10.48047/jocaaa.2024.33.08.84

Vibrato: An Innovation in Music Education through Graphical Representation and AI-Powered Composition

Neetu Anand[1], Kunal[2], Auchitya[3], Raymond[4]
[1]Associate Professor, Maharaja Surajmal Institute
[2,3,4]Student, Maharaja Surajmal Institute

Abstract:

The paper explores the development of "Vibrato - A Music Learning and Generation Web App," showcasing innovations in music education through graphical representation and AI-powered composition. It examines the inefficiencies of current technology in music learning and generation, motivating the need for a novel approach. The app seamlessly integrates cutting-edge technology, offering a visually enriched learning experience and pioneering a sophisticated method of music generation. The methodology involves graphical note representation, real-time audio mapping, and a state-of-the-art AI-powered music generation module. Through meticulous codebook projection, transformer decoding, and logits prediction, Vibrato creates a solid foundation for high-quality, controllable music generation. The research includes comprehensive frontend and backend development, AI-powered music generation, user feedback gathering, and plans for future enhancements, ensuring continuous improvement and adaptability. The Vibrato project encompasses a diverse scope within the realm of music learning and generation, aspiring to deliver a holistic and innovative user experience. At its core, the project introduces a graphical representation of musical notes, providing a user-friendly platform tailored for individuals at varying proficiency levels. This graphical representation serves as a pivotal tool, enhancing the understanding of pitch and rhythm and making the learning process accessible and beneficial to both novice and seasoned musicians alike. All the major issues have been addressed in this paper.

1. Introduction:

The realm of music education and composition has witnessed transformative advancements propelled by technological innovation. As we navigate the intricacies of musical learning and creation, the limitations of conventional methodologies become increasingly apparent. This paper introduces "Vibrato - A Music Learning and Generation Web App," a pioneering initiative that endeavors to redefine music education through the integration of graphical representation and AI-powered composition.

10.48047/jocaaa.2024.33.08.84

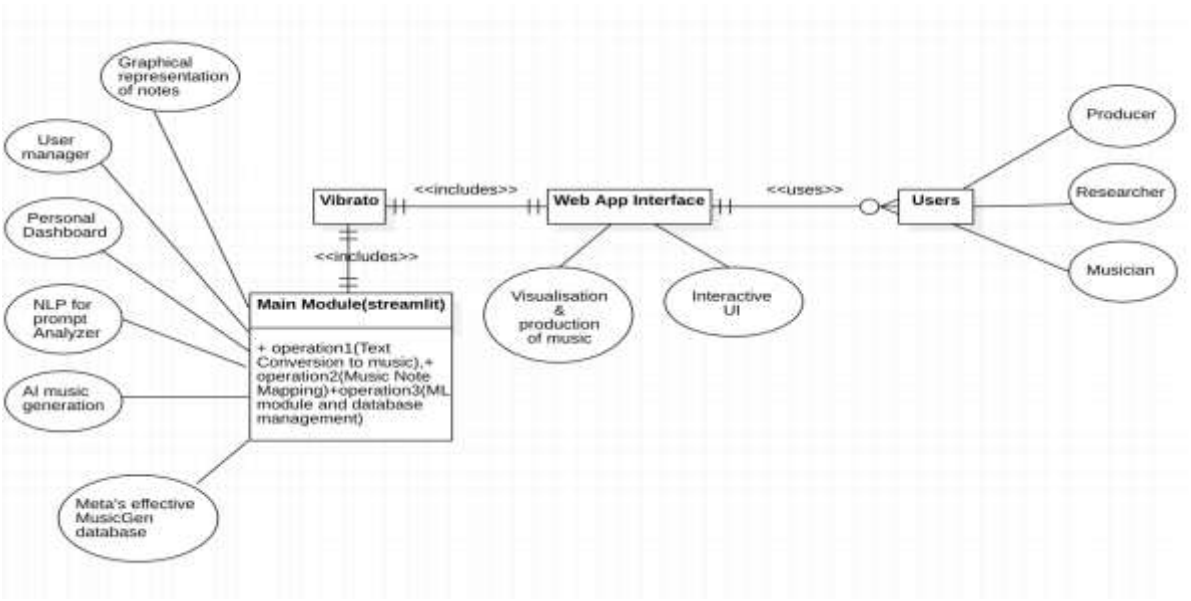


***Figure1*: Data flow diagram of Vibrato Application**

In the backdrop of contemporary technology, this web application not only addresses existing inefficiencies but sets a new standard for interactive and personalized music learning experiences.

The landscape of music education often grapples with challenges related to the comprehension of musical intricacies, especially for beginners. Traditional methods, reliant on notation and auditory exercises, may present barriers to entry for enthusiasts seeking a more intuitive and visually enriched learning path. Additionally, the process of music composition demands a nuanced understanding of various elements, and the limitations of current

technology can hinder the exploration of diverse musical landscapes.

The inefficiencies inherent in existing technology prompt the exploration of alternative approaches. "Vibrato" rises to this challenge, aiming to bridge the gap between traditional music education methods and the evolving needs of contemporary learners and creators. By leveraging graphical note representation and an AI-powered music generation module, this web app seeks to enhance the learning experience and unlock new possibilities for musical expression.

This introduction sets the stage for a comprehensive exploration of Vibrato's methodology, delving into its innovative features, including graphical note representation, real-time audio mapping, and the intricacies of the AI-powered music generation module. Through a meticulous examination of codebook projection, transformer decoding, and logits prediction, this paper unfolds the systematic approach
that positions Vibrato at the forefront of innovation in the intersection of technology
and music education. As we embark on this journey, the paper seeks to elucidate how Vibrato not only addresses the inefficiencies of current technology but propels music education into a new era of creativity and accessibility.

2. Literature Review:

2.1 Evolution of Music Education Technology:

The evolution of technology has profoundly impacted the field of music education, offering tools and platforms that cater to a diverse range of learners. Traditional methods, primarily reliant on sheet music and in-person instruction, have

10.48047/jocaaa.2024.33.08.84

gradually given way to digital solutions aimed at enhancing accessibility and interactivity. Notable advancements include the integration of multimedia elements and interactive interfaces to facilitate a more engaging learning experience.[1]

2.2 Challenges in Current Music Education Technology:

Despite the strides made in music education technology, challenges persist, particularly concerning the accessibility and comprehensibility of musical concepts. Conventional notation systems may pose barriers to beginners, hindering their ability to grasp complex musical elements such as pitch and rhythm. Additionally, existing technology may fall short in providing an immersive and visually intuitive platform for music composition, limiting the exploration of diverse musical landscapes.[2]

2.3 Graphical Representation in Music Learning:

Research has highlighted the benefits of graphical representation in aiding the understanding of musical concepts. Visualizing musical notes on a graph can significantly enhance the comprehension of pitch and rhythm, providing learners with a more intuitive pathway into the intricate world of music. The integration of graphical elements in music education platforms has shown promise in addressing the challenges associated with traditional notation systems.[3]

2.4 AI-Powered Music Generation:

The incorporation of artificial intelligence (AI) in music generation introduces a paradigm shift in creative expression. AI-driven models, particularly recurrent neural networks (RNNs), have demonstrated the ability to transform textual prompts into unique and expressive musical compositions. This technology opens avenues for users to explore diverse musical landscapes, pushing the boundaries of traditional composition methods.[4]

2.5 Tokenization Techniques in Music Data Representation:

Efficient data representation is crucial for the manipulation of musical information. Tokenization techniques, such as the use of compressed discrete music representation through EnCodec audio tokenizer with Residual Vector Quantization (RVQ), provide a concise and effective means of conveying musical elements. These techniques not only enable efficient data compression but also allow for the extraction of essential musical features, fostering a structured and organized exploration of composition.[5]

2.6 Real-Time Audio Mapping in Music Education Platforms:

Real-time audio mapping represents a significant development in music education technology, particularly in applications focused on singing and performance. Platforms utilizing frameworks like Streamlit offer simplicity and rapid prototyping capabilities, enabling immediate feedback on singing performances. This feature enhances the interactive nature of the learning experience, providing users with instant insights into their musical execution.[6]

2.7 Current State of AI-Powered Music Education Platforms:

While AI-powered music education platforms exist, a comprehensive integration of graphical representation, real-time audio mapping, and sophisticated AI-powered music generation remains relatively scarce. This literature review identifies the need for a holistic approach that combines these elements to address the

10.48047/jocaaa.2024.33.08.84

inefficiencies of current technology and provide a seamless and innovative learning experience.[7]

2.8 Summary:

In summary, the literature review underscores the evolving landscape of music education technology, emphasizing the persistent challenges and limitations of current approaches. The integration of graphical representation and AI-powered music generation stands out as a promising avenue for overcoming these challenges, offering a potential solution to enhance the accessibility, comprehension, and creativity within the realm of music education. The subsequent sections of this paper will delve into the methodology adopted by "Vibrato" to address these issues and pave the way for a transformative music learning and generation experience.

Graphical Representation of Musical Notes:

The methodology employed in "Vibrato" begins with a revolutionary approach to data representation. Musical notes are transformed into visually intuitive graphical formats, bridging the gap between traditional notation systems and a more accessible learning platform. The graphical representation enhances the comprehension of pitch and rhythm for learners, catering not only to beginners but also providing a valuable tool for musicians at various skill levels.[8]

AI-Powered Music Generation Module:

At the core of "Vibrato" lies an AI-powered music generation module, driven by a recurrent neural network (RNN). Trained on a vast dataset curated by Meta, this neural network possesses the capability to transform text-based prompts into unique and expressive musical compositions. Leveraging the power of artificial intelligence, this system introduces a new dimension to music creation, enabling users to explore diverse musical landscapes effortlessly.[9]

3. Methodology:

3.1 Data Representation:

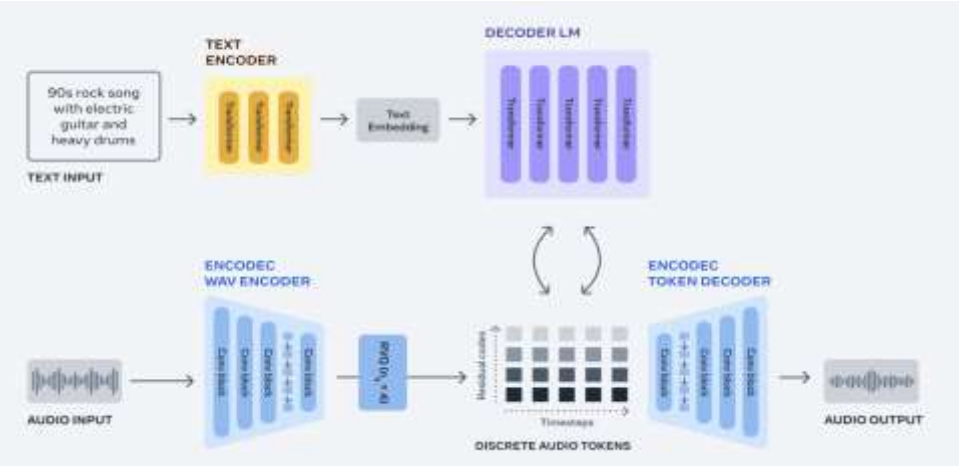


Figure2: **Model of Text-to-Music of Meta's MusicGen [20]**

10.48047/jocaaa.2024.33.08.84

Compressed Discrete Music Representation:
To facilitate the manipulation of musical information, the project adopts a compressed discrete music representation using tokens. Acquired through the EnCodec audio tokenizer, these tokens offer a concise and efficient means of conveying musical elements. The tokenizer employs Residual Vector Quantization (RVQ), a state-of-the-art technique that facilitates the compression of musical data. The outcome is a set of multiple parallel streams, each comprised of discrete tokens drawn from different learned codebooks.[10]

3.2 Codebook Projection and Positional Embedding:
Codebook Patterns:
The primary objective is to define specific patterns (Ps) for codebook projection at
each step. Unique patterns are developed by carefully considering the codebook indices relevant to each step in the generation process. These patterns serve as a guide for the model to understand which codebooks to focus on during different stages of music generation.

Codebook Retrieval:
Efficient retrieval of codebooks from the input tensor Q based on predefined patterns (Ps) is implemented. Learned embeddings enhance the representation of codebook values, and a special token is introduced to handle scenarios where certain codebooks might be absent.[11]

Sinusoidal Embedding:
Inclusion of sinusoidal embeddings is essential to encode the current step (s) and provide temporal context to the generated music. Integration of sinusoidal embeddings ensures that the model considers the temporal aspect of music generation, enhancing the understanding of the sequential order of events.

3.3 Transformer Decoder:
Transformer Architecture:
A key objective is to implement a transformer with L layers and dimension D. The design revolves around capturing sequential dependencies in the music generation process. The inclusion of multiple layers and a specified dimension ensures that the model can effectively learn and represent intricate patterns in the input data.[12]

Cross-Attention Block:
Integration of a cross-attention block fed with the conditioning signal C enhances the model's ability to generate music based on external criteria. This step enables the model to consider external conditioning signals, such as melodic features or textual prompts, during the music generation process. The cross-attention block facilitates the incorporation of additional context, making the generated music more versatile and controllable.

Fully Connected Block:
The conclusion of each layer with a fully connected block and a residual skip connection aims to process information within each layer efficiently. Fully connected blocks are implemented to process information at a local level within each layer. The addition of residual skip connections aids in efficient learning, allowing the model to retain essential information across layers and enhance convergence.[13]

Layer Normalization:
The application of layer normalization to each block before summing with the residual skip connection contributes to stabilizing the training process. Layer normalization is introduced to each block to mitigate internal covariate shifts and stabilize the training process, ensuring more consistent and reliable learning throughout the model.[14]

10.48047/jocaaa.2024.33.08.84

3.4 Logits Prediction:

Pattern Transformation:

Transformation of the output into logits predictions for Q at indices given by Ps+1 is crucial for generating the next set of codebooks. This step involves developing a mechanism to transform the output of the model into predictions specific to the next pattern step Ps+1, ensuring coherence in the overall composition.

Codebook-Specific Linear Layer:

The application of a codebook-specific linear layer from D channels for each present codebook in Ps+1 ensures tailored predictions for each codebook. Codebook-specific linear layers are implemented to customize predictions for each codebook in the next pattern step Ps+1, allowing the model to generate diverse and expressive musical compositions based on the specified codebook patterns.[15]

3.5 Frontend Development with React:

User Interface Design:

Development of a user-friendly interface for both learning and generation modules is achieved using React, a JavaScript library for building user interfaces. Key features for the learning and generation modules, emphasizing graphical note representation, are identified. Collaboration with UX/UI designers results in visually appealing and intuitive interface components.

Graphical Note Representation: Visualization of musical notes on a graph enhances learning. A graphical representation of notes is integrated using SVG or canvas elements. Musical notes are mapped to the graph, allowing users to visually understand pitch and rhythm. Interactive features are implemented for users to engage with the graphical representation.

3.6 Backend Implementation with Streamlit:

Real-time Audio Mapping:

To handle real-time audio input and facilitate immediate feedback during singing, Streamlit is chosen as the backend framework for its simplicity and suitability for rapid prototyping. Real-time audio mapping is implemented by integrating audio input processing libraries, enabling users to receive instant feedback on their singing performance through the graphical interface.

3.7 AI-powered Music Generation:

Recurrent Neural Network (RNN):

Implementation of an AI-powered music generation module is achieved using an RNN architecture. The RNN is chosen for its ability to capture sequential dependencies in music. Training on Meta's dataset, which likely includes diverse musical patterns, the RNN is integrated to transform text-based prompts into unique musical compositions.[16]

3.8 User Feedback and Evaluation:

Gathering User Feedback:

Objective: Collect feedback to evaluate the effectiveness of Vibrato.

Approach: Deploy the web application to a user testing environment. Solicit user feedback through surveys, interviews, or user testing sessions. Focus on understanding the user experience with graphical note representation and real-time audio mapping.

AI-powered Music Generation Evaluation:

Objective: Evaluate the performance of the AI-powered music generation module.

Approach: Define objective metrics for evaluating the generated musical pieces. Conduct human studies to gather subjective assessments of the quality and uniqueness of the generated music. Use the gathered data to refine the AI model and improve its capabilities.[17]

3.9 Future Enhancements:

Expansion of Music Generation Capabilities:

10.48047/jocaaa.2024.33.08.84

Objective: Plan for future enhancements to expand music generation capabilities.
Approach: Identify areas for improvement, such as generating more diverse and complex musical compositions. Explore advanced AI techniques or larger and more diverse datasets for training the music generation model.[18]

Incorporation of Advanced Audio Analysis Tools:
Objective: Plan for future integration of advanced audio analysis tools.
Approach: Investigate advanced audio analysis tools, such as pitch detection or harmony analysis. Evaluate how these tools can enhance the learning and generation experience in Vibrato. Plan for the seamless integration of these tools into the existing framework.

3.10 Conclusion:
The outlined methodology adopts a meticulous approach to codebook projection, transformer decoding, and logits prediction. Through the definition of codebook patterns, integration of sinusoidal embeddings, and incorporation of attention mechanisms, "Vibrato" can effectively generate music conditioned on various factors.

4. Testing Methods and User Survey Results

4.1 Testing Methods:

The Vibrato project underwent a rigorous testing process to ensure the reliability and functionality of its key features. The testing methods employed included both automated testing for backend processes and manual testing for user interactions. The automated testing focused on codebase verification, ensuring that backend functionalities and algorithms operated correctly. Manual testing covered a wide array of user scenarios to evaluate real-world usage, including scale and pitch training, instrument accompaniment, real-time feedback, progress tracking, customized practice plans, community and collaboration features, and the innovative Text-to-Music Generation.[19]

4.2 User Survey Results:

To gather comprehensive feedback from users, a survey was conducted among 50 users from our campus. Participants were asked to rate various features on a scale from 1 to 10, with 1 being the lowest and 10 being the highest. The results provided valuable insights into user satisfaction and areas that may require further attention.

1. Scale and Pitch Training:
   - Rating: Average
   - Feedback: Scale and pitch training features received an average rating, suggesting that users perceived them to be functional but possibly requiring enhancements for a more satisfying user experience.

2. Instrument Accompaniment:
   - Rating: Highest
   - Feedback: Instrument accompaniment received the highest rating among users. Users appreciated the seamless integration with instruments like piano, guitar, and drums, indicating a positive and satisfying experience.

3. Real-Time Feedback:
   - Rating: Highest
   - Feedback: Real-time feedback received the highest rating among users. This suggests that users found the real-time feedback feature to be effective and beneficial in their practice sessions.

4. Progress Tracking:
   - Rating: Average
   - Feedback: Progress tracking features received an average rating, indicating

10.48047/jocaaa.2024.33.08.84

that users found them functional but with potential room for improvement.

5. Customized Practice Plans:
- Rating: Average
- Feedback: Customized practice plans received an average rating, suggesting that users found the feature functional but possibly requiring additional enhancements for a more personalized experience.

6. Community and Collaboration:
- Rating: Low
- Feedback: Community and collaboration features received a lower rating, indicating areas that may need attention to improve user engagement and satisfaction.

7. Text-to-Music Generation:
- Rating: Below Average
- Feedback: The Text-to-Music Generation feature received a below-average rating, highlighting potential concerns or challenges that users encountered with this specific feature.

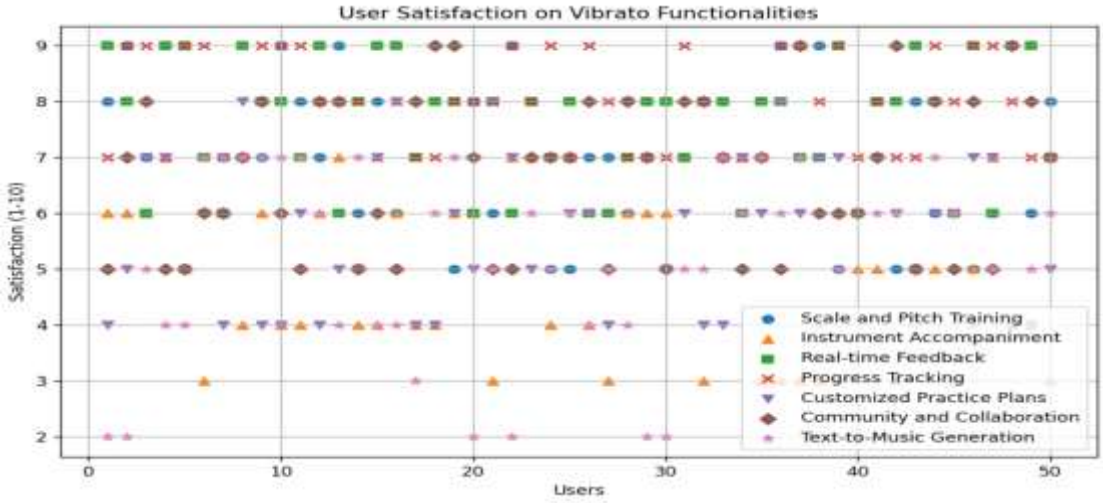


Figure3: Graphical Representation of testing

4.3 Conclusion:

The user survey results provide valuable insights into the strengths and areas of improvement for the Vibrato project. While features like instrument accompaniment and real-time feedback were well-received, the Text-to-Music Generation feature and community collaboration features require attention based on the user ratings. These findings emphasize the importance of addressing user feedback to enhance user satisfaction and overall usability. The survey results will be instrumental in guiding future updates and optimizations for the Vibrato project, ensuring its continued success among users.

5. Scope

The Vibrato project encompasses a diverse scope within the realm of music learning and generation, aspiring to deliver a holistic and innovative user experience. At its core, the project introduces a graphical representation of musical notes, providing a user-friendly platform tailored for individuals at varying proficiency levels. This graphical representation serves as a pivotal tool, enhancing the understanding of pitch

10.48047/jocaaa.2024.33.08.84

and rhythm and making the learning process accessible and beneficial to both novice and seasoned musicians alike.

In tandem with graphical representation, the project integrates a sophisticated AI-powered music generation module. Fueled by a recurrent neural network, this module empowers users to explore a spectrum of musical possibilities by transforming text-based prompts into unique and expressive compositions. The incorporation of artificial intelligence not only introduces a novel dimension to music creation but also encourages users to delve into personalized and creative musical endeavors.

Efficient data representation is a key aspect of the project's scope, achieved through advanced tokenization techniques such as the EnCodec audio tokenizer utilizing Residual Vector Quantization (RVQ). This ensures a streamlined approach to conveying musical elements, offering users a structured and organized method for manipulating musical information within the application.

Real-time audio mapping is seamlessly integrated to provide users with immediate feedback during singing practice sessions, contributing to an enriched learning experience. The development of a comprehensive user interface, facilitated by React for frontend development, underscores the project's commitment to ensuring a seamless and visually engaging platform. The graphical note representation, supported by React, aids users in comprehending pitch and rhythm intuitively, enhancing the overall learning process.

The project's scope extends beyond individual learning to foster community and collaboration. Community features, including forums, collaborative projects, and the sharing of user recordings, aim to create a vibrant musical community within the platform. This collaborative environment encourages users to connect, share experiences, and embark on musical journeys together.

An additional innovative feature, the Text-to-Music Generation, enables users to convert textual inputs into musical compositions, providing a unique avenue for creative expression and experimentation. In summary, the Vibrato project's scope encompasses graphical note representation, AI-powered music generation, efficient data representation, real-time audio mapping, user-friendly interface development, community engagement features, and a pioneering Text-to-Music Generation feature. The project aspires to be a dynamic and inclusive platform, facilitating music learning and empowering users to actively engage in the creation and sharing of musical content within a vibrant community.

6. Conclusion

In conclusion, Vibrato stands as a pioneering force in music education and creation, seamlessly integrating graphical note representation, AI-powered composition, and advanced tokenization. While the model exhibits areas for improvement, user feedback remains instrumental for refinement. The project's commitment to fostering a vibrant musical community through forums and collaborative features is commendable. Vibrato's innovative Text-to-Music Generation adds a unique creative dimension. Continuous iteration, addressing identified inefficiencies, and incorporating user insights will propel Vibrato to further excellence. As a dynamic platform, it aspires to redefine music learning, encourage creativity, and build a collaborative space for musicians of all levels.

10.48047/jocaaa.2024.33.08.84

References

1. Jade Copet, Felix Kreuk (7, November 2023). Simple and Controllable Music Generation. arXiv:2306.05284v2 [cs.SD]

2. Meinard Müller. Fundamentals of music processing: Audio, analysis, algorithms, applications, volume 5 Springer, 2015.

3. Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. arXiv preprint arXiv:2210.13438, 2022.

4. Shahaf Bassan, Yossi Adi, and Jeffrey S Rosenschein. Unsupervised symbolic music segmentation using ensemble temporal prediction errors. arXiv preprint arXiv:2207.00760, 2022.

5. Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561, 2021.

6. Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. arXiv preprint arXiv:2209.15352, 2022.

7. Chris Donahue, Antoine Caillon, Adam Roberts, Ethan Manilow, Philippe Esling, Andrea Agostinelli, Mauro Verzetti, Ian Simon, Olivier Pietquin, Neil Zeghidour, et al. Singsong.

8. Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. arXiv preprint arXiv:2301.02111, 2023.

9. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1):5485–5551, 2020.

10. Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416, 2022.

11. Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. arXiv preprint arXiv:2301.12503, 2023.

12. Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. arXiv preprint arXiv:2301.12661, 2023a.

10.48047/jocaaa.2024.33.08.84

13. Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023.

14. Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2023.

15. Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

16. Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In Advances in Neural Information Processing Systems, 2022.

17. Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by d-adaptation. arXiv preprint arXiv:2301.07733, 2023.

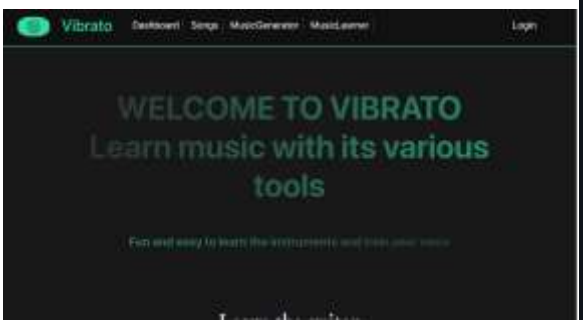
18. Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. arXiv preprint arXiv:2110.05069, 2021.

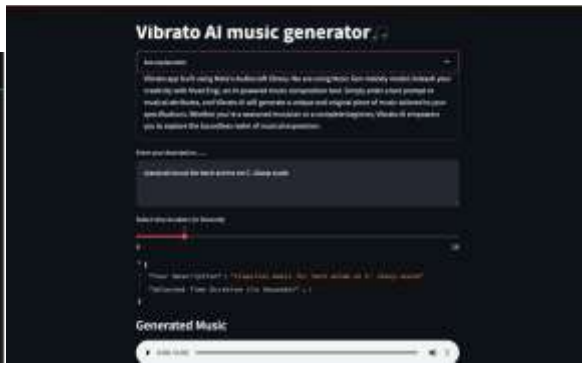
19. Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

20. Adrien Ycart, Emmanouil Benetos, et al. A study on lstm networks for polyphonic music sequence modelling. ISMIR, 2017.

Look and Feel of the Project :

10.48047/jocaaa.2024.33.08.84





How Graph is plotted of voice in real time: