Towards AI Rapper: Creating an Interactive Rap Battle Experience with Generative AI

Nikita Kozodoi
Amazon Web Services

Elizaveta Zinovyeva
Amazon Web Services

Zainab Afolabi
Amazon Web Services

Egor Krasheninnikov
Amazon Web Services

Abstract

This demo lets you experience a real-time interactive rap battle between a human and AI or two AI opponents. We develop an end-to-end solution that leverages multiple data modalities and ML models. Our system integrates speech recognition, emotion detection, language modeling, vision, text-to-speech, and voice cloning. The system combines multiple models to process user audio input, generates appropriate and contextually-aware text and synthesizes audio responses aligned with the background beat music. The demo demonstrates the potential of AI systems in creative music applications and offers insights into challenges and opportunities associated with integrating AI systems for interactive entertainment.

1 Background

Freestyle rap battles are a competitive form of artistic expression well-established in hip-hop culture [3]. The battles typically involve two or more rap performers who improvise rhyming verses, aiming to outperform opponents through clever wordplay and rhythmic delivery. The battles serve as a platform for rappers to showcase their linguistic dexterity, quick thinking, and cultural knowledge.

As generative AI continues to advance in areas such as language modeling and speech synthesis [2], there is growing interest in developing AI systems capable of engaging in creative tasks traditionally performed by humans. This demo explores this emerging field and presents an end-to-end solution for creating an interactive rap battle experience between a human and an AI opponent.

Creating a coherent rap battle experience presents several challenges, including generating contextually appropriate lyrics, synthesizing speech with proper prosody and emotion, integrating generated vocals with background music, and achieving sufficiently low latency for real-time interaction. Our work builds upon recent advances in LLMs, speech synthesis, and music generation, while drawing inspiration from prior research on ML-driven rap generation [4; 5] and musical experiences [1].

2 Demo

The demo implements a web-based application deployed on AWS. The solution runs the rap battle on a device with Internet access, using built-in microphone and speakers. The application orchestrates multiple AI models running on a GPU-powered server that work together to record and transcribe user lyrics, generate AI responses, and enable real-time voice-to-voice rap exchanges with low latency.

The video demonstrating the demo flow is available at `https://youtu.be/M7l5RjHEons`.

Figure 1 provides a high-level overview of the end-to-end interactive rap battle demo. The solution takes the following user inputs across different modalities:

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

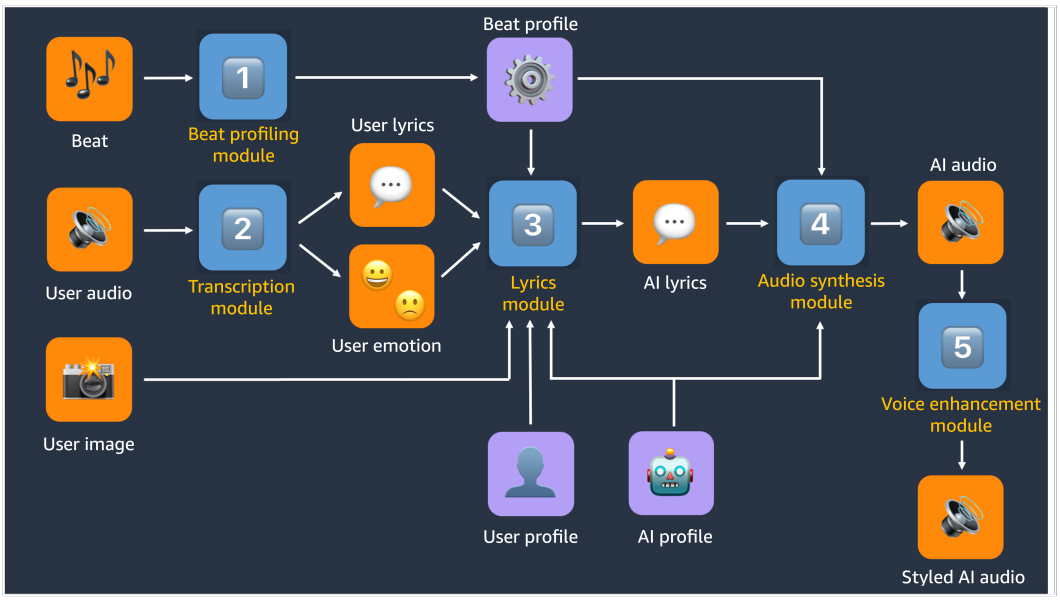

Figure 1: AI rap battle system architecture: processing multi-modal user inputs, generating contextually-aware LLM responses, and synthesizing beat-aligned vocal outputs.

1. **Beat audio:** Consistent with a typical rap battle setting, a constant beat plays in the background during both AI and human turns. We process the beat using audio processing tools and extract a beat profile (i.e., downbeat times, BPM, beat size).
2. **User audio:** The user's vocal input is recorded and processed to extract both the lyrical content and emotional tone of the audio. We use Whisper Large for transcription and Emotion2Vec for emotion classification.
3. **User image:** User photo is optionally captured to provide visual context to the AI rapper.
4. **User and AI profiles:** The user may optionally provide context data on rapper personas (e.g., name and short bio). If present, this data is passed as additional context.

After processing the user data, we generate AI rap audio in these steps:

1. **Lyrics generation:** Given user lyrics, photo, and additional context, we use an LLM on Amazon Bedrock to generate creative and context-aware rap lyrics. The beat profile data defines the number of lines, their length and rhyme structure.
2. **Content check:** Amazon Bedrock Guardrails check the LLM output and user inputs for prompt attacks and content filters. In case of violations, we regenerate the AI lyrics until the content check is passed before moving to the next step.
3. **Audio synthesis:** The AI kyrics are converted into expressive audio line by line using Kokoro TTS or Chatterbox TTS model. The audio is enhanced by inserting special sounds (e.g., laughter) between the lines. The beat profile is used to align each line with the corresponding downbeat timestamp and adjust the line speed if required.
4. **Voice cloning:** As a post-processing step, we use OpenVoice to further enhance the audio's expressiveness and allow for the imitation of typical rap vocals.

Our demo synthesizes and extends various research threads, presenting a novel integration of multimodal AI technologies to create an interactive rap battle experience. By combining real-time processing of user input across text, audio, and vision modalities with expressive audio synthesis and music alignment, we aim to push the boundaries of AI for artistic and entertainment applications.

References

[1] Briot, J.P., Hadjeres, G., Pachet, F.D.: Deep learning techniques for music generation, vol. 1. Springer (2020)

[2] Łajszczak, M., Cámbara, G., Li, Y., Beyhan, F., van Korlaar, A., Yang, F., Joly, A., Martín-Cortinas, Á., Abbas, A., Michalski, A., et al.: Base TTS: Lessons from building a billion-parameter text-to-speech model on 100K hours of data. arXiv preprint arXiv:2402.08093 (2024)

[3] Oware, M.: Battle rap: An exploration of competitive rhyming in hip hop. In: African Battle Traditions of Insult: Verbal Arts, Song-Poetry, and Performance, pp. 147–164. Springer (2023)

[4] Savery, R., Zahray, L., Weinberg, G.: Shimon the rapper: A real-time system for human-robot interactive rap battles. arXiv preprint arXiv:2009.09234 (2020)

[5] Wu, D., Addanki, K.: Learning to rap battle with bilingual recursive neural networks. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)