Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 514–521
https://doi.org/10.54941/ahfe1006863

Dramatizing Everyday Conversations: A Context-Aware BGM Recommendation System Using Generative AI

Maki Sakamoto, Shota Takahashi, and Haruka Matsukura

The University of Electro-Communications, Chofu, Tokyo 182-8585, Japan

ABSTRACT

Conventional music recommendation systems often rely on predefined emotional values or direct user interaction, making it difficult to incorporate nuanced conversational context. To address this limitation, we propose a novel system that recommends background music (BGM) for everyday conversations based on contextual analysis using a generative AI model, Gemini. Our system transcribes spoken dialogues into text, analyzes the content using Gemini, and then identifies similar scenes and BGMs from a preconstructed dataset composed of 12 BGMs derived from the Japanese TV drama "Ichiban Sukina Hana (My Favorite Flower)". By matching real-life conversations with relatable dramatic contexts, the system aims to enhance the immersion and emotional resonance of ordinary dialogues.

Keywords: Context-aware recommendation, Generative AI, Conversational BGM, Emotional computing

INTRODUCTION

In everyday conversation, having background music (BGM) can often make the interaction more lively by, for example, promoting relaxation. In such settings, users frequently search for and play tracks that match the atmosphere of the moment. With the recent spread of diverse music-streaming services, an enormous number of tracks have become available. Platform examples include Apple Music, Spotify, YouTube Music, and LINE MUSIC. Typically, users install these platforms on their smartphones and rely on keyword search to find tracks to add to their playlists. When they know a song title or artist name, discovery is straightforward. Conversely, when one wishes to find music that "fits the mood" without concrete information, searching within a massive catalog becomes difficult. Accordingly, systems have been studied in which users input impressions using affective words or natural language, but representing music with affective terms or natural language is limited and may yield results misaligned with user intent. For these reasons, we consider there to be a need for systems that estimate conversational atmosphere and recommend matching BGM.

In this study, we aim to enhance conversational experiences by recommending BGM that stages everyday dialogue as if it were a drama. Rather than having users actively search for BGM, the system proposes

© 2025. Published by AHFE Open Access. All rights reserved.

appropriate tracks from users' utterances and behaviors, and plays BGM suited to the ongoing conversation—thereby potentially improving the conversational experience and elevating the ambience as well.

We aim to build a system that omits search operations otherwise required of users and automatically recommends tracks suitable for their conversation. Concretely, from the input conversation we use today's generative AI, "Gemini," to search for drama scene texts with high similarity to the conversation, and realize a mechanism that plays the BGM used in the corresponding scenes. By using this system to play BGM that matches the conversation content, we expect to create the feeling of being inside a drama. Because BGM in dramas and films is known to strongly appeal to human emotions, automatically recommending appropriate BGM can deepen immersion in the conversation and provide a richer communication experience.

Given a search input, the system recommends tracks to the user. In our experiment, because we seek to optimize results from intuitive, impressionistic inputs, we first review current recommendation systems, focusing on YouTube.

Google has published papers on YouTube's video recommendation algorithms (Covington, 2016; Beutel, 2018; Chen, 2019). Covington (2016) studies how to apply deep learning to a large-scale platform like YouTube. Beutel (2018) and Chen (2019) investigate mechanisms that improve app usage time and click-through rates using user behavior data. On services as large and diverse as YouTube, collaborative filtering is often employed; content-based filtering is considered difficult to apply due to cost and the sheer diversity of content types. Biases in YouTube recommendations have been noted in Kirdemir (2021) and Liu (2021). Kirdemir (2021) points out that a small subset of videos strongly influences recommendation generation. Liu (2021) reports a Matthew Effect (popular videos are more likely to be recommended). Because training optimizes for rewards such as watch time using collaborative-filtering-like methods, such drawbacks readily emerge. Consequently, YouTube's recommendation system reflects the influence of majority-user behavior. While content popular with the majority tends to be high-quality and in demand, this can also skew exposure.

Given that music fundamentally acts on human sensibility, it is crucial not only to search using objective, factual information such as acoustic features, but also to prioritize affective information—how a song feels. To address the limitations of attribute-based MIR, research on content-based approaches that rely on the music content (e.g., acoustic feature vectors) has been active in recent years (Nakamura, 2011).

Approaches that automatically retrieve tracks aligned with a user's musical preferences (like/dislike) have long been proposed. Yoshii (2006) proposed a method that uses user ratings on a subset of registered tracks to deliver preference-aligned music. They integrate collaborative filtering—leveraging ratings from users with similar evaluations—and content-based recommendations that consider acoustic similarity, using a Bayesian network to represent user preference as a latent variable. Kaji (2004) proposed combining user preference with situational context. They focus on lyrics as features, recommending tracks whose lyrics resemble those of songs the user

likes; they also prepare labels for listening context (time of day, location, psychological state) and recommend songs similar both in preference and context, with labels obtained via web-based surveys.

However, preference-based search has four major issues. First, when user preference is the query, tracks outside those preferences are not retrieved, limiting exploration beyond one's taste. Second, when affective words or rating scales on affective words are used as the query, the intended information represents the user's mental state when listening; expressing such vague states precisely is difficult. Third, affective-term-based search requires users to possess a sizable affective vocabulary; repeated use of the same terms hampers discovery of new music. Fourth, adjective-pair rating scales demand many dimensions to capture subtle nuance, increasing user burden and undermining usability. Moreover, because such systems require prior learning of user traits, they are not easy to use casually. In sum, while emphasizing affective information is important, rather than relying on predefined affective vocabularies or scales, we should incorporate new, more intuitive affective signals into music search.

We therefore propose a system targeting everyday conversation that recommends music based on conversational context and flow. The system obtains text from the input conversation and analyzes it with a generative AI. Concretely, it leverages Gemini to measure similarity of topics and content at high precision. Beyond conventional emotion analysis, we utilize Gemini's advanced language-processing capabilities to perform detailed contextual analysis and, based on that, select optimal drama scenes and their corresponding BGM. We prepare a variety of drama scenes reflecting everyday situations and preanalyze them to allow flexible matching across diverse conversational content. A key feature of this approach is that drama scenes can comprehensively cover varied situations—relaxed chats among close friends, formal workplace exchanges, or emotionally charged moments—providing scripts whose style and content are close to natural conversation. Because drama BGM can amplify emotions, it can emphasize the ambience and foster deeper immersion in the moment. This, in turn, can naturally highlight conversational flow and enhance mood.

Our goal is to realize an interactive experience that goes beyond mere music recommendation. By understanding conversational context and dynamically recommending music accordingly, we aim to make everyday conversations richer and more enjoyable. This system not only deepens users' ordinary conversational experiences but also proposes a new form of entertainment through the fusion of music and dialogue.

Conversation-Aware BGM Recommendation: System Design and Setup

The system implemented in this study is a BGM recommendation system that stages everyday conversations as if they were scenes from a drama. It targets general, everyday conversations and utilizes generative AI to recommend drama BGMs suited to the conversational content based on speech-to-text transcripts. The input is conversational audio, which is transcribed to obtain conversation texts. From a pre-prepared drama dataset, the system searches for items that exhibit high contextual similarity to the transcribed

conversation. This search is handled by "Gemini," Google's large language model (LLM). The specific system flow is shown in Figure 1. By retrieving drama BGMs with high similarity and recommending those that fit the conversational context, we expect to improve the ambience of the interaction as well as participants' immersion and concentration.

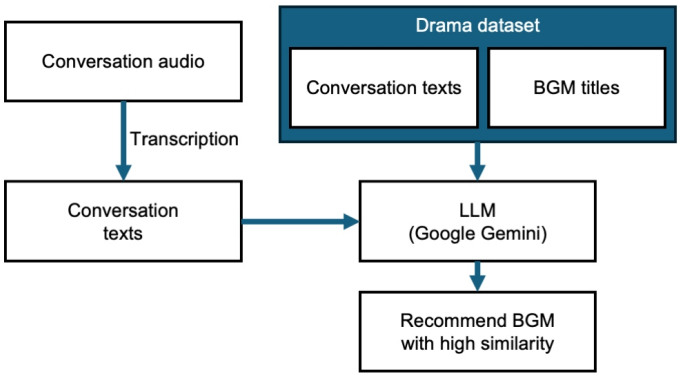

Figure 1: Flow of the system.

We first select the drama(s) to be used in the experiment. In this study, we use one drama work as a sample dataset. The selection follows three criteria:

- The drama should not focus on specialist domains such as medicine or court trials.
- Dialog should involve multiple speakers with few silent or monologic segments.
- Conversations should be close to real, everyday speech.

Based on these criteria, we use a Japanese TV drama "Ichiban Sukina Hana" ("My Favorite Flower") as the sample (hereafter, the "sample drama"). We watch the sample drama and construct the dataset used by the system.

While watching the sample drama, for each scene in which BGM is played we record the "conversation texts" (via Google Cloud Speech-to-Text) and the corresponding "BGM title." This constitutes the dataset. The sample drama has a soundtrack of 23 tracks, and roughly 10 BGMs are used per episode. In this study we watch Episodes 1–5 and target the BGMs played within them. This yields approximately 50 sets of conversation-texts and BGMs for our dataset. Because the soundtrack playlist is publicly available, we use the BGM audio published on Spotify during the evaluation experiment.

Thus, based on the sample drama we collect the conversations for scenes where BGM is present and prepare a dataset of paired conversation texts and BGM titles. Additionally, using ChatGPT (an LLM), we generate and attach a "conversation theme" similar to each drama conversation texts. This is solely to facilitate scenario reproduction during the experiment—i.e., to test whether the expected BGM is recommended when participants conduct conversations aligned with the drama scenarios. The conversation themes are not used by the recommendation system itself and have no effect on recommendation.

Because the frequency of BGM reuse varies, similar conversations could otherwise lead to the same BGM being recommended repeatedly. To mitigate this, when constructing the dataset we limit each BGM to one linked conversation text. In selecting that conversation text per BGM, we ensure the linked "conversation theme" has distinctive context that does not overlap with other BGMs. Following this policy, we prepared 16 such dataset entries.

Our system uses "Google Cloud Speech-to-Text" to convert conversational audio to texts in real time. The service recognizes 120 languages and dialects and can automatically identify the spoken language. It supports both streaming and batch recognition, accurately transcribing proper nouns, dates, and phone numbers, and it allows customizable language models to improve recognition of domain-specific terminology. After transcription, the texts are stored within the system in CSV format and—together with the dataset—is sent to Gemini via an appropriately designed prompt. Gemini offers multiple models, including experimental ones. In this system, we use Gemini 1.5 Flash, a balanced multimodal model that delivers strong performance across most tasks. Inputs and outputs are in text form, and the model performs well as a language model. When making requests to Gemini, the API request body includes both the dataset and the input conversation text. The English-translated prompt used by the system is shown below. By enforcing a specified output format, we can explicitly obtain the recommended tracks. Further, by requesting both similarity scores and rationales, we can account for cases in which a top-ranked item nevertheless has relatively low similarity. Prompt used by the system is as follows:

The following is annotation data (A) from a drama, capturing the dialog segments during which a particular BGM is played.
It includes "BGM Title" and "Conversation Text."
Annotation Data (A): (dataset file)
Next, using the actual conversation text (data below), and following the "Conditions" specified, search Annotation Data (A) for the most similar conversations to this data.
[Conditions]
* Ignore characteristics of the BGM itself.
* Because Annotation Data (A) and the conversation data were produced by a speech-to-text tool, please compensate for unnatural context where necessary.
* Compute similarity using a text-similarity measure and present the top five BGMs.
* Include numeric similarity scores. Approximate values are acceptable; rate on a 1–10 scale.
* Use the following output format: "Rank," "BGM Title," "Similarity (1–10)," "Rationale."
Conversation Data: (conversation-text file)

Evaluation Experiment With Human Participants

We conducted a system evaluation experiment with participants engaging in actual conversations using the constructed system. 17 participants (16 male, 1 female; Mean age: 28.94 years) participated for the experiment. They were divided into six groups of two or three and asked to hold two 15-minute conversations on pre-specified topics. This experiment involves our system that recommends background music (BGM) for participants' conversations based on similar dialogues found in TV dramas. Specifically, the system identifies the five most relevant BGMs that were played during comparable drama scenes. The purpose of this evaluation experiment is to verify whether the "Expected BGM" — the one anticipated by the researchers — is included among these top five AI-recommended tracks.

The results of BGM recommendation across all 12 conversational sessions (six groups, two sessions each) are shown in Table 1. "Recommendation Rank" indicates the rank position of the expected BGM within the system's output list.

Table 1: Recommendation ranks of expected BGMs across experiments.

Exp. No.	Expected BGM Top: Original Japanese Title Bottom: English Translation of Original Title	Reasoning Behind the Recommendation	Recommendation Rank
1A	……二人になれなかった4人…… Four who failed to pair off	Episodes from childhood at school or home	Below 4th place
1B	いちばん好きな花 My favorite flower	Episodes of going out or buying food together with someone	3rd
2A	しょうもないけど、好きだから It's nothing special, but I love it.	About favorite foods and drinks	1st
2B	一人で大丈夫にならなきゃ I need to be self-sufficient.	Thoughts on family and children	3rd
3A	友だちならよかった I wish we were just friends.	The difference between distinction and discrimination	3rd
3B	好かれる努力 Efforts to be liked	The idea of pursuing a career based on personal interests	1st
4A	小さな花束 A little bouquet	Episodes of giving or receiving flowers	1st
4B	いちばんすきな花-4人のひとりたち- My favorite flower -Four people, alone-	About school festival group activities and interactions	1st
5A	【オクサマ】といたアイスの棒 An ice cream stick with "wife" written on it	Discussions regarding problem solving	Below 4th place
5B	いちばん好きな花 My favorite flower	About interpersonal distance	2nd
6A	ひとりぼっち All alone	How to get along with friends	2nd
6B	学校嫌いです。I hate school.	About praising and its effects	2nd

Out of 12 trials, in 10 cases (83.3%) the expected BGM was successfully recommended within the top three ranks. In two cases (16.6%), however, the expected BGM did not appear in the ranking. For example, in Experiment 1A (ranked out), the top-ranked BGMs were associated with conversation texts about "school memories" or "friendship." Although this is partly aligned with the assigned theme of "childhood episodes at school or home," the system emphasized the "school" context more strongly than the "childhood" aspect. Consequently, other BGMs were prioritized, and the expected BGM was not recommended.

CONCLUSION

In prior conversation-based music recommendation systems, methods relying solely on emotion values or agent-mediated dialogues have been mainstream. However, they struggle to sufficiently consider conversational context and flow, making natural recommendation difficult. To address these issues, this study built a system that targets everyday conversations and recommends music by understanding conversational context. The system analyses conversation text with high accuracy using Gemini, a generative AI model. Furthermore, by combining pre-prepared drama scenes—reflecting everyday situations—with their BGMs, the system flexibly accommodates diverse conversational contents and atmospheres. This approach enables experiences in which speakers can deeply immerse themselves in the ongoing interaction—whether relaxed exchanges, formal settings, or emotionally charged scenes. To verify the advantages of this approach, we implemented the system and conducted an evaluation experiment.

We developed a system that takes conversational audio as input and recommends BGMs suited to the conversational context. Using twelve conversation themes, we conducted live conversations and tested whether the expected BGM would be recommended from the audio input. As a result, in 10 out of 12 trials (83.3%), the expected BGM was recommended within the top three ranks. For the trials that fell out of rank, although the conversations were related to the assigned themes, more specific sub-contexts were emphasized (partly diverging from the original intent of the theme), which likely caused other BGMs to be prioritized. Additionally, the actual conversational content did not always match what was anticipated, contributing to recommendations that differed from the target. These findings suggest that refining conversation themes to be more concrete and reproducible would increase the likelihood that BGMs aligned with the themes are recommended appropriately.

REFERENCES

Beutel, A., Covington, P., Jain, S., Xu, C., Li, J., Gatto, V., & Chi, E. H. (2018). "Latent Cross: Making Use of Context in Recurrent Recommender Systems", Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM, 2018).

Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., & Chi, E. H. (2019). "Top-K Off-Policy Correction for a REINFORCE Recommender System", Proceedings

of the 12th ACM International Conference on Web Search and Data Mining (WSDM '19), pp. 456–464. New York, NY: ACM.

Covington, P., Adams, J., & Sargin, E. (2016). "Deep Neural Networks for YouTube Recommendations", Proceedings of the 10th ACM Conference on Recommender Systems. New York, NY: ACM.

Kaji, K., Hirata, K., & Nagao, M. (2004). "An online music recommendation system based on annotations of context and user preference", IPSJ SIG Technical Report on Music Information Science, Vol. 2004, No. 127, pp. 33–38.

Kirdemir, B., Kready, J., Mead, E., Hussain, M. N., & Agarwal, N. (2021). "Examining Video Recommendation Bias on YouTube", in Boratto, L., Faralli, S., Marras, M., &Stilo, G. (eds.), Advances in Bias and Fairness in Information Retrieval, pp. 106–116. Cham: Springer International Publishing.

Liu, Y. C., & Huang, M. Q. (2021). "Examining the Matthew Effect on YouTube Recommendation System", Proceedings of the 2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), pp. 146–148.

Nakamura, T., Kawanishi, K., & Sakamoto, M. (2011). "Possibility of music recommendation based on lyrics and colors", IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, J94-A (2), pp. 85–94.