Supporting Smooth Interruption in a Video Conference by Dynamically Changing Background Music Depending on the Amount of Utterance

Haruki Suzawa
Osaka Prefecture University
Sakai, Japan
sdb01068@st.osakafu-u.ac.jp

Ko Watanabe
University of Kaiserslautern & DFKI GmbH
Kaiserslautern, Germany
ko.watanabe@dfki.de

Masakazu Iwamura
Osaka Metropolitan University
Sakai, Japan
masa.i@omu.ac.jp

Koichi Kise
Osaka Metropolitan University
Sakai, Japan
kise@omu.ac.jp

Andreas Dengel
University of Kaiserslautern & DFKI GmbH
Kaiserslautern, Germany
andreas.dengel@dfki.de

Shoya Ishimaru
University of Kaiserslautern & DFKI GmbH
Kaiserslautern, Germany
ishimaru@cs.uni-kl.de

ABSTRACT

Interrupting a speaker at the right moment during a meeting is an advanced skill, and not everyone can do it. It is not a rare case that one person keeps talking for a long time, particularly in a video conference, due to limited bandwidth and latency. In order to solve this problem, this paper presents a proof of concept and a working prototype of *DiscussionJockey*, an online meeting bot that measures the amount of speech of each meeting participant and provides an acoustic stimulus selected by the measurement. On the basis of a literature review, we hypothesized that the timing of speech can be implicitly manipulated by playing background music (BGM) with specific beats per minute (BPM). We conducted a pilot study using the proposed system and observed it made the utterance rate of participants closer to each other. The result of this pilot study has revealed the potential and challenges of meeting interventions.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**.

KEYWORDS

Affective Computing, Intelligent User Interface, Videoconferencing

ACM Reference Format:
Haruki Suzawa, Ko Watanabe, Masakazu Iwamura, Koichi Kise, Andreas Dengel, and Shoya Ishimaru. 2022. Supporting Smooth Interruption in a Video Conference by Dynamically Changing Background Music Depending on the Amount of Utterance . In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3544793.3560384

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09…$15.00
https://doi.org/10.1145/3544793.3560384

1 INTRODUCTION

Talking with others is an essential activity to transfer knowledge and come up with new ideas. Although COVID-19 has dramatically changed working styles, people value meetings as much or more than ever. Even after this pandemic situation ends, this trend will be likely to continue. However, compared to a face-to-face meeting, a video conference causes various problems such as a lack of real-life communication with one's boss or colleague, video delay, audio quality, and so on. Among those problems, we focused on the decrease in the quality of communication. For instance, it is always more difficult to understand the other participants' feelings in recent video conference platforms. It is usually confusing if participants understand what a speaker is talking about, especially in online lessons. In addition, it is hard to notice even if others feel like saying something. This is because it lacks the information which comes from eye gaze, the direction of sound, or body language [6].

As typical trouble, a video conference makes participants harder to cut into the conversation and leads to a collision of conversation or deviation of engagement in the meeting. This comes from difficulties in judging when is the best time to interrupt the speaker. As a solution to such trouble, we propose a system that supports passing the batons of speech to others. We have designed an online meeting bot that manipulates the speech rate of a speaker using background music and makes space for a participant who wants to start talking. We chose background music as an intervention because we wanted to give implicit feedback rather than direct suggestion. For example, Zoom has a raise hand button but some people are not willing to use that button to not interrupt the speaker.

Our research goal is to use a bot to do what people hesitate to do by themselves. People mind interrupting other people's speech so we let the speaker notice implicitly that some others are trying to say something. This work is organized as follows. We first describe our findings from the survey to support that music can control speech rate. We then describe how the system we designed works.

Next, we review the setting and result of the pilot study. Finally, we discuss our findings including the problems we need to solve.

2 BACKGROUND

Researchers have explored many ways to make online meetings more productive. To monitor audience reaction while delivering presentations, Murali *et al.* developed a system that dynamically spotlights the most expressive participant [4]. Results of this study showed that presenters self-assessed the quality of their talk more similarly to the audience members when they use the proposed system. Samrose *et al.* designed an intelligent feedback dashboard that supports to have effective and inclusive meetings [6]. They observed that the dashboard made the attendees more aware of effective and inclusive meetings.

The idea of manipulating the speech rate by changing the beats per minute is inspired by the research by Otsubo *et al.* [5]. This system manipulates the walking speed by changing the playback speed of BGM after synchronizing the tempo of walking and BGM. Researches about the relation between background music and speech rate are mainly discussing the effect of a metronome on stuttering by synchronizing a metronome and speech rate. It has been proved that stuttering can be dramatically reduced with a metronome. Although the slowing effect is not a direct factor of this treatment, they control the speech rate of people who suffer from stuttering using a metronome [1].

The question arises of how we can define an effective meeting. In this research, we will focus on the *utterance rate* and the results of Alternative Uses Task(AUT). We define *utterance rate* as the percentage of speech of each participant from the beginning of the experiment. When we evaluate the meeting with AUT, we use the idea of evaluation with AUT which is described in a paper by Hosseini *et al.* [3]. They used three categories of indices; fluency, originality, and Index of Convergence (IOC). Fluency is the total number of ideas. Originality means how creative is the idea over all groups: it scores higher if the idea occurred in fewer groups. IOC is defined as how many times the idea remained in the same category as a previously mentioned idea by the other participants.

3 PROPOSED SYSTEM

Figure 1 shows an overview of the proposed system. Our system is implemented as a Web application composed of client-side and server-side, that are connected via Web Socket. We assume that meeting participants use a video conference application like Zoom, Teams, or Google Meet. We leverage the necessary functions for video conferencing to the respective applications and perform speech amount acquisition and audio stimulation on the proposed system.

3.1 Measurement of the Amount of Speech

On the client-side, we mainly used JavaScript especially React to develop the application. It first captures the utterance of a participant with the Web API getUserMedia method. We limit the scope to the frequency of utterances of human beings. The graphs described on the left side of Figure 1 are examples of the descriptions of Fast Fourier Transform (FFT). It is updated around 60 times per minute and sums up the whole value of every frequency in the range that

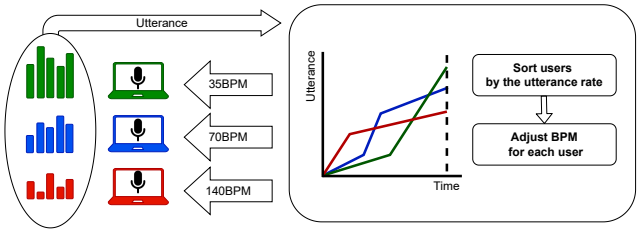


Figure 1: An overview of the workflow behind the proposed system

we have limited. After that, if the score is more than the threshold, it will be regarded as *Talking*. When it is judged as *Talking*, a variable is incremented and this value will be sent every five seconds. The value of the threshold depends on each environment so participants can adjust it with the slider on the Web page. They need to find the place where it properly says if they are *Talking* or *Not Talking* before the experiment starts.

3.2 Intervention in Meeting

On the server-side, it arranges the participants in a row according to the amount of utterance as a total score since the beginning of the experiment as Figure 1 shows. This ranking dynamically changes because the amount of utterance is sent from the client-side consistently. According to the ranking, the server-side sends commands to the client-side in order to adjust the BPM for each user. If there are three participants, the person who speaks the most in total will hear the slowest BGM (35 BPM). The second engaged person will hear the medium fast BGM (70 BPM) and the last person will hear the fastest BGM (140BPM). This BGM will also dynamically change when the ranking is updated.

We decided to use BGM with 70 BPM as a basis because Dr. Emma Gray, a cognitive behavioral therapist, worked with Spotify and found that listening to music with 50 - 80 beats per minute helps you be more creative and productive[1]. We then tried some beats and decided to use 70 BPM and simply made it double and half for other interventions. It remains a controversial issue so we will find the best BPM in the future experiment.

4 PILOT STUDY

4.1 Experimental Design

We conducted a pilot study involving three volunteers participating in an Alternative Uses Task (AUT) [2]. In AUT, participants are asked to list as many alternative usages of a given object (e.g., tennis ball, spoon, hanger). We utilized AUT as a measurement of creativity and activeness of a group meeting. In order to evaluate our hypothesis, they joined sessions with the following three conditions (10 minutes for each). In the first condition, participants discussed without BGM. This condition is prepared to record behaviors in a usual meeting situation. In the second condition, the same specific BGM (70BPM) was played in the background for all participants. By comparing with the first condition, we can investigate the general

[1]https://medium.com/taking-note/can-music-make-you-a-productivity-powerhouse-9161721fced6

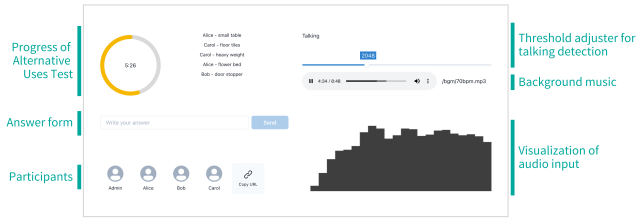
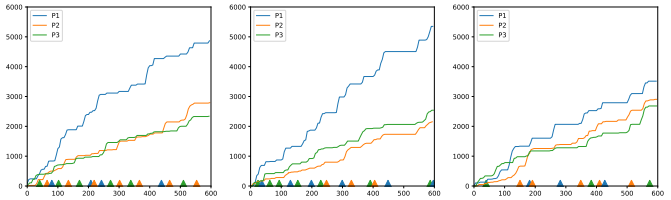


Figure 2: An experimental web page for the proposed system and Alternative Uses Task (discussing usages of a brick)



(a) Conition 1 - without BGM **(b) Conition 2 - static BPM BGM** **(c) Conition 3 - dynamic BPM BGM**

Figure 3: The cumulative amount of speech and timing of answers during 600 seconds

impact of the audio stimuli during a meeting. In the last condition, participants heard BGM with 35, 70, and 140 in order from the speaker who speaks the most.

Figure 2 shows all components that are displayed for all participants. During an experiment, participants submitted their answers about the AUT to the server by using a form. These data about utterance ranking, music intervention, and answers to the task were recorded constantly with the time stamp. In addition, an experiment conductor (admin) could access the other web page that contains buttons for start/stop recording. When the admin presses the start button, the 10 minutes timer shown in Figure 2 starts. When the timer stops, CSV files about utterances, music, and answers are downloaded.

4.2 Results

Figure 3 illustrates the transition of the amount of utterance and the timing of answers for each condition. The AUT objects were set to a tennis ball, spoon, and hanger, respectively. Although the amount of utterance of P1 was quite more than the other participants in Conditions 1 and 2, it got closer to the rest of the participants in Condition 3. Table 1 summarizes their statistics. We calculated one participant's total utterances divided by the total utterances of all participants as *utterance rate*. The *utterance rate* of each participant in Condition 3 was close to each other compared to the other two conditions. However, the total score of the number of answers was smaller than in the other conditions so we suppose this is because the topic discussed in Condition 3 was more difficult than the other questions.

4.3 Discussion

Our pilot study has revealed some challenges. First, we need to consider the characteristics of the meeting participants. The idea of this research is to give the best timing to cut into the conversation to a participant who is trying to speak by interrupting the speaker. Therefore, this system does not work if other participants have no intention to speak. For instance, if one participant speaks only when he comes up with an answer and does not communicate with others, there is no meaning in making some space to start talking, and the utterance rate does not change. Another solution to this problem is to find different tasks that require more communication with others. We also should consider the difficulties of problems and the order of conditions to conduct a larger-scale experiment and evaluate the result statistically.

Table 1: Utterance rate and the number of answers of each condition

Condition	C1: without BGM			C2: static BPM BGM			C3: dynamic BPM BGM		
Participant	P1	P2	P3	P1	P2	P3	P1	P2	P3
Utterance rate [%]	48.7	27.9	23.4	53.3	21.4	25.3	38.6	31.9	29.5
The number of answers	4	7	6	6	3	7	3	5	3

There remain controversial issues such as whether this way of intervention was appropriate or not. We played the fastest BGM for a participant who speaks the least, but there are some possibilities that it made the participant rush too much and talk rather less. Considering this, intervention might have to be the opposite of the original idea: fast music for a person speaking the most and slow music for a person speaking the least. As a different idea from using background music, we can also use haptic feedback like vibration as a way of intervention. The advantage of using audio feedback is that we don't have to prepare extra hardware devices. However, it sometimes takes too much attention and distracts participants more than needed. In contrast, if we use mobility devices, we can intervene in the real world and interrupt the speaker more effectively than background music.

5 CONCLUSION

This paper introduced DiscussionJockey, a video conference bot that helps people cut into the conversation using dynamically changing background music. Even if it is still a work in progress, our prototype and pilot study demonstrates the potential of audio stimuli during video conferences. Our underlying theme is leaving things that people hesitate to do to Artificial Intelligence. We hope to find an optimal way of intervention to realize natural conversation. In the proposed method, there is no function to identify who wants to say something, and who speaks the most is constantly interrupted by the background music. Therefore, in future work, we are planning to integrate other feature modalities such as heart rate, facial expression, and eye gaze in order to identify who wants to start talking.

ACKNOWLEDGMENTS

This work was supported by DFG ANR JST International Call on Artificial Intelligence *Learning Cyclotron* and JASSO Student Exchange Support Program.

REFERENCES

[1] John Paul Brady. 1969. Studies on the metronome effect on stuttering. *Behaviour Research and Therapy* 7, 2 (1969), 197–204.

[2] Joy P Guilford. 1967. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior* 1, 1 (1967), 3–14.

[3] Sarinasadat Hosseini, Xiaoqi Deng, Yoshihiro Miyake, and Takayuki Nozawa. 2019. Head Movement Synchrony and Idea Generation Interference – Investigating Background Music Effects on Group Creativity. *Frontiers in Psychology* 10 (2019). https://doi.org/10.3389/fpsyg.2019.02577

[4] Prasanth Murali, Javier Hernandez, Daniel McDuff, Kael Rowan, Jina Suh, and Mary Czerwinski. 2021. Affectivespotlight: Facilitating the communication of affective responses from audience members during online presentations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[5] Atsushi Otsubo, Hirohiko Suwa, Yutaka Arakawa, and Keiichi Yasumoto. 2019. BeatSync: walking pace control through beat synchronization between music and walking. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, 367–369.

[6] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. 2021. Meetingcoach: An intelligent dashboard for supporting effective & inclusive meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.