缺的多模态统一模型的问题，总结当前的尝试的改进方案，和对未来的展望。

关键词:　语音合成; 深度学习；生成模型；

一. 研究意义

语音合成（Text-to-Speech）是研究将文本转换成语音的技术，在日常生活中有着广泛的应用，比如智能助手、人声阅读、虚拟歌声等。这些应用能够提供更加智能化的服务，增强用户的交互体验。具体而言，TTS 技术可以用清晰无干扰的语音来进行指导，提高信息传递的质量与降低说明成本；TTS 技术可以自由定义语音时长，匹配视频产品的制作，提升工作效率； TTS 技术可以对人声进行模拟，制造不同音色的音频，丰富产品的种类与观感。该技术在目前有着广泛的使用范围与较好的发展前景。

二. 研究概览与发展脉络

2.1 研究概览

传统的TTS技术，将合成过程分为前端和后端两个模块，前端模块主要负责对输入文本进行预处理分析，从文本归纳出语音特征，转为声学可用数据；后端模块则根据这些数据，生成相应的梅尔频谱图，再由声码器生成音频。在TTS技术中，前端模块使用了自然语言处理、语音信号处理等技术，后端模块则包括训练器，合成模型、声码器组件[1]。

展开而言，前端模块需要文本分词、停顿分析、语调分析，其作用是确定词句的轻重读与语调上扬下降的幅度；此外，还需要对文字进行韵律分析，确定每个音素在不同语境下的发音方式，以解决多音问题。这些处理的结果将被转换为音素序列，为后续的

语音生成提供基础。

后端模块则是接收前端模块输送的语料，根据转换的要求与对应的声学特征，最终形成一张囊括所有声学所需信息的梅尔频谱图。再由声码器（Vocoder）接受梅尔频谱图，将其转换为最终的语音信号。声码器可以将模型中的声学特征转换为语音波形，生成高质量的语音信号。其实现方式有多种，如基于谱重构的方法、基于神经网络的方法等，目前常用的用HifiGan声码器等。

如[2]所言，TTS技术中的关键便在于形成一张将韵律、表现力等特征都囊括其中的梅尔频谱图，其是否平滑，粒度是否显示足够信息，由TTS模型决定，而最终声音的清晰度则由声码器决定。如何开发一个良好的TTS模型，便成为了目前的关键。

2.2 发展脉络

TTS模型的主要研究发展中，综述文献[3]概括了使用的方法经过了拼接法、参数法，直到目的深度学习法。

前两种方法为传统研究方法。拼接法主要通过前期录制大量音频，以尽可能完整地覆盖所有音节音素，合成时直接进行音素单位的拼接。参数合成法则主要对已有的录音样本，通过人工或自动设定声学参数（如基频、谐波、共振峰等），构建从声学参数到完整语音之间的映射合成器。

而近年的改进，则集中在利用各类深度学习模型，基于其强大的特征提取能力和序列数据处理能力，突破了传统观念下必须先形成中间语料，再生成梅尔频谱图转换音频的阶段技术。文献[4]里也详细论证了，开发端到端的技术里，让模型从文本直接生成梅尔频谱图，既加快了模型速度，也保证了音频质量。

端到端TTS模型发展非常迅速，研究的主体也经历了从自回归结构模型发展为非自回归结构的模型的阶段。文献[5]中分析，早期的自回归TTS模型发展于深度学习兴起之时，认为语音的生成必须有前后之别，生成每个梅尔频谱帧时，都必须考虑前面的帧对该帧的影响，这往往导致训练时间长、鲁棒性低的问题。

而随着硬件技术的发展以及对并行速度的追求，则出现了新的非自回归的TTS模型，它们注重并行生成梅尔频谱帧，极大加快了生成时间。例如基于Transformer的FastSpeech[6]是最先出现的非自回归模型，基于并行注意力头实现了时间上的重大突破。

三. 目前方法与问题

3.1 传统语音合成

综述文献[3]对传统的方法，也有详细的介绍。拼接法（concatenative synthesis）通过连接小的、事先录好的语音单元（如音素、双音素、三音素等）并经过韵律修饰（prosodic modification）来拼接整合成完整的语音。这种技术通过波形处理使得言语的超音段特征发生改变，而音段特征（谱包络）保持不变。拼接法最大限度地保留了原始发音人的音质，但直接拼接的方法导致语音听起来生硬，韵律修饰处理出的转折生硬，导致边界处明显不连续，容易产生错误。这种方法合成效果不稳定、音库容量大，构建周期长，可扩展性太差。总体而言，拼接法已无法满足当前的主要需求。

参数法（parametric speech synthesis）方法中，常见的是基于隐马尔可夫模型的参数语音合成方法，采用统计建模、特征预测、参数生成和音频生成的流程。该方法首先需要从训练数据的语音文件中提取各帧对应的声学参数，然后对待转换文本进行分析，得到各音素相应的上下文属性。接着，根据这些属性，通过对之前训练出参数的决策树进行决策，得到待合成语句对应的隐马尔可夫模型序列。最后由声码器对模型进行转换。然而，这种基于传统概率算法的声码器的问题是，它反映复杂映射的能力有限，只在对一些简单的语音任务上较为适用，对于复杂情况下重建出来的声音，质量较为差。

近年来，基于深度学习的模型在语音质量、模型轻量、转换速度上都更胜一筹，一般分为自回归模型与非自回归两类模型。

3.2 基于自回归模型的声码器

3.2.1 Wavenet

自回归声码器兴起于机器学习的较早阶段，主要有两个具有代表性的声码器模型：Wavenet与WaveRNN。

文献[7]指出，WaveNet是由Google DeepMind于2016年推出的自回归模型，其主模型基于PixelCNN架构实现。WaveNet首次实现了端到端的训练，并通过学习音频波形的概率分布来生成新的语音波形。具体地说，模型采用了一种逐步增量的采样策略，按顺序预测每个时刻的输出，直到生成整个波形。

展开而言，在训练时，模型首先将语音序列联合概率 $x = \{x1, x2, \cdots, xT\}$ 分解为各时刻条件概率的乘积，如公式（1） 所示，其中x 是语音波形值序列，x_t 是一个时刻波形值：

$$p(x)=\prod_{t=1}^{T} p(x_t|x_1,x_2,\cdots,x_{t-1}) \quad (1)$$

下一步，需要根据公式 2 对声码器进行建模：

$$p(x|h)=\prod_{t=1}^{T} p(x_t|x_1,x_2,\cdots,x_{t-1},h) \quad (2)$$

在建模时，Wavenet有两种方法：全局方法和局部方法。全局方法是指模型接收单额外输入条件h ，该条件在所有的时间点上影响模型的输出；局部条件建模方法是指模型有第二种时间序列 ht ，通过对原始数据的低采样率获得。

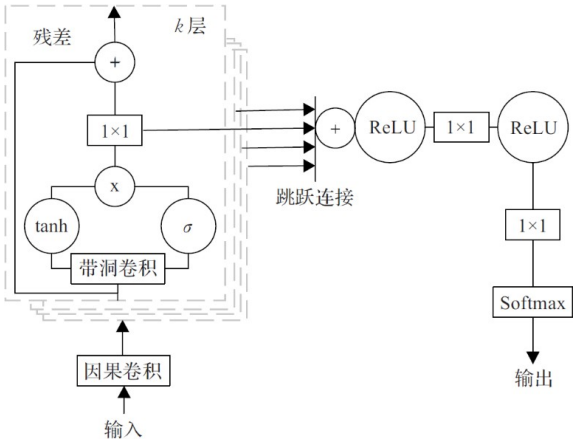

图 1：Wavenet 模型

在WaveNet声码器中，由于卷积神经网络难以捕捉语音信号的动态信息，可能会导致合成语音中出现噪音和失真等问题。文献[8]通过引入基于感知权重的噪声整形技术来解决这一问题。具体地，该方法利用预测残差的梅尔频谱特征进行感知权重建模，并将感知权重应用于形状参数的预测中，从而产生更加自然流畅的语音样本。

综合而言，WaveNet作为最早的自回归模型，训练时间和计算资源消耗较大，需要大量的语音样本和高性能计算机，生成语音的速度较慢，每秒只能生成24,000个样本；然而它的优势在于生成的合成语音比传统的文本转语音系统更加逼真，并且训练的更为简单。

3.2.2 WaveRNN

与WaveNet相比，文献[9]提出的WaveRNN的出现则是为了解决CNN网络架构复杂，生成效率低下的问题。

在技术上，如图2，WaveRNN使用了简化模型、矩阵稀疏化、并行序列等技术，显著提升了序列生成速度。其采用了双路 softmax 层，在保持语音合成质量的同时，相比

WaveNet 更加紧凑，可以在 GPU 上 4 倍快速地生成 24 kHz 16 位音频。其次，WaveRNN 通过权重剪枝技术，大幅度降低了权重矩阵的占比。

实验结果显示，对于固定数量的参数，大型稀疏网络比小型稠密网络表现更好，这种关系在稀疏度超过 96% 时仍然成立。而正由于稀疏 WaveRNN 的权重数量较少，甚至能够在移动 CPU 上实时采样高保真音频，甚至在手机上都可以部署使用。

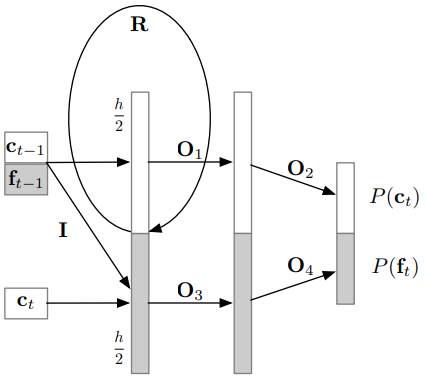

图 2：WaveRNN 模型

3.3 非自回归模型

3.3.1 FastSpeech

尽管例如WaveRNN已在速度上有了一定的提升，但是其仍然不满足于GPU发展后高并行化快速生成的要求，因此，文献[6]提出的FastSpeech则是非自回归模型中首先出现的佼佼者。

如下的图3可以看出，FastSpeech主要采用基于Transformer的并行多注意力机制[10]，编码器将输入的文本转换为一组隐藏表示，解码器将这些表示转换为梅尔频谱图。FastSpeech的长度调节器可以确保文本和梅尔频谱图的长度匹配。具体而言，通过使用均方误差（MSE）损失从自回归模型中提取注意力，模型可以预测音素持续时间，并使用六个前馈变换块，用于将音素序列转换为mel-spectrogram序列。经测试，FastSpeech将mel-spectrogram的生成速度提高了270倍，端到端语音合成速度提高了38倍。

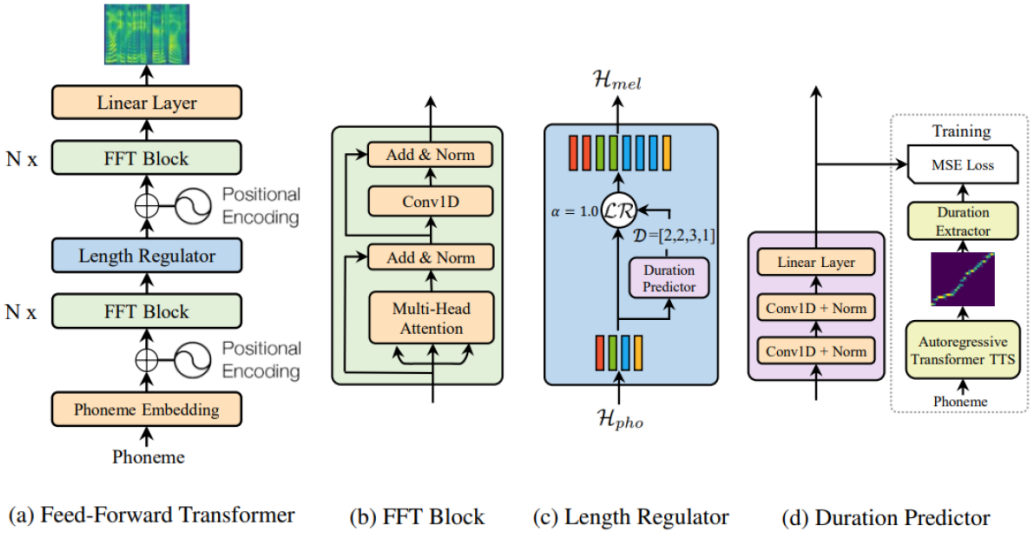

图 3：FastSpeech 模型

3.3.2 FastSpeech2

文献[11]提出的FastSpeech 2，是FastSpeech的更新版本。FastSpeech的具体技术细节中，曾依赖于一个自回归的教师模型来进行持续时间预测和知识蒸馏，这导致了一些基础信息的损失。如图4，FstSpeech2直接从基准目标中训练以避免信息损失，而不是使用教师模型中的简化数据[8]。该模型从目标语音中确定持续时间、音调和能量的值，并使用连续小波变换方法将音高轮廓转换为音高谱图。

在训练过程中，FastSpeech 2使用了一个称为DiffWave的非自回归声码器，该声码器将梅尔频谱图转换为实际的波形信号。与FastSpeech相比，FastSpeech 2具有更快和更准确的语音合成速度，并产生更高质量的语音输出。

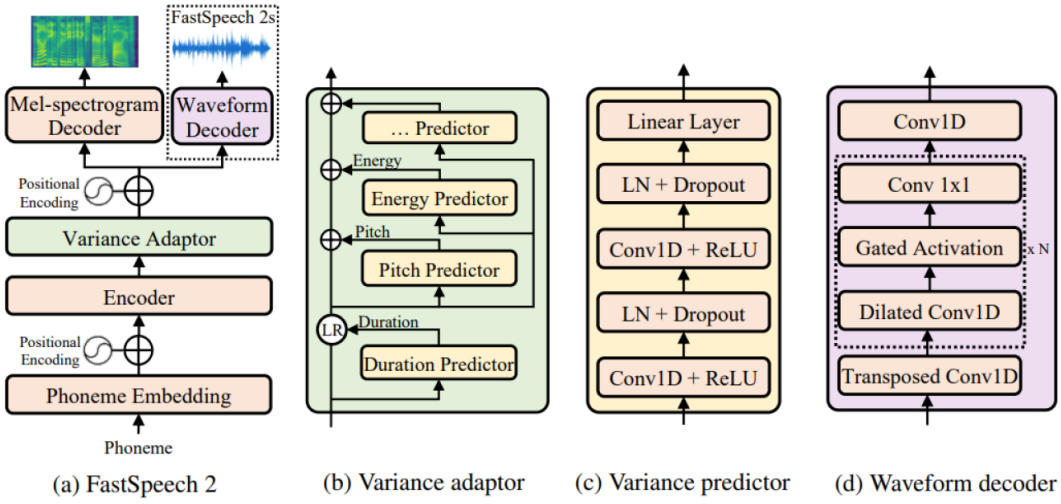

图 4：FastSpeech2 模型

3.3.3 PortaSpeech

近年来对语音合成模型的大小也提出了更高的要求，要求模型要具备高质量的合成效果且尽可能地轻量化。在这方面，文献[12]提出的PortaSpeech是一种具有高质量、轻量级特点的生成式语音合成模型。如图[5]，该模型利用了轻量级的变分自编码器（VAE）和正则化流等方法来捕捉语音的隐含特征和变化，使用基于Transformer的生成器来从隐含特征中生成mel-spectrogram等特征。此外，PortaSpeech还支持通过控制VAE和normalizing-flow的采样参数来调节语音的风格、情感和速度等属性。

与其他TTS模型相比，主观和客观实验结果表明PortaSpeech在语音音质和韵律建模方面都具备优异的性能。此外，当将模型参数量压缩到6.7M（相当于在FastSpeech 2的基础上进行4倍的模型压缩和3倍的运行内存压缩）时，PortaSpeech也能保持稳定的合成效果。

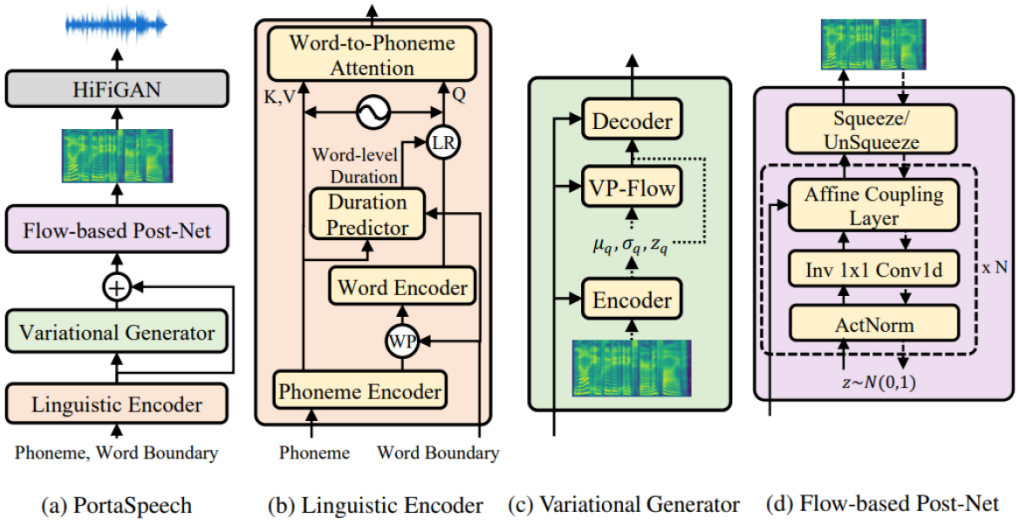

图 5：PortaSpeech 模型

3.4 训练与转换问题

如文献[3]所言，在使用深度学习训练的声码器中，依旧存在两个共同的问题。首先，在许多域适应语音合成方法的训练中，常常基于一个理想性假设，即在不同数据源之间对齐特征。这些方法希望生成模型能够在实际使用中对域移动保持不变。然而，在实际情况下，由于不同的语音文本可能需要表达不同的感情，一个语者的发音音调也可能随着文本而发生变化。这些因素都会对域的移动产生影响，难以保持不变，训练时往往需要海量的音频文本去进行训练，才能克服这一问题。

其次，如文献[13]指出，对于语音转换技术的使用范围。虽然声码器在对语音模型进行转换的过程中已经取得了一定的成功，并且近年来也有将TTS技术应用于歌声合成、视频配音等方面，但是其音色往往还较为单一，无法自由调配，而是受训练数据影响。

四. 发展方向

4.1 模型数据集优化

针对第三部分提出的问题，文献[14]指出，可以采用多域自适应的方法。具体来说，设计者可以使用多个域的数据训练生成模型，以使其对不同域的语音和语言特征具有更好的通用性。同时，可以通过引入辅助任务、正则化等技术来使得模型对于感情和语气等方面的变化具有一定的鲁棒性。以此来使得模型可以对域的迁移匹配做到更好。

4.2 使用方向拓展

多模态语音合成是一种集成多种模态信息的语音合成技术。除了文本，现有模型也在研究结合图片、视频和语音等多种模态信息进行语音合成，使得合成语音更加丰富和多样化。

在多模态语音合成中，图像信息可以包括人物头像、表情动作、身体姿态等，语音信息可以包括声音的节奏、音调、韵律等，这些都可以通过多模态语音合成技术被整合起来。在视频制作方面，多模态语音合成可以通过结合图片和语音信息，让虚拟角色的声音与其嘴部动作、面部表情等保持同步，从而增强用户体验。在虚拟现实方面，多模态语音合成可以结合用户的视觉和听觉感受，为用户提供更加沉浸式的体验。在语音助手方面，多模态语音合成可以通过结合用户的语音指令和显示屏信息，让语音合成的内容更具有表现力和交互性。

4.3 统一的集成模型

如[15]，随着chatgpt的兴起，集成的大模型也成为了研究的主流方向之一。TTS技术目前在语音合成、歌声合成等多个子任务方面，都有了较为优秀的模型。但是尽管以音频作为输出，却依旧缺少一个统一的模型，能够一次性的完成这些任务，这无疑为TTS的实用性打了折扣。因此，统一的多任务音频合成模型也势必会成为未来研究重点方向。

五. 个人见解

Chatgpt已显示了AIGC对于生活的巨大影响，在人机交互愈发频繁的今日，从文字交流到语音交流，显然也是未来的一个主流方向，而这一切也需要TTS技术的支撑。因此，未来的TTS模型将会进一步进入人类的生活之中，有着十分优秀的发展前景。

在技术上，我认为未来的发展趋势是仍让会在非自回归模型领域发展，并且发展的方向将会是数据集训练与统一模型的开发之上。正如第四部分中所言，多域自适应的方法可以有效提高模型的鲁棒性，解决模型当今的问题，而TTS的训练语料目前也较为短缺和固定，因此，为了更加稳定的模型，日后将会涌现更多优秀的音频数据集。此外，为了实操中的更加方便，一个如同chatgpt一样可以同时完成多种TTS任务的ttsgpt，也显然是一个十分需要被研究的领域。

在日后，相信随着TTS技术的进一步发展，必然会为我们的生活带来更多的便利与改善。

六. 参考文献

[1]Sercan O Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. arXiv preprint arXiv:1702.07825, 2017.

[2]Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In Advances in Neural Information Processing Systems, pages 1171–1179, 2015.

[3]Xu Tan.A Survey on Neural Speech Synthesis.arXiv preprint arXiv:2106.15561 2021.

[4]Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In Advances in Neural Information Processing Systems, pages 1171–1179, 2015.

[5]Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Nonautoregressive neural machine translation. arXiv preprint arXiv:1711.02281, 2017.

[6]Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In Advances in Neural Information Processing Systems, pages 3165–3174, 2019.

[7]Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. arXiv preprint arXiv:1711.10433,2017.

[8]Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4779–4783. IEEE, 2018.

[9]Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, Koray Kavukcuoglu.Efficient Neural Audio Synthesis.arXiv preprint arXiv:1802.08435 2018.

[10]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998–6008, 2017.

[11]Yi Ren, Chenxu Hu, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text-to-speech. arXiv preprint arXiv:2006.04558, 2020

[12]Yi Ren, Jinglin Liu, Zhou Zhao. PortaSpeech: Portable and High-Quality Generative Text-to-Speech. arXiv preprint arXiv:2109.15166, 2022

[13]Nishimura, M.; Hashimoto, K.; Oura, K.; Nankaku, Y.; and Tokuda, K. 2016. Singing Voice Synthesis Based on Deep Neural Networks. In Interspeech, 2478–2482.

[14]Blaauw, M.; and Bonada, J. 2020. Sequence-to-sequence singing synthesis using the feed-forward transformer. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7229-7233. IEEE.

[15]Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren , Zhou Zhao, Shinji Watanabe. AudioGPT: Understanding and Generating Speech,Music, Sound, and Talking Head. arXiv preprint arXiv:2304.12995, 2023