A Retrieval Augmented Approach for Text-to-Music Generation

Robie Gonzales[1] **Frank Rudzicz**[1,2]

[1]Dalhousie University [2]Vector Institute
{robie.gonzales,frank}@dal.ca

Abstract

Generative text-to-music models such as MusicGen are capable of generating high fidelity music conditioned on a text prompt. However, expressing the essential features of music with text is a challenging task. Furthermore, the limited set of text-music pairs leads to distributional shift, resulting in a consistent audio quality degradation with underspecified prompts. In this paper, we present a retrieval-augmented approach for text-to-music generation. We first pre-compute a dataset of text-music embeddings obtained from a contrastive language-audio pretrained encoder. Then, given an input text prompt, we retrieve the top k most similar musical aspects and augment the original prompt. This approach consistently generates music of higher audio quality as measured by the Frechét Audio Distance. We compare different retrieval strategies and find that augmented prompts dislay high text adherence Our findings show the potential for increased control in text-to-music generation.

1 Introduction

Modeling discrete representations of audio obtained from a neural audio codec has been an effective approach in tasks such as audio generation (Borsos et al., 2023; Kreuk et al., 2023), speech synthesis (Wang et al., 2023; Zhang et al., 2024) and self-supervised learning (Pepino et al., 2023). In particular, text-to-music generation (Copet et al., 2023; Agostinelli et al., 2023; Lam et al., 2023; Huang et al., 2023a; Schneider et al., 2024; Liu et al., 2023) has seen widespread adoption.

Despite their impressive capabilities, these models still suffer from distributional shifts, where underspecified user prompts lead to an audio quality degradation. Furthermore, constructing text prompts that accurately capture the user's creative intent while also expressing the essential features of music remains a challenge.

Inspired by the success of retrieval augmented generation (RAG) in natural language processing tasks, we present a retrieval augmented approach for text-to-music generation. While relatively simple, we show our approach consistently generates music of higher audio quality, while also displaying high text adherence. Our findings show potential for increased control in text-to-music generation.

2 Related Work

2.1 Music Generation

Recent generative music models can be roughly separated into two categories: transformer-based and diffusion-based models. MusicLM (Agostinelli et al., 2023) adopts a similar approach to AudioLM (Borsos et al., 2023), which represents audio using multiple streams of "semantic tokens" and "acoustic tokens" obtained from SoundStream (Zeghidour et al., 2021). MusicGen (Copet et al., 2023) adopts a single stage approach, where a transformer decoder is trained to predict multiple streams of discrete audio tokens using codebook interleaving patterns. MAGNeT (Ziv et al., 2024) extends this approach by introducing a masking schedule during training in which spans of tokens are predicted.

Conversely, diffusion models such as MeLoDy (Lam et al., 2023), Moûsai (Schneider et al., 2024), and AudioLDM (Liu et al., 2023) operate on learned, continuous representations of the audio signal. DITTO (Novack et al., 2024) and Music ControlNet (Wu et al., 2024) enable tailored music creation by directly optimizing control features in the latent space, whereas Mustango (Melechovsky et al., 2024) integrates textual metadata controls within the reverse diffusion step. In this work, we focus on the transformer based MusicGen (Copet et al., 2023).

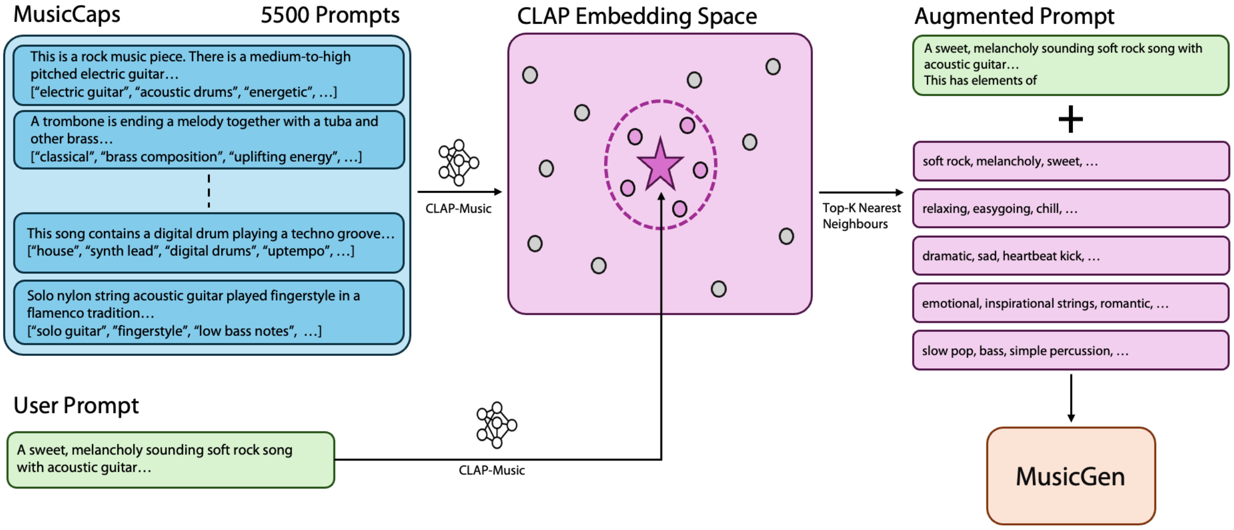
MusicCaps
5500 Prompts
This is a rock music piece. There is a medium-to-high pitched electric guitar…
["electric guitar", "acoustic drums", "energetic", …]
A trombone is ending a melody together with a tuba and other brass…
["classical", "brass composition", "uplifting energy", …]
This song contains a digital drum playing a techno groove…
["house", "synth lead", "digital drums", "uptempo", …]
Solo nylon string acoustic guitar played fingerstyle in a flamenco tradition…
["solo guitar", "fingerstyle", "low bass notes", …]
CLAP-Music
CLAP Embedding Space
Top-K Nearest Neighbours
Augmented Prompt
A sweet, melancholy sounding soft rock song with acoustic guitar…
This has elements of
+
soft rock, melancholy, sweet, …
relaxing, easygoing, chill, …
dramatic, sad, heartbeat kick, …
emotional, inspirational strings, romantic, …
slow pop, bass, simple percussion, …
User Prompt
A sweet, melancholy sounding soft rock song with acoustic guitar…
CLAP-Music
MusicGen

Figure 1: Overview of our retrieval augmented approach. We encode the text captions of MusicCaps using embeddings from CLAP. Given an input text prompt, we retrieve the top k most similar items. We extract their musical aspects list and concatenate them to the original prompt. This is fed as input to MusicGen for text-to-music generation.

2.2 Retrieval Augmented Generation

Retrieval augmented generation (RAG) has been a popular approach for integrating external knowledge from a retrieval module into a parametric language model, particularly for knowledge intensive tasks (Lewis et al., 2020). In this framework, contextually relevant documents from an external corpus are retrieved according to a query. This information is then augmented to the original input and guides the generation process. While there has been some work in applying RAG for general text-to-audio generation (Yuan et al., 2024) and speech (Wang et al., 2024), no work yet has focused on text-to-music generation.

3 Methodology

3.1 Dataset

For our experiments using retrieval and text-to-music generation, we use the MusicCaps dataset (Agostinelli et al., 2023) which consists of roughly 5500 text-music pairs. Each 10-second music clip is paired with a free-text caption describing the music (*This is a rock music piece*), and a list of musical aspects describing genre, mood, instrumentation, etc (*electric guitar, acoustic drums, energetic*).

3.2 Retrieval

The goal of the retrieval module is to retrieve a set of textual music aspects that are similar to the input text prompt. We first pre-compute a dataset of text embeddings obtained from a contrastive language-audio pre-trained encoder (CLAP) (Wu et al., 2023). We use the `music-audio` checkpoint [1], which is trained on AudioSet (Gemmeke et al., 2017), LAION-Audio-630k (Wu et al., 2023), and a dataset of music samples. We encode each free-text caption in MusicCaps which results in a fixed sized 512-dimensional embedding. To store these embeddings, we use Spotify's Annoy [2], an approximate nearest neighbour search library.

Given an input text prompt, we retrieve the top k most similar captions ranked by Euclidean distance. We then extract their musical aspects and perform preprocessing to remove duplicates and words that signal low quality (*low quality, poor audio quality, amateur recording*). Finally, we combine all retrieved musical aspects, prefix them with *"This has elements of "* and concatenate them to the original prompt.

We experiment with various retrieval strategies. We vary the number of retrieved items using $k = 3, 5, 10$. We also experiment with retrieving similar items using the CLAP text embedding of the musical aspect list, as well as retrieving random musical aspects.

[1] `https://huggingface.co/lukewys/laion_clap/blob/main/music_audioset_epoch_15_esc_90.14.pt`

[2] `https://github.com/spotify/annoy`

Method	$\text{FAD}_{\text{CLAP-Audio}} \downarrow$	$\text{FAD}_{\text{CLAP-Music}} \downarrow$	$\text{FAD}_{\text{VGGish}} \downarrow$	KL ↓	$\text{CLAP}_{\text{score}} \uparrow$
Unconditional	0.4668 ± 0.0061	0.5063 ± 0.0033	7.1027 ± 0.0048	2.1013 ± 0.0027	-
First Aspect	0.4055 ± 0.0024	0.4401 ± 0.0009	5.3287 ± 0.0029	2.0567 ± 0.0113	0.1038 ± 0.0063
First Sentence	0.3520 ± 0.0031	0.4055 ± 0.0013	4.8330 ± 0.0024	1.5212 ± 0.0031	0.0735 ± 0.0054
Full Caption	0.3443 ± 0.0026	0.4027 ± 0.0007	4.7895 ± 0.2830	1.3044 ± 0.0011	0.0997 ± 0.0113
Caption's Nearest Neighbours					
Augmented ($k = 3$)	0.3363 ± 0.0021	0.4093 ± 0.0009	4.8390 ± 0.0021	1.3485 ± 0.0017	**0.2863** ± 0.0026
Augmented ($k = 5$)	**0.3189** ± 0.0014	**0.3878** ± 0.0004	**4.3458** ± 0.0028	1.2556 ± 0.0014	0.2810 ± 0.0121
Augmented ($k = 10$)	0.3496 ± 0.0130	0.4116 ± 0.0012	4.6627 ± 0.0017	1.3538 ± 0.0013	0.2854 ± 0.0070
Aspect's Nearest Neighbours					
Augmented ($k = 3$)	0.3519 ± 0.0049	0.4187 ±, 0.0102	5.1123 ± 0.0007	1.2776 ± 0.0059	0.2756 ± 0.0024
Augmented ($k = 5$)	0.3407 ± 0.0068	0.4131 ± 0.0087	4.9794 ± 0.0076	1.3611 ± 0.0112	0.2790 ± 0.0015
Augmented ($k = 10$)	0.3507 ± 0.0178	0.4095 ± 0.0100	4.8538 ± 0.0390	**1.2272** ± 0.0011	0.2688 ±0.0032
Random Aspects					
Augmented $k = 3$)	0.3640 ± 0.0033	0.4403 ± 0.0018	5.8419 ± 0.0006	1.6052 ± 0.0042	0.2453 ± 0.0031
Augmented ($k = 5$)	0.4433 ± 0.0270	0.4511 ± 0.0024	6.1379 ± 0.0041	1.7267 ± 0.1520	0.2358 ± 0.0026
Augmented ($k = 10$)	0.4159 ± 0.0198	0.4801 ± 0.0028	6.6805 ± 0.0023	1.5736 ± 0.0371	0.2199 ± 0.0033

Table 1: Quantitative evaluation results. Mean values and 95% confidence intervals are reported. For the augmented caption, we experiment with retrieving the caption's nearest neighbours, the musical aspect's nearest neighbours and random aspects. A low FAD score indicates the generated music is plausible. A low KL score indicates the generated music shares similar concepts with the reference set. A high CLAP score indicates the generated music adheres to the text prompt.

3.3 Text-to-Music Generation

An audio language model is composed of two components: (i) a compression model, which handles a mapping between audio signals and discrete audio tokens, and (ii) a transformer decoder language model, which operates on these audio tokens. To facilitate text conditioning, a pre-trained text encoder is integrated into the cross-attention blocks of the transformer decoder.

Given a discrete representation of the audio signal z, our goal is to model the joint probability distribution $p_\theta(z \mid y)$, where y is a semantic representation of the condition. This can be computed as a product of its conditional probabilities:

$$p_\theta\left(z_1, \ldots, z_n \mid y\right) = \prod_{i=1}^{n} p_\theta\left(z_i \mid z_1, \ldots, z_{i-1}, y\right) \tag{1}$$

In this work, we are interested in the effect of augmenting y with relevant musical information and how it affects the generation process.

We generate baseline music samples using several methods: (1) no text prompt (unconditional), (2) using only the first musical aspect in the list, (3) using only the first sentence in the text caption, (4) the full text caption.

3.4 Evaluation

Evaluating generative music models remains a challenge (Gui et al., 2024). Given we are interested in the effect of an augmented prompt in text-to-music generation, we aim to capture two important aspects: the audio quality and the adherence to the text description.

Frechét Audio Distance (FAD) The Frechét Audio Distance (Kilgour et al., 2019) is a reference-free audio quality metric which correlates well with human perception. The FAD is computed by comparing a reference set of audio samples to an evaluation set in terms of their distributions in an embedding space. A low FAD score indicates the audio of the evaluation set is plausible. We use the FAD toolkit [3] to evaluate our generated music samples in three embedding spaces: $\text{CLAP}_{\text{Audio}}$, $\text{CLAP}_{\text{Music}}$ and VGGish.

Kullback-Leiber Divergence (KL) The KL-divergence is a measure of how one probability distribution diverges from a second, expected probability distribution. Due to the complex nature of music, there is a many-to-many relationship between text descriptions and music clips. Therefore, we use a classifier (Koutini et al., 2022) trained for

[3] https://github.com/microsoft/fadtk

multi-label classification on AudioSet to compute the KL-divergence over the class probabilities between the reference set and generated music. The generated music is expected to share similar concepts with the reference set when the KL is low.

CLAP Score As a joint text-audio embedding model, CLAP can be used to quantify the similarity between text-audio pairs. We compute both the text embedding $f_{text}(\cdot)$ of the text caption, and the audio embedding $f_{audio}(\cdot)$ of the generated music sample. Similar to the MuLan Cycle Consistency (MCC) (Huang et al., 2022), the CLAP score is then defined as the average cosine similiarity between these embeddings (Copet et al., 2023; Huang et al., 2023b). A high CLAP score indicates the generated music adheres to the text prompt.

4 Results

Table 1 presents the results of the RAG approach against baseline generation methods. In general, FAD scores computed in the $\text{CLAP}_{\text{Audio}}$ and $\text{CLAP}_{\text{Music}}$ embedding spaces are lower than the VGGish embedding space. This could be attributed to VGGish being trained on a classification task at 16kHz, while CLAP is trained on a contrastive task at 48kHz. The higher dimensional CLAP features may also capture more complex musical features.

For the baseline generation methods, unconditional generation consistently generates music of poor audio quality. Specifying a single musical aspect or first sentence of the caption improves quality, while the full text caption achieves the best scores. This suggests that the conditioning text encoder plays a key role in influencing the generation process. The CLAP score is highest when specifying a single musical aspect, likely because matching the audio to a single word presents a simpler task than aligning with a more complex sentence.

For the RAG methods, retrieving similar items based on the caption outperforms retrieving similar items based on the musical aspects. This is reasonable as the musical aspects aim to capture more qualitative features and as a result could diverge more from the intended caption description. Retrieving five similar items based on the caption achieves the best FAD scores, suggesting a trade-off with prompt length. Despite their relevance, too many aspect qualities may hinder the models performance instead of focusing on a select few. The best CLAP score is achieved by retrieving three similar items based on the caption, which aligns with the notion that matching fewer relevant words presents a simpler task. Interestingly, the best KL score is achieved by retrieving ten similar items based on the musical aspect list. This could be due to our implementation of calculating KL, where we use a classifier trained for a multi-label task on AudioSet. By retrieving many diverse aspects, we increase the probability of matching with multiple labels.

Finally, retrieving random items generates music of worse or comparable audio quality to baseline generation methods. Again, this suggests a trade-off with the prompt length and relevance. This also demonstrates the ability of the pre-trained text encoder to transfer useful representations when generating diverse music.

5 Future Work

In this work, we explored the overall effect of an augmented prompt when generating music with MusicGen. However, it would be more valuable to investigate specifically how the augmented tokens affect the internal representations. SMITIN (Koo et al., 2024) trains classifier probes to identify self-attention heads that perform instrument recognition. Then, they introduce an inference time intervention technique for steering the generated output towards the desired musical trait. Extending this approach to other specific control methods for various musical features is a desirable goal.

People value agency and control over creative collaborations with generative AI models. As such, we want to build systems that promote interactive, human-centered approaches. Equipping MusicGen with the ability to refine and build upon previously generated output is another valuable direction.

6 Conclusion

In this paper, we presented a retrieval augmented approach for text-to-music generation. While relatively simple, we show our method consistently generates music of higher audio quality while displaying high text adherence. We compare the trade-offs of various retrieval strategies and suggest extensions to this work. Our findings show the potential for increased control in text-to-music generation.

7 Ethics Statement

Large scale generative models raises questions regarding ethics and societal consequences of their use. Generative text-to-music models can represent an unfair competition for artists which is an open problem. Another potential bias is the lack of diversity in the MusicCaps dataset, which contains a larger proportion of Western music. Through open research, we hope that such generative models can become useful as a tool for amateur musicians and professionals.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. Musiclm: Generating music from text. *Preprint*, arXiv:2301.11325.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2523–2533.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. 2023. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 47704–47720. Curran Associates, Inc.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. 2024. Adapting frechet audio distance for generative music evaluation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1331–1335.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. 2022. Mulan: A joint embedding of music audio and natural language. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, pages 559–566.

Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, Jesse Engel, Quoc V. Le, William Chan, Zhifeng Chen, and Wei Han. 2023a. Noise2music: Text-conditioned music generation with diffusion models. *Preprint*, arXiv:2302.03917.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023b. Make-an-audio: text-to-audio generation with prompt-enhanced diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Proc. Interspeech 2019*, pages 2350–2354.

Junghyun Koo, Gordon Wichern, Francois G. Germain, Sameer Khurana, and Jonathan Le Roux. 2024. Smitin: Self-monitored inference-time intervention for generative music transformers. *Preprint*, arXiv:2404.02252.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. 2022. Efficient training of audio transformers with patchout. In *Interspeech 2022*, pages 2753–2757.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2023. Audiogen: Textually guided audio generation. In *The Eleventh International Conference on Learning Representations*.

Max W. Y. Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, Jitong Chen, Wang Yuping, and Yuxuan Wang. 2023. Efficient neural music generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 17450–17463. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8293–8316, Mexico City, Mexico. Association for Computational Linguistics.

Zachary Novack, Julian Mcauley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. 2024. DITTO: Diffusion inference-time t-optimization for music generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38426–38447. PMLR.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2023. Encodecmae: Leveraging neural codecs for universal audio representation learning. *Preprint*, arXiv:2309.07391.

Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. 2024. Moûsai: Efficient text-to-music diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8050–8068, Bangkok, Thailand. Association for Computational Linguistics.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *CoRR*, abs/2301.02111.

Mingqiu Wang, Izhak Shafran, Hagen Soltau, Wei Han, Yuan Cao, Dian Yu, and Laurent El Shafey. 2024. Retrieval augmented end-to-end spoken dialog models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12056–12060.

Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan. 2024. Music controlnet: Multiple time-varying controls for music generation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:2692–2703.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. 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)*, pages 1–5.

Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. 2024. Retrieval-augmented text-to-audio generation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 581–585.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:495–507.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. 2024. Masked audio generative modeling. In *The Twelfth International Conference on Learning Representations*.