FALL-E: GAUDIO FOLEY SYNTHESIS SYSTEM

Minsung Kang, Sangshin Oh*, Hyeongi Moon, Kyungyun Lee, Ben Sangbae Chon*

Gaudio Lab, Inc., Seoul, South Korea. bc@gaudiolab.com

ABSTRACT

This paper introduces FALL-E, Gaudio's Foley Synthesis System, which is submitted to the DCASE 2023 Task 7 - Foley Synthesis Challenge (Track A). The system employs a cascaded approach comprising low-resolution spectrogram generation, spectrogram super-resolution, and a vocoder. We trained every sound-related model from scratch using our extensive datasets, and we utilized a pre-trained language model. We conditioned the model with dataset-specific texts, enabling it to learn sound quality and recording environment based on the text input. Moreover, we leveraged external language models to improve text descriptions of our datasets and performed prompt engineering for quality, coherence, and diversity. We report the objective measure with respect to the official evaluation set, although our focus is on developing generally working sound generation models beyond the challenge.

***Index Terms*—** Generative models, DCASE, sound synthesis

1. INTRODUCTION

Generative AI has seen significant progress in recent years, particularly in the domains of images and text. However, the progress in sound generation has been comparatively slower. To address this gap, we proposed a challenge for foley synthesis, which was later incorporated into an official proposal to DCASE and accepted [1]. In this report, we introduce Gaudio's Foley Synthesis System, which is our submission to the challenge.

Generative models have advanced rapidly in recent years, particularly in the domains of images and texts. In the field of sound generation, however, progress has been comparatively slower. To address this gap, numerous impressive works have been introduced including text-to-sound models such as AudioGen [2] and AudioLDM [3]. In addition, several works can be used as modules of the whole system such as HiFi-GAN [4], SoundStream, EnCodec [5, 6], latent diffusion [7], and spectrogram super-resolution [8].

Furthermore, in text-input and text-conditioned generation, models such as T5 [9], GPT [10, 11], text prompt engineering [12, 13], and diffusion with conditioned generative models [14, 15, 2, 3] have been introduced. As the behavior of large deep learning models is somewhat difficult to analyze, these works enable us as users to steer the model using carefully selected text inputs.

In this context, we present a novel approach to foley synthesis that utilizes a cascaded system composed of low-resolution spectrogram generation, a super-resolution module, and a vocoder. Our system represents our submission to the DCASE 2023 Task 7 - Foley Synthesis Challenge (Track A) [16]. While we report objective measures with respect to the official evaluation set, our ultimate goal is to develop sound generation models that extend beyond the challenge's scope.

In Section 2, we introduce our model architecture, FALL-E, detailing the function of each module and how they work in tandem. In Section 3, we provide an in-depth analysis of our evaluation results, showcasing the effectiveness of our approach in various settings. Lastly, in Section 4, we summarize our contributions and highlight future directions for our work. Overall, we believe our system represents a significant step forward in foley synthesis and we are excited to share our findings with the research community.

2. FALL-E

2.1. Architecture

Cascaded systems with intermediate features have been widely used for sound synthesis applications such as symbolic music generation [17, 18], text-to-speech synthesis [19, 20, 21], and audio generation [3]. The ease of training and interpretability make cascaded systems a preferred approach in the field. Several cascaded systems for speech and audio generation have used mel-spectrogram as their intermediate feature. We adopt this approach to generate foley sound. Our proposed system, called FALL-E, consists of three separately-trained models: a Glide-based [15] feature generation model, a diffusion-based upsampling model, and a mel-spectrogram inversion model based on the HiFi-GAN neural vocoder [22].

Feature generation model is based on Glide, a diffusion generative model for text-to-image generation [15]. To generate sound signal from the text prompt, we borrow the main branch of the Glide except the image-related text encoder.

*Equal contribution

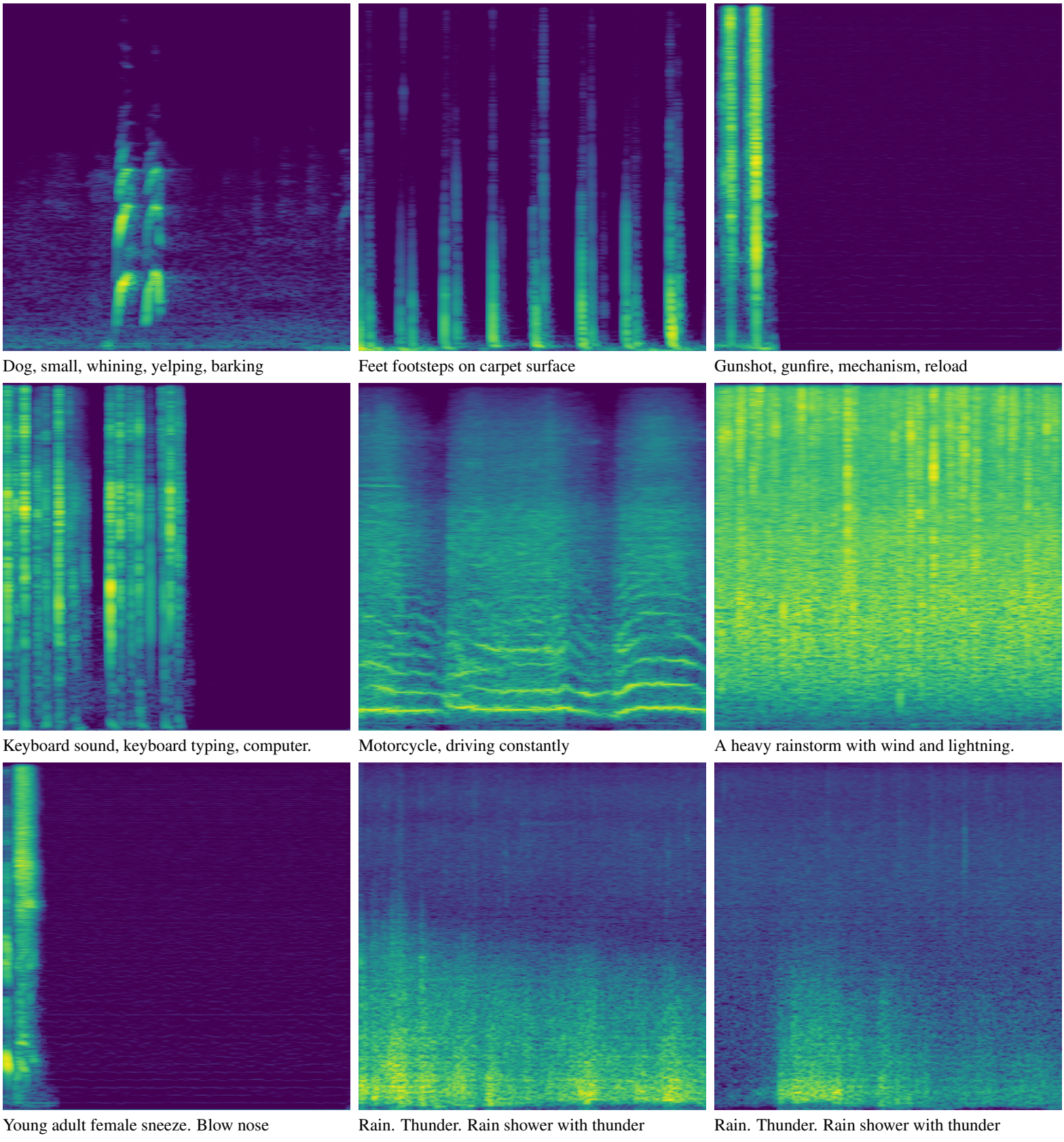


Figure 1: Selected 1024-bin mel spectrograms, all with the *clean* prefix, from one of the final versions of our model.

The architecture of the model includes 5 blocks for each side of the U-Net shaped network and convolution channels increases 2 times as it goes to the deeper block starting from 192.

Text encoder of the system is a pretrained Flan-T5, an instruction finetuned-variants of a T5 model which shows better performance for various applications [23]. We feed a sequence of text embedding from this model to the Glide-based feature generation model with omitting the sentence-level embeddings. To generate category-based audio signals, predefined text prompts are used.

Upsampling model is another diffusion-based generative model that synthesizes mel-spectrograms from the generated features (or low-resolution spectrograms). The overall architecture of this model is also a U-Net, that is similar to that of the feature generation model, but the hyperparameters differ to make it a smaller size. It uses 4 blocks in each side, and starting channel size is 128.

Mel inversion model converts the generated mel-spectrogram into waveforms. This model has a similar structure to HiFi-GAN vocoder, but we add skip connections to every convolution block to improve phase reconstruction for general audio signals.

The whole system (including the mel inversion model and the text encoder) has 642M parameters. The system can be effectively served with a single GPU.

2.2. Training and Inference Details

Datasets for the training include various sources across public and private audio datasets, including AudioSet [24], CLOTHO [25], Sonniss,[1] WeSoundEffects,[2] ODEON,[3] and FreeToUseSounds.[4] To prevent the data imbalances or the potential risks of model misbehavior, audio samples with speech or musical contents are filtered out based on their metadata. After the filtering, we use 3815 hours of audio signals for training.

Text conditioning can be optimized or engineered to improve the model behavior for both training and inference. One of our focus was to control the recording condition / environment of the generated signals so that the model can learn from crowd-sourced noisy (low SNR level) datasets as well, while being able to produce high-quality audio. Among the datasets we used, AudioSet was the only "*noisy*" dataset. We append a special token that indicates *noisy dataset* to the text input during training. For the other datasets, we append *clean dataset* token. The impact of this additional token will be discussed later. We also clean the text label (i.e., text normalization) by dropping some stop words and numbers.

[1]https://sonniss.com/gameaudiogdc

[2]https://wesoundeffects.com/we-sound-effects-bundle-2020

[3]https://www.paramountmotion.com/odeon-sound-effects

[4]https://www.freetousesounds.com/all-in-one-bundle/

Sound class	**FAD ↓ (ours)**	**FAD ↓ (baseline)**
Dog bark	8.685	13.411
Footstep	5.644	8.109
Gun shot	2.633	7.951
Keyboard	3.835	5.230
Moving motor vehicle	6.540	16.108
Rain	5.464	13.337
Sneeze & cough	2.390	3.770
Average	5.027	9.702

Table 1: The FAD results were obtained by evaluating the FAD scores for 100 generated audio samples in each sound class.

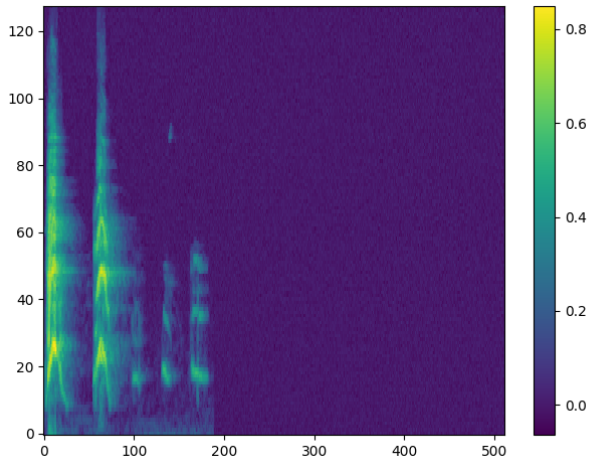
3. EVALUATION AND ANALYSIS

Table 1 presents the FAD scores of our proposed approach compared to the baseline approach across all sound classes using the official evaluation repositories.[5] Our approach outperforms the baseline approach in all classes, with notable improvements observed in the rain and moving motor vehicle classes. These classes are characterized by steady sounds, which lack onsets and offsets. Furthermore, the subjective quality is significantly improved by our model in all classes. It should be acknowledged that FAD scores may not be indicative of other important aspects of audio quality such as clarity, high-SNR, and high-frequency components. Also, as FAD measures similarity between a reference set and a test set, improvement beyond reference is mis-measured as a degradation.

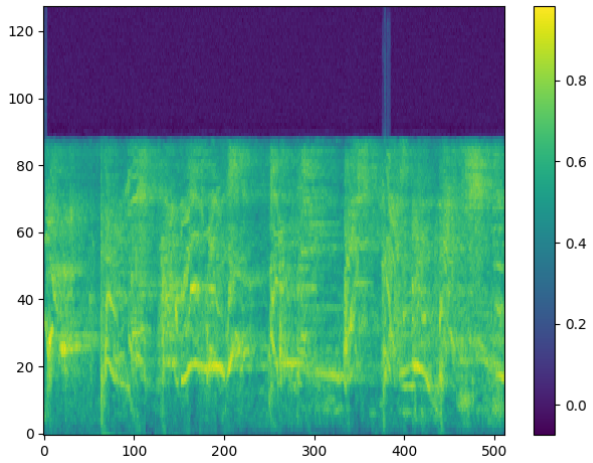
Our model was developed to generate high-quality audio suitable for real-world scenarios using environment and audio quality prefixes. Despite most of the audio samples in our training dataset exhibiting poor audio quality due to background noise, babble noise, wind noise, device noise, and codec distortion, we confirmed our model produces high-quality audio. As discussed in Section 2.2, we controlled the audio sample quality by adding a special token as a prefix to the original text. Given that audio quality cannot be evaluated objectively, we conducted a listening test for the same text with both *clean* and *noisy* prefixes. Depending on the prefix used, we observed impressive improvements in sound quality across all sound classes. As illustrated in Figure 2, we can clearly observe that the use of the *clean* prefix had a discernible impact on the audio quality, as indicated by the mel spectrogram images. This type of model steering by prompting has been popular in other domains, and to our best knowledge, our work is the first work that successfully shows it in audio generation.

Despite the fact that DCASE 2023 Task 7 does not include generating audio based on natural language text, our

[5]https://github.com/DCASE2023-Task7-Foley-Sound-Synthesis



(a) Dog bark with the *clean* prefix



(b) Dog bark with the *noisy* prefix

Figure 2: Mel spectrograms of the generated audio samples using different recording environment prefixes.

model is designed to do so; similar to AudioGen [2] and AudioLDM [3]. However, in our experiments with various text prompts, we also have observed some shortcomings where the details depicted by the text do not translate well into the generated audio. For example, generating audio for a sentence such as "A man wearing sneakers and a woman wearing high heels are walking together in a church" requires the model to have a complex understanding of both temporal and frequency sequences. The model needs to distinguish between the sounds produced by each object, consider factors such as the similarity in footstep velocity, and account for the long reverberation of sound in a church. Currently, the performance of our audio generation model showed some limitation for complex texts like the example. However, as text-audio multi-modal representation learning advances and larger training datasets are used, we expect the audio generation performance to improve further.

4. CONCLUSION

In this paper, we have presented FALL-E, Gaudio's foley synthesis system. FALL-E employs a cascaded approach with low-resolution spectrogram generation, a super-resolution module, and a vocoder. Our system was submitted to the DCASE 2023 Task 7 - Foley Synthesis Challenge (Track A), and we have reported the objective measure with respect to the official evaluation set. Through our extensive dataset and language model conditioning, as well as prompt engineering, we have achieved high-quality, diverse, and coherent sound generation results.

There is a vast potential for the development of generative AI in the audio domain. As technology continues to advance, new possibilities for sound generation arise, and the potential applications of this technology are vast. For example, in film and game production, foley synthesis could be used to produce more realistic sound effects, saving time and resources compared to traditional foley artistry. We believe that FALL-E, along with other works in the field, will pave the way for future advancements in generative audio technology, and we look forward to the continued development of this exciting area of research.

Acknowledgement

The authors are grateful for Naver D2 Startup Factory and Naver Cloud Platform for supporting the GPU resources for this research.

We would like to highlight the clear arrangement implemented to ensure fairness and prevent any unfair advantage in the task. The conflict of interest with one of the organizers of this task was openly disclosed to the organizers, and the co-organizer affiliated with the institution in question remained uninvolved once the finalists were objectively determined. Additionally, during the subjective evaluation phase, other organizers were kept blind to the submission numbers to maintain impartiality. These measures were put in place to uphold the integrity and impartiality of the task evaluation process.

5. REFERENCES

[1] K. Choi, S. Oh, M. Kang, and B. McFee, "A proposal for foley sound synthesis challenge," *arXiv preprint arXiv:2207.10760*, 2022.

[2] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

[3] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," 2023.

[4] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[5] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[6] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[8] L. Sheng, D.-Y. Huang, and E. N. Pavlovskiy, "High-quality speech synthesis using super-resolution mel-spectrogram," *arXiv preprint arXiv:1912.01167*, 2019.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[12] V. Liu and L. B. Chilton, "Design guidelines for prompt engineering text-to-image generative models," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–23.

[13] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

[14] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021.

[15] T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, and J. L. Banks, "Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening," *Journal of medicinal chemistry*, vol. 47, no. 7, pp. 1750–1759, 2004.

[16] K. Choi, J. Im, L. Heller, B. McFee, K. Imoto, Y. Okamoto, M. Lagrange, and S. Takamichi, "Foley sound synthesis at the dcase 2023 challenge," *In arXiv e-prints: 2304.12521*, 2023.

[17] G. Mittal, J. Engel, C. Hawthorne, and I. Simon, "Symbolic music generation with diffusion models," *arXiv preprint arXiv:2103.16091*, 2021.

[18] A. Muhamed, L. Li, X. Shi, S. Yaddanapudi, W. Chi, D. Jackson, R. Suresh, Z. C. Lipton, and A. J. Smola, "Symbolic music generation with transformer-gans," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 1, 2021, pp. 408–417.

[19] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[20] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.

[21] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[22] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[23] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," 2022.

[24] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[25] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.