TEXT-TO-AUDIO GENERATION: A COMPARATIVE ANALYSIS FOR PERCEPTUAL PERFORMANCE EVALUATION

Emma Coletta, Federico Ferreri, Lorenzo Previati

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano
Piazza Leonardo Da Vinci 32, 20122 Milano, Italy
[emma.coletta, federicoangelo.ferreri, lorenzo.previati]@mail.polimi.it

ABSTRACT

Advancements in the field of Artificial Intelligence Generated Content (AIGC) have led to the development of sophisticated Text-to-Audio (TTA) models. These models synthesize audio from textual descriptions, with potential applications ranging from video game sound effects to acoustic scene setting in movies. This study presents a comparative analysis of five TTA models: *AudioGen*, *AudioLDM* (in two versions: 'm-full'and 'l-full'), and *AudioLDM2* (in two versions: base and 'large'). We evaluated these models using both objective and subjective methods. Objective evaluation involved calculating FAD scores with embeddings extracted using both the VGGish and CLAP models to measure alignment with input text and audio quality of the generated audio datasets. Subjective evaluation was conducted via a web-based MUSHRA listening test administered to participants of varied ages and musical backgrounds. Our findings emphasize strengths and limitations of each generative model and of the proposed automatic perceptual evaluation, revealing its inaccuracy in assessing TTA models whose training datasets do not include the one used to compute the FAD score. This study contributes to the field by providing insights into the accuracy of current automatic quality assessment methods for TTA models. Our implementation and demos are available at https://github.com/federicoalferreri/MAE_Capstone_Text-To-Audio-Generation.git.

Index Terms— audio generation, machine learning, text-to-audio generative models, quality assessment, perceptual evaluation

1. INTRODUCTION

In recent years, AIGC (Artificial Intelligence Generated Content) technology has developed rapidly, giving rise to AI models fully responsible for the creation of digital content such as images, videos, audio, text, etc. without the need for human involvement in the creative process [1]. The growing research interest in building deep learning models capable of synthesizing audio led to the development of many AI-based audio generation models, each focusing on a particular subdomain of the addressed matter such as the generation of music [2], speech [3] or more specific types of sounds. AI-based audio generation systems have many potential applications in fields such as the automatic creation of sound effects for video game or for acoustic scene setting in movies and audio/video productions [4]. As of now, in the audio synthesis field, significant progress has been made in the implementation of Text-To-Audio (TTA) models, sound generative systems that use natural language descriptive prompts to guide the synthesis of the desired audio samples [5].

Making a direct comparison between TTA systems is notoriously difficult due to the usage of different evaluation methodologies and metrics when reporting results of their performance. Motivated by it, with the following experimental study we aim to provide a valid performance analysis of different TTA generative models by comparing objective and subjective perceptual evaluations of them. More specifically, we'll proceed to investigate the performance of the following models: AudioGen [6], AudioLDM [7] (whose analysis is proposed on two different versions of it, '*AudioLDM-m-full*' and '*AudioLDM-l-full*') and AudioLDM2 [4] (in two versions, '*AudioLDM2*' and '*AudioLDM2-large*').

Over time, various metrics have been proposed for automatic music evaluation, some of them prioritizing alignment to the input text, including the CLAP [8] score, whereas other prioritize how close the quality of the generated audio is to a reference 'studio-quality' one, i.e. the FAD (Frechet Audio Distance) [9] score. In the following paper, on the objective quality assessment side, we'll proceed employing both scores to perform automatic evaluation of the audio samples generated by the selected models. On the subjective perceptual evaluation side, we administered a web-based MUSHRA [10, 11] listening test featuring a limited set of generated audio samples to participants of different ages and musical background and collected results later interpreted and compared to the ones deriving from the previously mentioned objective evaluation.

The aim of this study is to verify the reliability of the objective automatic evaluation in relation to the results obtained from the subjective test, which we consider the reference for assessing the actual quality in generating audio of the analyzed models.
Core contributions of this paper are: (i) an overview of the currently used metrics for the evaluation of audio synthesis quality, (ii) a comparative objective analysis of five popular TTA generative systems, (iii) a listening test to perform perceptual quality evaluation of these systems and (iiii) a comparison between objective and subjective metrics in assessing quality of the afore-mentioned models.

2. RELATED BACKGROUND

The development of Text-To-Audio (TTA) models has been significantly influenced by advancements made in Text-To-Image (TTI) generative systems: inspired by the success of GANs (Generative Adversarial Networks) and VAEs (Variational Autoencoders) in managing to synthesize images from textual descriptors [12–14], researchers began exploring the application of similar architectures in the sound synthesis field. Earlier audio generative models include GanSynth [15], Text-To-Speech (TTS) models such as WaveNet [3] and Text-To-Music (TTM) models such as MuseNet [2] which, despite not being strictly TTA model, laid the foundation for subsequent work in their ability to synthesize high-fidelity and coherent audio. More recently, diffusion models gained popularity in generative tasks: initially employed in image generation [14,16], they have soon been adapted for audio generation, offering a new outlook on the synthesis of high-fidelity audio from text descriptions. In this category but not object of our evaluation, particularly relevant models are DiffSound [17], which started investigating the generation of sound conditioned on a text prompt, Make-An-Audio [18] and Tango [19].

2.1. Text-To-Audio Generative Models

This section briefly introduces the TTA models object of our evaluation analysis. Text-To-Audio (TTA) systems are advanced generative models that produce audio outputs conditioned on textual input. The system is fed a text prompt which is first processed to extract meaningful features and then guides the generative process to ensure the resulting audio signal aligns with the given description.

AudioGen [6] leverages autoregressive methods to generate audio sample-by-sample and transformer-based architectures to effectively capture and model long-range dependencies in audio sequences trying to ensure the produced audio is both high-definition and semantically consistent with the input caption.

AudioLDM [7] performance was evaluated on two different versions of the same model, '*AudioLDM-m-full*' and '*AudioLDM-l-full*'. *AudioLDM* leverages Latent Diffusion Models (LDMs) to generate high-quality audio from textual descriptions. Being a diffusion model, audio is represented using a latent space, which is iteratively refined and gradually transformed from random noise into a coherent audio signal incorporating textual conditioning to guide the diffusion progress.

AudioLDM2 [4] is an improved version of the previously described *AudioLDM*. In this newer implementation, the diffusion process has been enhanced to obtain overall better audio quality. It also incorporates more advanced techniques to better capture temporal dependencies and audio details. It maintains the same capability as *AudioLDM* of generating a wide range of audio types, including sound effects and speech. Once again, the evaluation was performed on two different versions of it, '*AudioLDM2*' and '*AudioLDM2-large*'.

2.2. Datasets Generation

Clotho [20] is the reference dataset selected for this study. It was designed for audio captioning and the complete version contains 4981 labeled audio samples, 15 to 30 seconds long, covering various environmental sounds and everyday activities. Audio data to create it was sourced from the FreeSound [21] platform. Each audio clip is associated to 5 different corresponding descriptive captions (for a total of 24905 captions). To conduct our experiment, we chose to use the 'evaluation' portion of the dataset which only features 1045 audio clips and the corresponding 5225 text prompts to limit generation and consequent evaluation time.

Other popular datasets commonly employed in training and evaluating the performance of TTA models are AudioCaps [22], AudioSet [23], WavCaps [24], ESC50 [25], FSD50K [26] and UrbanSound8K [27].

3. EXPERIMENTAL SETUP

In this section, we outline the experimental setup designed to perform the perceptual analysis of the selected generative systems.

3.1. Prompts Selection

To obtain the text prompts needed for audio clips generation and so model evaluation, we implemented the following algorithm to select only one descriptive caption for each audio sample out of the five originally provided by the *Clotho* [20] evaluation dataset:

1. We calculated the frequency of each word across all captions in the original dataset;
2. For each caption, we computed a final score by summing the frequencies of the words that make up the caption. Words such as adverbs, conjunctions, prepositions and auxiliary verbs were excluded from this count;
3. The caption with the lowest overall score, thus the one comprising less frequent words overall, was selected to use in the audio generation step.

This method ensured that the selected captions provided the most specific and detailed descriptions possible, which is critical for generating accurate and contextually relevant audio samples.

3.2. Text-To-Audio Models

The audio generative process was implemented using the models mentioned in Section 2.1. and their setup based on each model's default settings. The notebooks for each model are available and ready to use, after importing the desired input dataset, at the following link: [28].

4. MODEL EVALUATION

Evaluating the performance of TTA models often involves assessing various aspects of the generated audio including its quality and adherence to the input text description.

4.1. Objective Evaluation

Particularly challenging in completing the task of objectively assessing the quality of audio signals, is finding a metric able to evaluate and quantitatively compare different approaches adopted in audio generation with respect to the perceived quality of their output. Standard measures such as *Signal to Noise Ratio* (SNR), *Signal to Distortion Rario* (SDR) or *Signal to Inference Rartio* (SIR), typically used to evaluate signal separation algorithms, fail at taking into account the perceptual quality of the newly produced signal [9]. To overcome this problem, Kilgour et al. revised the *Fréchet Inception Distance* (FID), used to evaluate image generation systems, to introduce the *Fréchet Audio Distance* (FAD).

The FAD is a metric designed to measure how a given audio clip (generated sample) compares to the clean, denoised version of it (target sample) [9] and since its introduction, FAD has become a commonly used objective evaluation metric for estimating quality of audio generative models. A common approach to calculate FAD is to extract audio embeddings using the VGGish model [29], yielding a single FAD score for a whole set of generated music samples. In our experiment, the FAD was computed against the Clotho evaluation dataset [20] as reference.

VGGish is an audio feature extraction model based on the VGG image recognition architecture [30] and trained on a large dataset of YouTube videos as an audio classifier. It was designed to recognize a wide range of acoustic events and transform waveforms into embeddings capturing semantic content of the audio.
A less common approach, that we decided to adopt nonetheless, is computing the FAD using embeddings extracted from the CLAP [8] model which consists of a Contrastive Language-Audio Pretraining pipeline to develop audio representation by combining audio data with natural language descriptions. It's characterized by a flexible feature space (embeddings can adapt to various contexts) and by a multi-modal understanding (its text-audio representations allow it to better capture relationships between text descriptions and corresponding audio). To compute the FAD with CLAP embeddings, we used a smaller portion of the generated audio datasets. A resampling of these samples was also necessary: implementation of the provided notebook [28] requires generated audios to be at a sample rate of 48kHz (original samples were at 16kHz).

There are advantages to both models: VGGish is now a well-established audio feature extractor providing a standard benchmark for audio quality despite its poor sensitivity to temporal features, whereas CLAP captures richer information about the audio, potentially leading to more accurate and sensitive FAD scores.

4.2. Subjective Evaluation

To fulfill the purpose of the study of comparing an objective automatic perceptual evaluation with a subjective one, we conducted a *MUltiple Stimuli with Hidden Reference and Anchor* (MUSHRA) [11] experiment using a compliant web-audio API and administered a listening test to a total of 34 people of diverse ages and musical backgrounds. Participants were asked to evaluate audio clips generated with 10 different input text prompts, from the 5 different analyzed models, twice: once to assess the alignment with the descriptive caption and the second time to evaluate the perceived audio quality of the sample.

Audio clips could be rated according to a five-star likert scale, often used for evaluation purposes of the perceived overall listening experience (OLE) when assessing audio systems [10]. Possible ratings were: 'Bad', 'Poor', 'Fair', 'Good' and 'Excellent'. Moreover, audio samples were selected so that their respective generative captions described acoustic events from different categories and typical of diverse environments and everyday activities. The listening test is momentarily available here[1].

5. RESULTS AND DISCUSSION

In this section, we present and analyze findings from our evaluation analysis of the different TTA generative models. We compare the models performance based on objective metrics presented in Section 4.1 and subjective assessments obtained from the listening test presented in Section 4.2, providing insights into their strengths and weaknesses in generating contextually accurate and high-quality audio. We also provide a comparison between results obtained from FAD perceptual evaluation and the subjective one, highlighting how the former is not particularly reliable in assessing quality of the models whose training datasets do not include the one used as the evaluation set to compute the FAD.

5.1. FAD - Objective Evaluation

Table1 reports the objective metrics comparison between FAD scores obtained for each model and computed with embeddings extracted using both the VGGish [29] and the CLAP [8] models.

Model	**VGGish** ↓	**CLAP** ↓
AudioGEN	**1.80312**	0.33796
AudioLDM-m-full	5.57813	0.56389
AudioLDM-l-full	5.40306	0.46311
AudioLDM2	5.95507	0.42067
AudioLDM2-large	3.91272	**0.33601**

Table 1: *FAD scores of the assessed models computed using embeddings extracted with both the VGGish and CLAP models.*

Firstly, we observe that *AudioGen* has the lowest, and therefore the best, FAD score according to the VGGish model. This is expected, as *AudioGen* has been trained using the *Clotho* dataset, among others. The resulting score computed by us is very close to the one reported in the employed model[2] documentation, equal to 1.77. Regarding the *AudioLDM* model (in all its versions), its improved version *AudioLDM2* demonstrates overall better performance in terms of audio quality, as indicated by its authors. This is reflected by the better FAD scores obtained for both VGGish and CLAP embeddings compared to *AudioLDM*. Additionally, the FAD score generally improves with an increase in model size. It's worth noticing how, compared to *AudioGen*, *AudioLDM* and *AudioLDM2* models achieve higher, and thus worst, FAD scores according to our computation, likely because they were not trained using the Clotho dataset. Moreover, our computed FAD VGGish scores show worse perceptual audio quality compared to the values reported in their respective original documentation. On a more general note, CLAP embeddings outperform VGGish's, potentially indicating that the joint audio-text embeddings from CLAP capture audio features and text-audio alignment more accurately than the VGGish model: this may be be due to VGGish's limitations in capturing temporal dynamics of audio. Comparing scores obtained with VGGish and CLAP embeddings, the quality assessment results appear consistent in classifying *AudioLDM* as the worst generative model among those analyzed. In contrast, *AudioGen* and *AudioLDM2-large* emerge as the most successful ones in terms of achieving the best quality, with each achieving the best scores depending on which embedding model was used to compute the FAD.

5.2. MUSHRA - Subjective Evaluation

In this section, we continue to analyze results obtained from the perceptual subjective evaluation described in Section 4.2. At first, we will make a distinction between the two investigated aspects:

[1] https://text-to-audio-subjectivetest.000webhostapp.com/

[2] https://github.com/gudgud96/frechet-audio-distance

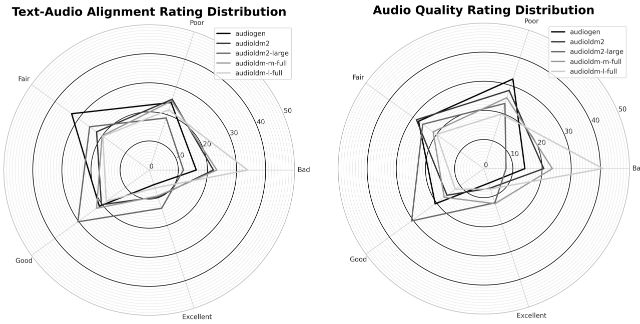

Figure 1: *Polar distribution diagram of the each model's quality assessment based on text-audio (left) alignment and audio quality (right) obtained with results from the subjective perceptual evaluation test.*

the alignment of the audio with the descriptive text prompt used to generate it and the overall audio quality of each sample as perceived by the test participants. Subsequently, we will compare the FAD scores presented in the previous Section (5.1) with the undifferentiated results obtained from the subjective test. All considerations were based on a percentage obtained by performing a weighted average, where each rating was assigned a weight from 1 (for 'Bad') to 5 (for 'Excellent').

Figure 5.2 shows the distribution-per-model quality assessment based on adherence of the generated audio with the text prompt responsible for its generation (on the left) and on overall audio quality (on the right). First of all, we can observe how *AudioLDM-l-full* has the highest concentration of 'Bad' ratings when assessing audio-text alignment. Conversely, *AudioLDM2-large* received the highest number of 'Good' and 'Excellent' ratings compared to all other models. *AudioLDM-m-full* and *AudioLDM2* show rather similar results: in both cases, evaluations were evenly distributed across all possible ratings, which contrasts with the results from objective analysis in the previous section that indicated these were the worst models among all. Regarding *AudioGen*, the majority of ratings given by participants fall within the middle range ('Poor', 'Fair' and 'Good') with no prevalence of either 'Bad' or 'Excellent' ratings. This suggests that *AudioGen* is not the best model among those analyzed, contrary to what resulting FAD scores might imply.

Proceeding with the perceptual analysis of audio quality, the observations are consistent with those made for audio-text alignment: *AudioLDM2-large* stands out, once again, as the model with the highest concentration of positive ratings, showing a notable peak of 'Good' evaluations. On the other hand, *AudioLDM-l-full* proves to be the worst model for assessing audio sample quality as well, with a rather significant prevalence of 'Bad' (41%) and 'Poor' (20%) ratings overall. *AudioGen* and *AudioLDM2* display once again a balanced distribution across all possible evaluations.

Combining all results from the subjective test, it is clear that *AudioLDM2-large* is the model with the highest concentration of positive ratings; in particular, it has a rather high percentage of 'Excellent' ratings compared to other models (with '*AudioLDM-m-full*' being the closest). As shown in Figure 2, *AudioLDM2* and *AudioLDM-m-full* have the most balanced distribution of ratings among all possible options, while *AudioGen* has most of its ratings in the middle range, with very few extremes. *AudioLDM-l-full* is confirmed as the worst model among all. On a more general note, the ratings are overall skewed towards negative evaluations: across all models, there is a low percentage of 'Excellent' ratings.

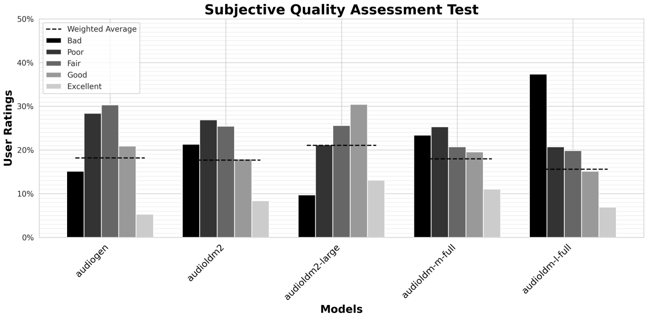

Figure 2: *Subjective test quality assessment obtained combining results relative to both audio quality and text-audio alignment.*

Comparing the results obtained from the objective and subjective perceptual tests, we observe that *AudioGen*, which was the only model (among those analyzed) trained with the Clotho dataset, achieves remarkably good FAD scores contrary to its rather mediocre ratings in the subjective test. On the other hand, there is consistency between objective and subjective test results in evaluating *AudioLDM2-large*, which is confirmed by either as the best model for both text-audio alignment and overall audio quality. From this, we can deduce that, training the model with the same dataset used as reference to compute the FAD, influences significantly the final score, positively affecting the automatic quality assessment of the model's output (in our case, *AudioGen*'s). It is reasonable to assume that if the other models had been trained with the same dataset used for evaluation, they would have achieved overall higher scores in the objective test (possibly even better than *AudioGen*'s), consistent with the results from the subjective test. Another difference between the scores obtained from the objective test and those from the subjective test is that the former do not reveal the negative rating disparity of *AudioLDM-l-full*, whose performance is evaluated as average whereas the subjective test indicates that it is the worst among all models for both audio quality and text-audio alignment.

6. CONCLUSIONS

With this study, we analyzed and compared TTA models, assessing their performance based on audio quality and adherence of audio clips with the textual descriptions used to generate them. Our findings reveal that the objective evaluation is not particularly reliable for assessing models whose training dataset do not include the one used for FAD evaluation. However, it is overall capable of identifying effectively the best model for both text-audio alignment and audio quality among those that do not employ the reference evaluation dataset in the training process. Additionally, the objective test struggles more to distinguish between models with intermediate performance and those yielding significantly poorer results but proves accurate in extracting relevant embeddings for FAD evaluation, as shown by the relatively good FAD scores associated with *AudioGen*, model trained with the same dataset (among others) used for evaluation.

7. REFERENCES

[1] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226*, 2023.

[2] A. Pal, S. Saha, and A. Ramalingam, "Musenet : Music generation using abstractive and generative methods," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, pp. 784–788, 04 2020.

[3] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[4] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[5] F. Ronchini, L. Comanducci, and F. Antonacci, "Synthesizing soundscapes: Leveraging text-to-audio models for environmental sound classification," *arXiv preprint arXiv:2403.17864*, 2024.

[6] 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.

[7] 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," *arXiv preprint arXiv:2301.12503*, 2023.

[8] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," 2024. [Online]. Available: https://arxiv.org/abs/2211.06687

[9] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," 2019. [Online]. Available: https://arxiv.org/abs/1812.08466

[10] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra—a comprehensive framework for web-based listening tests," 2018.

[11] Wikipedia contributors, "Mushra — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=MUSHRA&oldid=1178809176, 2023, [Online; accessed 28-June-2024].

[12] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016. [Online]. Available: https://arxiv.org/abs/1605.05396

[13] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021. [Online]. Available: https://arxiv.org/abs/2102.12092

[14] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, vol. 2, no. 3, p. 8, 2023.

[15] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio synthesis," *arXiv preprint arXiv:1902.08710*, 2019.

[16] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[17] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1720–1733, 2023.

[18] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 13 916–13 932.

[19] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction-tuned llm and latent diffusion model," *arXiv preprint arXiv:2304.13731*, 2023.

[20] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," 2019. [Online]. Available: https://arxiv.org/abs/1910.09387

[21] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 411–412.

[22] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[23] 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.

[24] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *arXiv preprint arXiv:2303.17395*, 2023.

[25] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[26] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[27] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[28] E. Coletta, F. Ferreri, and L. Previati, “Text-to-audio generation: A comparative analysis for perceptual performance evaluation,” https://github.com/federicoalferreri/MAE_Capstone_Text-To-Audio-Generation, 2024.

[29] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, “Cnn architectures for large-scale audio classification,” 2017, https://github.com/tensorflow/models/tree/master/research/audioset. [Online]. Available: https://arxiv.org/abs/1609.09430

[30] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015. [Online]. Available: https://arxiv.org/abs/1409.1556