# Vivid Background Audio Generation based on Large Language Models and AudioLDM

*Yiwei Liang*[1]*,Ming Li*[1]

[1]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Data Science Research Center, Duke Kunshan University, Kunshan, China

yiwei.liang@dukekunshan.edu.cn, ming.li369@dukekunshan.edu.cn

## Abstract

This paper describes a background audio and speech generation system for the Inspirational and Convincing Audio Generation Challenge 2024. Our system mainly includes three modules, namely, a text-to-speech (TTS), speech synthesis baseline, background text description extraction based on large language models, and the corresponding background audio generation based on latent diffusion. We compare the influence of text description extraction on the degree of correlation between background audio and its corresponding speech. At the same time, the results of different large language models on the background audio generated after description extraction are compared. We also propose an alternative evaluation metric named Overall Correlation Quality Score (OCQS) to evaluate the relevance and naturalness between speech text and its background audio. With the above evaluation metric, we test multiple models and find that the background audio generated by extracted speech text and summarized by large language models achieve better quality.

**Index Terms**: background audio generation, large language model, speech synthesis

## 1. Introduction

In recent years, the rapid development of science and technology has made significant progress in speech synthesis and background audio generation technology. However, generating natural and vivid background audio still faces many challenges and barriers.

Emotional speech synthesis technology has also made significant progress in recent years. For example, such as QI-TTS [1], and StyleTTS2 [2], these models can generate more natural and emotional speech, improving the experience and effect of human-computer interaction. Models such as SpeechT5 [3] provide a good foundation of pre-trained models for high-quality speech synthesis [4].

StableDiffusion [5] has made great progress in image synthesis, and it is proven to have strong expression ability and efficient training ability [6], which inspired researchers to apply latent diffusion models (LDMs) in audio synthesis [7]. The audio generation models based on the diffusion models have shown good results in the background audio generation tasks. Representative models in recent years include AudioLDM [8] and AudioLDM2 [9]. These models utilize latent diffusion to generate high-quality background audios based on the text prompts and adapt to various complex acoustic environments and needs. At the same time, latent diffusion for language [10] and NaturalSpeech 2 [11] show that the diffusion model also has a good effect on speech synthesis. However, it should be noted that although the above text-to-speech (TTS) and text-to-audio (TTA) models both use the diffusion model, there is still no single unified model that can generate the matching background audio when generating speech.

The superior performance of large language models in terms of text information processing capabilities has been validated in many tasks [12]. Tango [13] and Tango 2 [14] show that the application of a large language model as a text-encoder such as FLAN-T5 [15] in the field of TTA generation also has great improvement and promising prospectives.

The Inspirational and Convincing Audio Generation Challenge 2024 (ICAGC 2024) [16] focuses on pushing the boundaries of current TTS technology by enhancing the emotional depth and realism of generated audio. Its track 2 focuses on creating a truly immersive experience through background audio and speech fusion generation. Adding background audio can significantly enhance the user experience and make it more engaging [17]. However, no unified model currently exists for direct background audio generation with speech. Therefore, the approach is divided into two steps: first, synthesize the speech and relevant background audio, and second, merge them together.

In this work, we propose a system that combines text-to-speech (TTS), text-to-audio (TTA), and a large language model (LLM) together, which generates audio that has both background audio and speech. At the same time, the effectiveness of the system as well as the quality of the generated audio are validated through our proposed Overall Correlation Quality Score (OCQS) metric.

We adopt the AudioCaps dataset [18] to reproduce the results of AudioLDM2 [9], use the open source tool of GPT-SoVITS [1] to complete the speech synthesis work, and utilize of the large language model to extract and summarize the text description. Then, background audio is generated through the reproduced AudioLDM2 model. Finally, we evaluate the performance of multiple large language models for comparison purposes.

The remainder of this paper is organized as follows: Section 2 introduces the dataset that we used in this paper. Section 3 describes the background audio and speech generation system in detail. Section 4 describes the evaluation metrics and discusses the experimental results. Conclusions are summarized in Section 5.

## 2. Dataset

In this paper, we use the AudioCaps dataset to train and reproduce the AudioLDM2 model. AudioCaps is a comprehensive dataset designed for the task of audio captioning, which involves generating natural language descriptions for various au-

[1] https://github.com/RVC-Boss/GPT-SoVITS

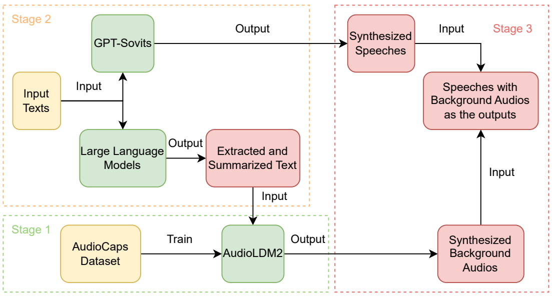

Fig. 1: *An overview of our system.*

dio clips [18].

The ICAGC2024 dataset contains 358 speech training utterances from 15 speakers and 389 sentences as the testing test data covering various situations such as poetry recitation and reading audiobooks [16]. The audios we evaluate are generated from the input texts in the testing set of this dataset.

## 3. System Description

### 3.1. Overview

The main challenge in generating natural and contextual background audio is how to extract useful background audio descriptions from the input text and pass it to the Text-to-Audio (TTA) model to generate background audio. Given the successful application of large language models in natural language processing tasks [19], we directly utilize large language models to process and summarize text information to generate practical and contextual descriptions of background audio.

Furthermore, currently, there is a lack of recognized objective indicators to evaluate the correlation and naturalness of the generated background audio along with the relevant speech. To fill this gap, this paper proposes an Overall Correlation Quality Score (OCQS) to evaluate the degree of correlation between the background audio and the associated speech content in terms of relevance and naturalness. OCQS considers the matching degree, contextual consistency, and naturalness of the background audio and speech. By analyzing the results, we can evaluate the performance of the model and make ablation studies.

In terms of experimental design, this paper reproduces the results of AudioLDM2 using the AudioCaps dataset for training and performs the speech synthesis work using the GPT-SoVITS toolkit. The large language model is used to extract the text information and generate a description of the background audio. The background audio is then generated through our trained AudioLDM2 model and the generated results are evaluated by OCQS.

Fig. 1 briefly illustrates the architecture of our system. It illustrates the three stages of our work, each representing a different focus of our efforts. In Stage 1, we train the AudioLDM2 model using the AudioCaps dataset. Stage 2 is the core of our work, where we send the same input text to the speech synthesis model GPT-SoVITS and a large language model. GPT-SoVITS is responsible for completing the TTS task, and we use the large language model's text processing capabilities to obtain a possible background audio description that is consistent with the corresponding input text. Then, we sent the background audio descriptions extracted by the large language model into the reproduced AudioLDM2 model to generate the background audio. Finally, in Stage 3, we merge the generated background audio with the synthesized speech together to obtain the final output with mixed background audio and speech.

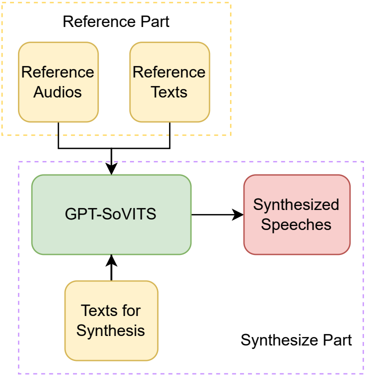

Fig. 2: *The workflow of Text-to-Speech*

### 3.2. Audio Generation Model

We train an AudioLDM2 model with the AudioCaps dataset. Our AudioLDM2 training process includes the following steps: First, we map the conditional information (in our case, text) to the Language of Audio (LOA) [9] using models such as CLAP [20]. This step generates estimations of the LOA, which are crucial for the subsequent audio synthesis. Next, Audio Masked Autoencoder (AudioMAE) [21] extracts features from the unlabeled audio data. Then, an improved Unet [9] is used to achieve self-supervised learning [22]. Lastly, a pre-trained HiFiGAN model [21] to convert the generated spectrum into high-quality audio.

### 3.3. The Synthesis Pipeline

We divide this section into two parts. The first part introduces the synthesis of emotional speech, and the second part is about the generation of background audio descriptions through a large language model.

#### *3.3.1. Emotional Speech Synthesis*

We choose the GPT-SoVITS[2] to perform emotional speech synthesis. GPT-SoVITS is a powerful multi-language text-to-speech (TTS) toolkit, that can synthesize high-quality speech by using short speech samples as reference audio and relevant text prompts as reference text, which fits well with our work needs.

Fig. 2 briefly illustrates our TTS workflow. We first select a clean speech audio with proper time duration and without noise interference as the reference for the speech synthesis model. After we choose the appropriate reference audio for each testing speaker, we adopt the GPT-SoVITS toolkit to perform the speech synthesis task.

We use the synthesized speech as part of our intermediate products, which will be mixed with the synthesized background audio in subsequent stages.

[2] https://github.com/RVC-Boss/GPT-SoVITS

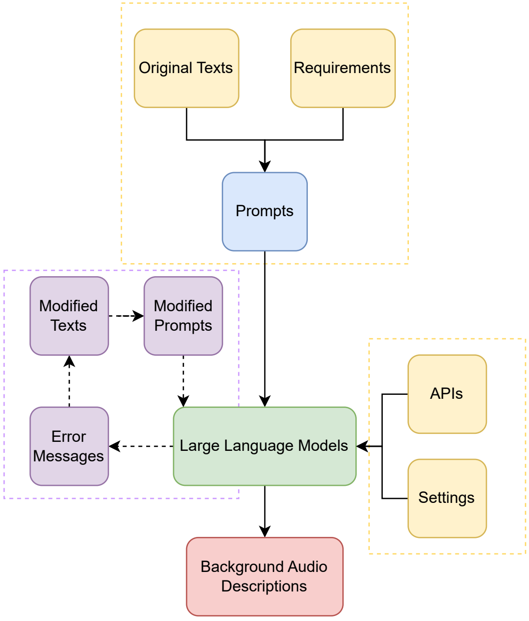

Fig. 3: *The workflow of background audio description generation.*

### *3.3.2. Background Audio Description Generation*

Since it is necessary to generate the corresponding background audio description through the input text, we select several different large language models to generate the background audio descriptions.

Fig. 3 briefly shows the workflow for generating audio descriptions. We select a total of four large language models related to our research for experimental testing, which are hunyuan-standard[3] developed by Tencent, ERNIE3.5-8K[4] developed by Baidu, Spark3.5Max[5] developed by iFlytek, and GPT-4o-mini[6] developed by OpenAI. We start by giving all of the large language models the same working instructions to achieve an objective prompt input. We take into account that using large language models can cause errors such as network connection errors, restricted words, etc., which will not return the desired background audio description correctly, so we make additional correction efforts when we encounter errors. By analyzing the returned error messages, we replace some words forbidden by the large language model with synonyms and regenerate the parts that cannot be generated properly due to the fluctuations of the genetic network. Eventually, we use the AudioLDM2 model with the description as input to generate the background audio.

### 3.4. Audio Mixing

We use the Pydub[7] library for audio mixing. We considered that when mixing audio if the generated background audio is too loud, it will become difficult to understand the speech content correctly. When we mix the audio, we reduce the background audio volume to a certain extent, so that the mixed audio can clearly hear the human voice, to achieve a better overall effect.

## 4. Results and Discussions

### 4.1. Evaluation Metric

Currently, there is no golden standard for evaluating the degree of Correlation between background audio and verbal speech in terms of relevance and naturalness. Therefore, we propose an alternative evaluation score, called the Overall Correlation Quality Score (OCQS). It is used to assess the degree to which the background audio relates to the content of the speech in terms of relevance and naturalness. The evaluation criterion of relevance requires the native speaker to comprehensively consider the background scenes in which the speech may occur, so as to evaluate whether the background audio matches with the current estimated acoustic scene. The degree of naturalness requires the judge to consider whether the background audio is harmonious in the audio. The judge mainly performs evaluation based on relevance, while naturalness is also considered. The scoring of OCQS requires the judge to make a score between one and five points.

Eight different native speakers used the OCQR rating to judge each audio generated by our system. The arithmetic mean of the scores given by these judges was labeled as the score for each audio.

### 4.2. Scoring validity

If more than one raters think there exists something wrong with the audio, such as missing audio or having difficulty in understanding the sentence, the audio will be marked as problematic audio, and the rating of the audio will be considered invalid.

We tested four different large language models on the required tasks in two parts. Due to the time and cost of scoring, we did not score all the produced audio. So only the first part scored all the generated audios, while the second part only randomly chose 100 generated audios for scoring.

The first part contains 389 audio per task, and only one audio file is reported as problematic. In the second part, each task contained 100 randomly selected audio, and only three files were reported as problematic.

Since the second part is scored by different groups of judges from the first part, there may be differences in determining whether there is difficulty in understanding the sentence content. According to the feedback during the scoring process, the judges of the second part had more doubts about the understanding of the sentences, so more files were reported as problematic.

### 4.3. Results

Fig 4 shows an example of audio generated by our system.

According to the experimental requirements, we define method 1 as requiring the large language model to generate the background audio description and define method 2 as directly translating the input text as the background audio description.

The results of the first part are shown in Tab. 1. Using the average score as the reference standard, method 1's results are significantly better than method 2's, GPT-4o-mini's performance on method 2 is slightly better than Spark3.5Max's, however, Spark3.5Max's performance on method 1 is much better than GPT-4o-mini's.

[3] https://hunyuan.tencent.com/
[4] https://qianfan.cloud.baidu.com/
[5] https://xinghuo.xfyun.cn/
[6] https://openai.com/
[7] https://github.com/jiaaro/pydub

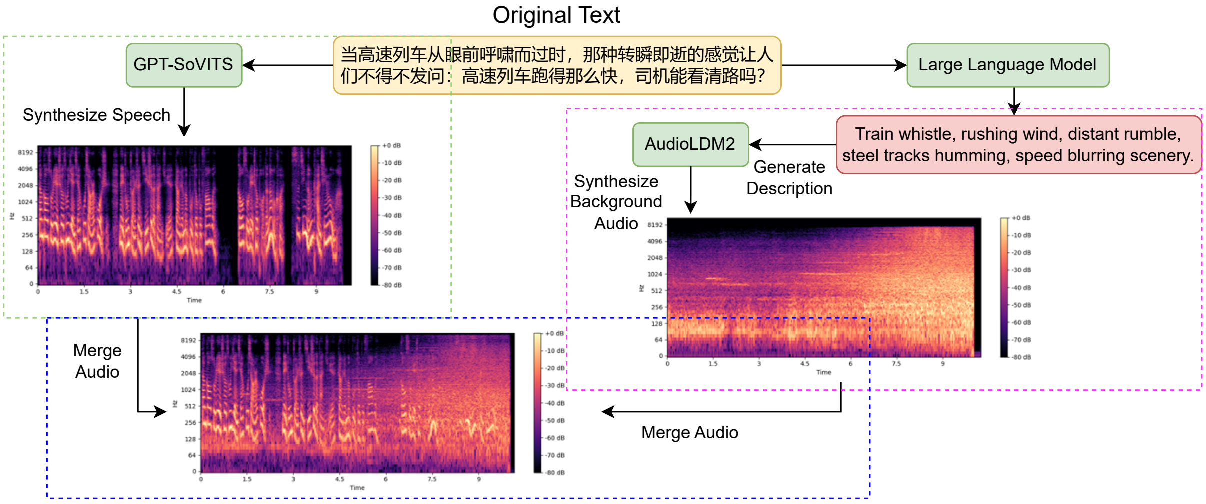
Original Text
当高速列车从眼前呼啸而过时，那种转瞬即逝的感觉让人们不得不发问：高速列车跑得那么快，司机能看清路吗？
GPT-SoVITS
Large Language Model
Synthesize Speech
AudioLDM2
Train whistle, rushing wind, distant rumble, steel tracks humming, speed blurring scenery.
Generate Description
Synthesize Background Audio
Merge Audio
Merge Audio
Hz
8192
4096
2048
1024
512
256
128
64
0
0
1.5
3
4.5
6
7.5
9
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

Fig. 4: *An example of our proposed vivid audio generation system*

Tab. 1: *Results in Part I evaluation with 389 sentences*

| Models&Methods | Average Score |
|---|---|
| `GPT-4o-mini:method 1` | 2.65 |
| `GPT-4o-mini:method 2` | 2.19 |
| `Spark3.5 Max:method 1` | 3.03 |
| `Spark3.5 Max:method 2` | 1.99 |

Tab. 2: *Results in Part II evaluation with 100 sentences*

| Models&Methods | Average Score |
|---|---|
| `Hunyuan-standard:method 1` | 2.80 |
| `Hunyuan-standard:method 2` | 2.34 |
| `ERNIE-3.5-8K:method 1` | 2.90 |
| `ERNIE-3.5-8K:method 2` | 2.25 |

In the second part, as shown in Tab. 2, the result of method 1 is still better than that of method 2, showing that an additional LLM-based text description for an acoustic scene is effective.

### 4.4. Discussions

Considering the performance of different large language models, we find that GPT-4o-mini is only slightly better than Spark3.5Max on method 2, but performs worse than the other large language models with method 1. We believe that the reason might be models are developed in China and thus have more powerful Chinese language processing capabilities.

## 5. Conclusion

This paper proposes a vivid background audio and speech generation system as well as an alternative evaluation metric for measuring the relevance and naturalness between speech text and its background audio. Our work finds that using generalized textual description leads to higher relevance and naturalness compared to directly using original input text for the AudioLDM2 model. We tested 4 different large language models in this study. Among the tested models, the background audio generated by text descriptions from Spark3.5 MAX achieves the best performance.

## 6. Acknowledgements

This research is funded in part by the DKU foundation project ”Emerging AI technologies for natural language processing” and DKU Summer Research Scholars program. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

## 7. References

[1] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, “Qi-tts: Questioning intonation control for emotional speech synthesis,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[2] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, “Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 19 594–19 621. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/3eaad2a0b62b5ed7a2e66c2188bb1449-Paper-Conference.pdf

[3] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, “Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing,” 2022. [Online]. Available: https://arxiv.org/abs/2110.07205

[4] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, “Neural codec language models are zero-shot text to speech synthesizers,” 2023. [Online]. Available: https://arxiv.org/abs/2301.02111

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” 2022. [Online]. Available: https://arxiv.org/abs/2112.10752

[6] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, “Listen, denoise, action! audio-driven motion synthesis with diffusion models,” *ACM Trans. Graph.*, vol. 42, no. 4, jul 2023. [Online]. Available: https://doi.org/10.1145/3592458

[7] Y. Yuan, H. Liu, X. Liu, X. Kang, P. Wu, M. D. Plumbley, and W. Wang, "Text-driven foley sound generation with latent diffusion model," 2023. [Online]. Available: https://arxiv.org/abs/2306.10359

[8] 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," *Proceedings of the International Conference on Machine Learning*, pp. 21 450–21 474, 2023.

[9] 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*, vol. 32, pp. 2871–2883, 2024.

[10] J. Lovelace, V. Kishore, C. Wan, E. Shekhtman, and K. Q. Weinberger, "Latent diffusion for language generation," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 56 998–57 025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/b2a2bd5d5051ff6af52e1ef60aefd255-Paper-Conference.pdf

[11] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," 2023. [Online]. Available: https://arxiv.org/abs/2304.09116

[12] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," 2024. [Online]. Available: https://arxiv.org/abs/2307.06435

[13] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction-tuned llm and latent diffusion model," 2023. [Online]. Available: https://arxiv.org/abs/2304.13731

[14] N. Majumder, C.-Y. Hung, D. Ghosal, W.-N. Hsu, R. Mihalcea, and S. Poria, "Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization," 2024. [Online]. Available: https://arxiv.org/abs/2404.09956

[15] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[16] R. Fu, R. Liu, C. Qiang, Y. Gao, Y. Lu, T. Wang, Y. Li, Z. Wen, C. Zhang, H. Bu, Y. Liu, S. Shi, X. Qi, and G. Li, "Icagc 2024: Inspirational and convincing audio generation challenge 2024," 2024. [Online]. Available: https://arxiv.org/abs/2407.12038

[17] D. Lobo, J. Dcruz, L. Fernandes, S. Deulkar, and P. Karunakaran, "Emotionally relevant background music generation for audiobooks," in *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, 2021, pp. 1–6.

[18] 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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 119–132. [Online]. Available: https://aclanthology.org/N19-1011

[19] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, 2023.

[20] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[21] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," 2023. [Online]. Available: https://arxiv.org/abs/2207.06405

[22] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep resunet for music source separation," 2021. [Online]. Available: https://arxiv.org/abs/2109.05418