Interspeech 2024
1-5 September 2024, Kos, Greece

An End-to-End Approach for Chord-Conditioned Song Generation

Shuochen Gao[1,†], *Shun Lei*[1,†], *Fan Zhuo*[2], *Hangyu Liu*[2], *Feng Liu*[2], *Boshi Tang*[1], *Qiaochu Huang*[1], *Shiyin Kang*[2], *Zhiyong Wu*[1,3,*]

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] Kunlun Skywork Technology Co., Beijing, China
[3] Peng Cheng Lab, Shenzhen, China

{gsc22,leis21}@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn

Abstract

The Song Generation task aims to synthesize music composed of vocals and accompaniment from given lyrics. While the existing method, Jukebox, has explored this task, its constrained control over the generations often leads to deficiency in music performance. To mitigate the issue, we introduce an important concept from music composition, namely *chords*, to song generation networks. Chords form the foundation of accompaniment and provide vocal melody with associated harmony. Given the inaccuracy of automatic chord extractors, we devise a robust cross-attention mechanism augmented with dynamic weight sequence to integrate extracted chord information into song generations and reduce frame-level flaws, and propose a novel model termed Chord-Conditioned Song Generator (CSG) based on it. Experimental evidence demonstrates our proposed method outperforms other approaches in terms of musical performance and control precision of generated songs[1].

Index Terms: song generation, chord-conditioned, attention with dynamic weights sequence

1. Introduction

Music, as a ubiquitous art form, plays a significant role in people's lives. Music that includes both accompaniment and vocals is referred to as songs, where well-designed songs necessitate a harmonious blend of vocals and accompaniment. The task of Song Generation, which synthesizes songs from lyrics, can play a critical role in the entertainment industry.

Early endeavors in music generation leveraged symbolic representations to produce score parameters [1], which were then rendered into music. However, symbolic music is confined to fixed instrumental timbres and lacks expressiveness. Recent years have witnessed an emergence of End-to-End Music Generation models that generate musical audio through text prompts [2, 3, 4, 5] and melody control [6]. Yet, End-to-End Music Generation often struggles to produce meaningful vocals, frequently resulting in gibberish. Conversely, the Singing Voice Synthesis (SVS) field focuses on generating singing voices from lyrics and scores, with existing efforts [7, 8, 9, 10] capable of producing high-quality vocals. However, SVS-generated singing often lacks accompaniment and requires users to provide music scores, revealing a deficiency in the models' song composition capabilities.

To address the challenges faced by Music Generation and Singing Voice Synthesis, Jukebox [11] introduces lyrics as a control condition on top of text prompts, enabling autonomous song generation. However, it primarily models based on acoustic feature sequences, which impedes its ability to assimilate high-level music theory knowledge. Moreover, it exhibits limited control over the music generation process, often resulting in outputs that often lack musicality.

In this work, we present an innovative end-to-end Chord-Conditioned Song Generator (CSG), which introduces chord condition for generating condition-compliant songs. Chord, as an important concept of the song, forms the foundation of accompaniment and provides vocal melody with associated harmony [12]. Even simple chords can sustain the basic auditory sensation of accompaniment and, combined with vocals, create melodious songs. For instance, guitar playing and singing involve coupling guitar chords with matching vocals. Given the integral relationship chord shares with both accompaniment and vocals, it serves as a straightforward and effective control condition for generating both components. To our knowledge, this is the first instance where chords have been used as a control condition in the domain of End-to-End Music Generation. Previously, only a portion of Symbolic Music Generation efforts utilized chord control [13, 14, 15]. Chord control simplifies manipulation significantly over melody and score control. Through CSG, even non-expert users, lacking formal musical theory knowledge, can employ standard chord progressions like '6451', '4536', or custom sequences, facilitating the creation of unique, harmonious songs. Following [16, 2], CSG employs a Self-Supervised Learning (SSL) model to extract semantic tokens, serving as substitutes for acoustic features. Moreover, given that existing methods for automatic chord extraction suffer from low precision issues, merely incorporating chords does not enable the model to effectively learn the relationship between chords and music, thereby impacting the musicality of the generated songs. To address this issue, we propose an innovative Attention with Dynamic Weights Sequence (DWS) that, while integrating chords with lyrics and songs, also assesses the correctness of chords frame by frame. By reducing interference from erroneous data and increasing the model's confidence in accurate chord data, this approach simultaneously enhances the musicality and control precision of the generated songs.

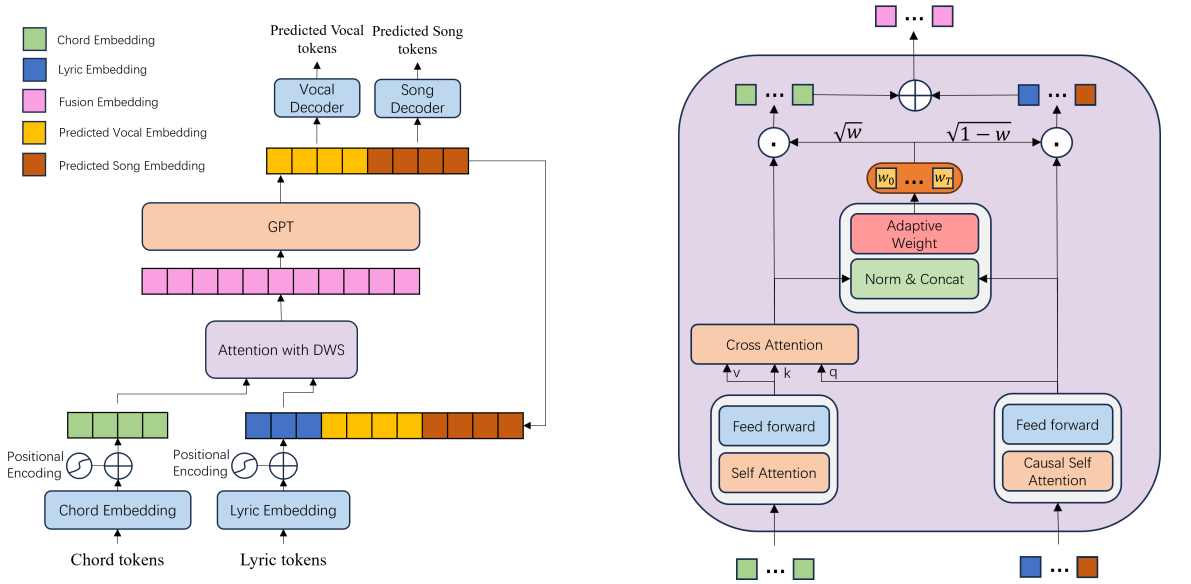
The main contributions of our paper are: (1) We introduce chords as the control condition for song generation, effortlessly and efficiently enhancing the musicality of the generated songs. (2) We propose an innovative Attention with DWS to improve the precision of control and the musical performance of the generated songs.

2. Methodology

As shown in Figure 1a, given chord and lyric tokens as inputs, CSG generates frames of songs in an autoregressive manner.

†Equal contribution.
*Corresponding author.
[1]Music sample: https://thuhcsi.github.io/interspeech2024-CSG

10.21437/Interspeech.2024-1837



(a) *Structure of the Chord-Conditioned Song Generator (CSG)* (b) *Details of Attention with Dynamic Weights Sequence (DWS)*

Figure 1: *The overall architecture of CSG with proposed Attention with DWS*

First, we map the input tokens to high-dimensional space. Then the chord and lyric embeddings are combined with vocal embeddings as well as song embeddings from previous frames, and fed into the Attention with DWS for semantics fusion. After that, the fusion embedding gets populated into a GPT module [17] to generate vocal and song embeddings for the current frame, which are finally decoded by two decoders for token prediction. At the training stage, the GPT and Attention with DWS get initialized from random weights and are trained together, while the lyric tokens are extracted by a tokenizer in pre-trained BERT[2] [18]. Vocal and song tokens are extracted by pre-trained BEST-RQ[3] [19, 20]. Chord token extraction will be explained in Section 2.1. During inference, we keep all the model weights fixed and get the chord&lyric tokens from users. At the end of inference, a diffusion vocoder, adjusted based on Stable Audio[4] [21], facilitates the restoration from tokens to audio. The following sections elaborate on the modules.

2.1. Chord Token Extraction

From the song data, we separate the background music and employ Autochord[5] to extract chord progressions from it. The extracted chord progressions consist of three components: chord roots, interval relations, and durations. Specifically, there are twelve possible chord roots: *C, C#, D, D#, E, F, F#, G, G#, A, A#, B* and four possible interval relations: *Major*, *Minor*, *Augmented*, and *Diminished*. The 48 possible combinations derived from them can cover the vast majority of chords. We further quantified the chord progressions into sequences, analyzing them at intervals corresponding to a 50Hz rate, with each frame quantized into one of 48 possible tokens. These tokens, determined by their root note and interval relation combination, are encoded as integers ranging from 0 to 47.

[2]BERT: https://github.com/google-research/bert/tree/master
[3]BEST-RQ: https://github.com/lucasnewman/best-rq-pytorch
[4]Stable Audio: https://github.com/Stability-AI/stable-audio-tools
[5]Autochord: https://github.com/cjbayron/autochord

2.2. Attention with Dynamic Weights Sequence (DWS)

To integrate chords with lyrics and songs, the simplest method is to concatenate the chord sequence with the audio sequence in the embedding dimension or to use cross-attention. However, concatenation allows the autoregressive prediction of the song to only see the preceding chord sequence, not the subsequent chords. Additionally, the measured chord extraction accuracy of the Autochord is 67.33%, which means our chord data inevitably contains some noise or even errors. These inaccuracies can interfere with the model's learning, whether through concatenation or cross-attention. To address this issue and improve the precision of control, we propose Attention with Dynamic Weights Sequence (DWS).

As illustrated in Figure 1b, Attention with DWS employs a dual-path architecture in which each path incorporates causal or non-causal transformer blocks to facilitate temporal alignment learning within sequences. Subsequently, a cross-attention mechanism is harnessed to synchronize chord embeddings with lyric-audio embeddings, resulting in the alignment output C:

$$C = softmax(\frac{Q_{lyric-audio}K_{chord}^{T}}{\sqrt{d_k}})V_{chord} \quad (1)$$

As a weighted average to chord embedding values, C is inevitably affected by the inaccuracies in chord data. This means that c_t, representing elements of C at specific time points, may undergo varying degrees of perturbation. To mitigate the impact of such inaccuracies, we introduce a temporally adaptive weighting network. This network is designed to assess the correlation between chord embeddings and audio embeddings sequentially, on a frame-by-frame basis, leading to the generation of the weight sequence W:

$$W = \sigma(M([C; A])) \quad (2)$$

Here, A denotes the lyric-audio embeddings computed by causal transformer blocks, and M is a mapping function that calculates temporal weights for each frame. σ is the sigmoid

Table 1: *Evaluation results of different song generation methods. The results of mean opinion scores (MOS) in user study are shown with 95% confidence intervals. The last row shows the ablation study. "w/o Attention with DWS (concatenation)" means using the concatenation method, and "w/o Attention with DWS (cross-attention)" means using cross-attention without DWS.*

Systems	**FAD**$_{vgg}$ ↓	**SIM**↑	**User Study**↑	
			MOS (MP)	MOS (CA)
Jukebox	14.06	-	2.46 ± 0.13	-
GPT-only	**6.80**	0.09	3.12 ± 0.09	2.65 ± 0.15
Ours	7.35	**0.61**	$\mathbf{3.74 \pm 0.09}$	$\mathbf{3.91 \pm 0.14}$
w/o Attention with DWS (concatenation)	7.67	0.52	3.17 ± 0.09	3.40 ± 0.14
w/o Attention with DWS (cross-attention)	7.71	0.46	3.24 ± 0.10	3.05 ± 0.15

activation function providing normalization. The generated weight sequence W then evaluates chord utilization across different moments. Consequently, the fusion embedding sequence F, which integrates both global chord information and preceding audio data, is represented as follows:

$$F = \sqrt{W} \circ C + \sqrt{1 - W} \circ A \tag{3}$$

Compared with concatenation and cross-attention, Attention with DWS enhances the robustness of the Fusion Embedding Sequence F towards chords containing inaccuracies while having access to global chord information. By employing a dynamic weight sequence to discern between correct and incorrect chords frame by frame, Attention with DWS ensures that the presence of incorrect chords during training does not diminish confidence in the correct chords. Thus, this enhances the precision of chord-conditioned control over the model. Furthermore, by minimizing the interference from incorrect chords, Attention with DWS enables the model to more effectively learn the musical correlation between chords and songs, thereby improving the musicality of chord-controlled song generation.

3. Experiments

3.1. Dataset

Given the lack of a large-scale, open-source lyrics-to-song dataset, we apply CSG to a proprietary dataset containing 554,467 English songs across various genres such as country, pop, rock, and rap. For preprocessing, the dataset is segmented into approximately 3 million vocal-only and accompaniment-only segments, each of which is annotated with the corresponding lyrics. We randomly sample 5% of the dataset and reserve it for validation and testing, while the rest is used for training.

3.2. Experiment Setup

We convert the chord, lyric, and audio tokens into 1024-dimensional embeddings. In Attention with DWS, both the self-attention block and the causal self-attention block consist of 2 layers, while the Adaptive Weight module is comprised of a single linear layer. The GPT model is constructed from 12 transformer blocks, each with a dropout rate of 0.1. Training of the proposed model and ablation baseline models are conducted for 500,000 steps on seven NVIDIA GeForce RTX4090 GPUs, with a batch size of 4 per GPU, utilizing the Adam optimizer and a learning rate warm-up scheduler with a target learning rate of 8×10^{-5} and a warm-up period of 32,000.

Jukebox is the sole work in the domain of song generation to date. We compare our model with the samples publicly available on the official Jukebox website. Moreover, to demonstrate how incorporating a basic chord condition can markedly enhance musicality, we trained a GPT-only model without the chord condition as a deterministic **baseline** for comparison, employing the same training methodology as our proposed model.

3.3. Results

3.3.1. Subjective Evaluation

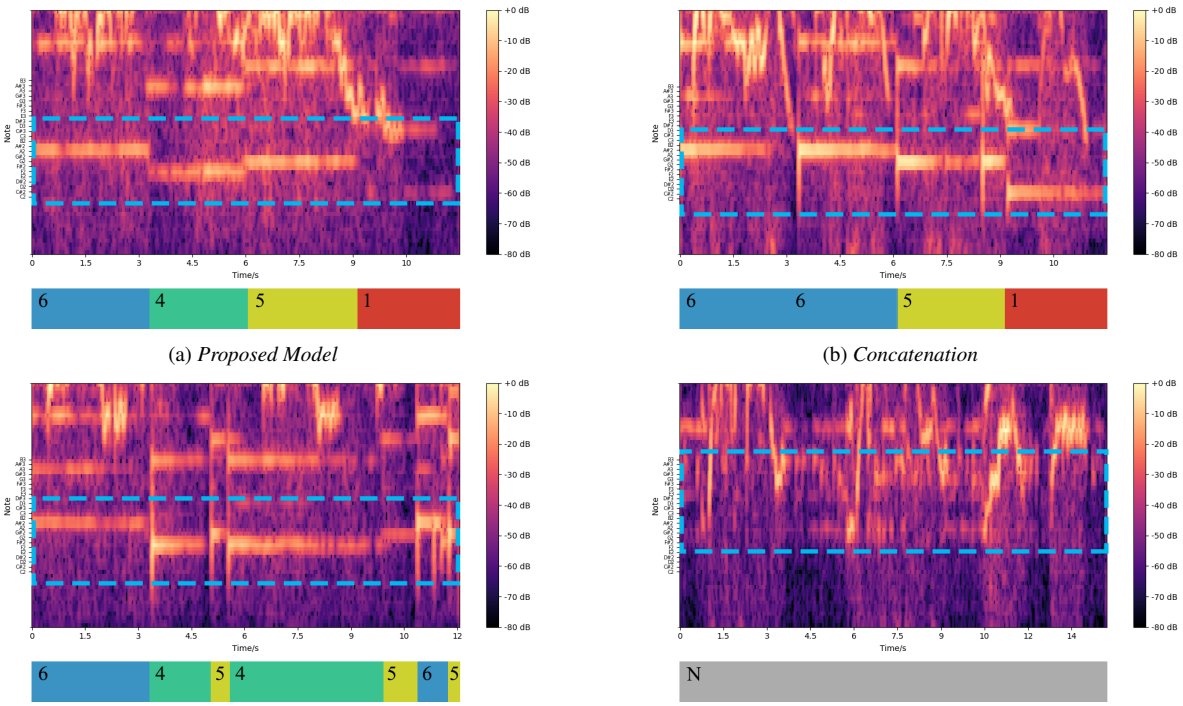
We employ two Mean Opinion Scores (MOS) to evaluate the capability of different models. 1) Musical Performance (MP): Evaluate the musicality of generated songs. 2) Chord Alignment (CA): Evaluate the correlation between the chords in the generated songs and the actual chords. For MP, we randomly select 15 sets of lyrics written by Jukebox researchers, which are not present in any existing dataset, to act as lyric inputs for each model. Additionally, we devise conventional chords, such as '6451', '4536', and '2516', as chord controls. Twenty-two participants are invited to rate these 15 song segments with identical lyrics, providing comprehensive scores for the songs' musicality. For CA, six sets of chords, including both conventional (e.g., '6451', '4536', '2516', '1564') and unconventional (e.g., '1111', '1234') chords, are used to generate 12 samples. Twenty participants evaluate the correlation between the chords in the generated songs and the actual chords, considering auditory perception.

The last two columns of Table 1 report the results of the MOS evaluation. Our proposed method achieves the highest MOS for both MP and CA. The MP of GPT-only music performance surpasses that of Jukebox, attributable to the modeling of semantic information by BEST-RQ, which improves the musicality of generated music. Moreover, the effective utilization of chord information is a key factor enabling our proposed model to outperform GPT-only in terms of musical performance. Owing to the GPT-only model's lack of chord conditioning, our proposed model significantly surpasses the GPT-only in terms of chord control precision.

3.3.2. Objective Evaluation

In terms of generation fidelity, we employ the Fréchet Audio Distance (FAD) [22] utilizing the VGGish model [23], where a lower FAD indicates higher audio fidelity. We randomly extract 103 10-second song segments with unseen lyrics from the Jukebox website, and generate 1000 12-second song segments using unseen lyrics and chords by both GPT-only and CSG, respectively, to calculate FAD.

Furthermore, control precision serves as a critical metric for assessing the effectiveness of control. We specify random chords and pair them with unseen lyrics to generate 400 song



(a) *Proposed Model* (b) *Concatenation*

(c) *Cross-attention* (d) *GPT-only*

Figure 2: *Synthesized spectrograms with note labels. Blue rectangular frames and the color bars below the spectrograms highlight the chords of the generated songs.*

segments. Utilizing Autochord, we extract the chords from these 400 segments and compare them with the specified chords to calculate the Similarity Index (SIM) for relative pitch accuracy, which calculates the proportion of the correct chords throughout the music. Meanwhile, we allow for a global shift in the key of the chords depending on the mode. Additionally, we compute the SIM for songs generated by the GPT-only model without chord conditions, serving as a reference baseline for uncontrolled generation.

The results for different methods are presented in Table 1. Both the GPT-only model and our proposed model exhibit lower FAD than Jukebox, attributable to the utilization of a vocoder based on Stable Audio, which enhances the fidelity of the generated audio. Besides, due to the introduction of chord control, our model exhibits a slight increase in FAD compared to GPT-only, while simultaneously demonstrating a significant improvement in SIM relative to conditions without chord control. The observed increase in FAD when incorporating control conditions is a typical phenomenon in music generation, and the improvement of SIM indicates that our model substantially enhances control over chords at the cost of a slight decrease in generation fidelity.

3.3.3. Ablation Study

We conduct ablation studies to further investigate the effectiveness of the attention mechanism with DWS. As shown in the last two rows of Table 1, simple integrated methods including concatenation and cross-attention fail to identify inaccurate chords during training. This failure leads to a disruption in recognizing correct chords during inference, resulting in diminished control precision, as indicated by SIM and CA. Additionally, this failure also hampers the model's capacity to accurately determine the relationships between chords and music, leading to a reduction in the musicality of the generated songs, as evidenced by MP. The inaccuracy also slightly decreases the generation fidelity, as shown by FAD.

3.3.4. Case Study

Beyond the quantitative metrics introduced earlier, the control precision of the proposed method can be directly demonstrated through qualitative analysis of spectrograms synthesized by various models under the same input conditions (Figure 2). When using chord condition as "6451", i.e., "A:min-F:maj-G:maj-C:maj", our proposed model generates a song with accurate chords, where the chord "1" concurrently resides in both notes "C2" and "C3". With the same chord and lyrics conditions, the concatenation model generates a song with chords "6651", and the cross-attention model generates a song with chords "6454565". Without using chords for control, controllable chord information is hard to be seen in the spectrogram.

4. Conclusion

In this work, we introduce a novel chord-conditioned song generation method, termed CSG, featuring our innovative Attention mechanism with DWS. This mechanism not only integrates chords with lyrics and songs but also reduces the impact of inaccurate chord data at the frame level. Experimental results demonstrate that CSG surpasses competing methods in musical performance through effective utilization of chord information, and Attention with DWS significantly enhances the musicality and control precision of generated songs.

5. Acknowledgements

This work is supported by National Natural Science Foundation of China (62076144), Shenzhen Science and Technology Program (WDZC20220816140515001, JCYJ20220818101014030) and the Major Key Project of PCL (PCL2022D01, PCL2023AS7-1).

6. References

[1] S. Ji, X. Yang, and J. Luo, "A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–39, 2023.

[2] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[3] K. Chen, Y. Wu, H. Liu, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies," *arXiv preprint arXiv:2308.01546*, 2023.

[4] Q. Huang, D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, Z. Zhang, J. Yu, C. Frank *et al.*, "Noise2music: Text-conditioned music generation with diffusion models," *arXiv preprint arXiv:2302.03917*, 2023.

[5] M. W. Lam, Q. Tian, T. Li, Z. Yin, S. Feng, M. Tu, Y. Ji, R. Xia, M. Ma, X. Song *et al.*, "Efficient neural music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[6] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[7] S. Zhou, S. Lei, W. You, D. Tuo, Y. You, Z. Wu, S. Kang, and H. Meng, "Towards Improving the Expressiveness of Singing Voice Synthesis with BERT Derived Semantic Information," in *Proc. Interspeech 2022*, 2022, pp. 4292–4296.

[8] Z. Zhang, Y. Zheng, X. Li, and L. Lu, "Wesinger 2: Fully parallel singing voice synthesis via multi-singer conditional adversarial training," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] H. Zhou, Y. Lin, Y. Shi, P. Sun, and M. Li, "Bisinger: Bilingual singing voice synthesis," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[10] Y. Lei, S. Yang, X. Wang, Q. Xie, J. Yao, L. Xie, and D. Su, "Unisyn: an end-to-end unified model for text-to-speech and singing voice synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 025–13 033.

[11] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[12] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, "Deep learning techniques for music generation–a survey," *arXiv preprint arXiv:1709.01620*, 2017.

[13] F. Li, "Chord-based music generation using long short-term memory neural networks in the context of artificial intelligence," *The Journal of Supercomputing*, pp. 1–25, 2023.

[14] S. Li and Y. Sung, "Melodydiffusion: chord-conditioned melody generation using a transformer-based diffusion model," *Mathematics*, vol. 11, no. 8, p. 1915, 2023.

[15] K. Choi, J. Park, W. Heo, S. Jeon, and J. Park, "Chord conditioned melody generation with transformer based decoders," *IEEE Access*, vol. 9, pp. 42 071–42 080, 2021.

[16] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "Audiolm: a language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[19] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, "Soundstorm: Efficient parallel audio generation," *arXiv preprint arXiv:2305.09636*, 2023.

[20] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *International Conference on Machine Learning*. PMLR, 2022, pp. 3915–3924.

[21] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," *arXiv preprint arXiv:2402.04825*, 2024.

[22] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms," in *Proc. Interspeech 2019*, 2019, pp. 2350–2354.

[23] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.