Effects of background audio changes on the perception of full-body motions of a group of synthetic characters

Fredrik Lundkvist
flundkvi@kth.se

Jesper Lundqvist
jeslundq@kth.se

Mattias Larsson
matlar4@kth.se

Yuwen Hu
yuwen@kth.se

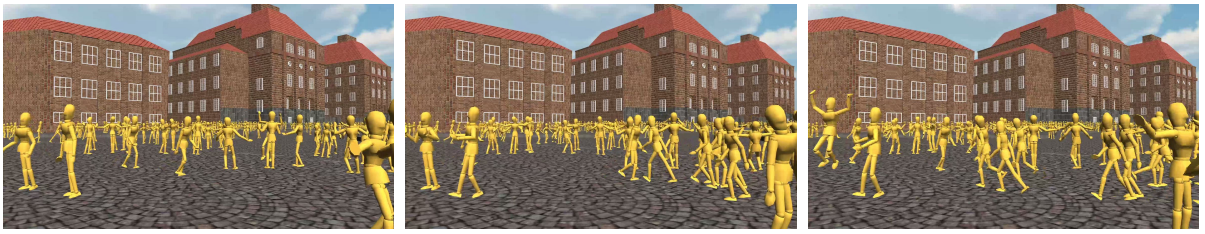


Figure 1. Screenshots of the visual stimuli shown in the experiment.

ABSTRACT

When creating realistic virtual crowds, it is important to not only create realistic stimuli, but also to keep in mind how these crowds are perceived by human observers. Most observers construct an overall perception of the crowds based on both visual and auditory stimuli. It is however unclear which of these two factors has the most impact on the impression as a whole.

To investigate how background audio in a video of a virtual crowd affected the perception of a crowd, an experiment where participants watched the same video of a virtual crowd of androgynous mannequins, rendered with the Unity game engine was conducted, with varying background audio; asking them to rate the perceived emotion of the crowd on a basic happy-neutral-sad 5-point likert scale.

The results show that the background audio has a significant effect on the perceived mood; for example a video of a happy crowd with sad background audio is perceived as sad by most of the human observers.

1. INTRODUCTION

Simulating the behaviour of crowds of people is an important element in computer graphics. Aside from technical solutions, it is important to learn how these simulated crowds are perceived by human observers.
Crowd simulations are an important element in computer graphics, with applications in visual effects, video games, architectural visualisation, and other fields. However, it is not enough to simply render a crowd and make it move naturally; to make the simulations as realistic as possible, one must also consider how human observers perceive these virtual crowds.

Since these simulations often include auditory elements, it is interesting to study how accompanying background audio affects the perception of a simulated crowd.
Earlier work has shown that when there is an emotional dissonance between visual and auditory stimuli, human observers seem to perceive the overall mood as the one communicated by the auditory stimuli [2].
The aim of this study is to build upon previous work, and examine if the same effect observed in [2] appears in the perception of full-body motion in synthetic crowds, instead of facial expressions of individual characters.

The question posed by this paper is thus:
How do changes in background audio affect a human observer's perception of a virtual crowd?

A pre-study was conducted (n=13, 10M:3F), as well as an experiment (n=22, 11M:11F), in order to find out whether differences in background audio affected perception of the mood of a virtual crowd presented in a video recording. The visual stimuli was created using 3D assets and scripts created by Ramos et al.[1], rendered using the Unity game engine, and recorded as digital video. The auditory stimuli was downloaded from the audio database Freesound[1], and edited to fit the length of the video clips. The audio clips expressed happy, neutral, and sad moods.

This paper is structured as follows: after introducing the topic of the study, some related work in this field are discussed in section 2. The methods used in the main experiment and the pre-study are presented in section 3. Section 4 shows the results of the study, which are further discussed in section 5. Section 6 concludes the paper with final remarks as well as proposals for further research.

2. BACKGROUND

On the topic of virtual crowds, several studies have investigated how variables such as physical expression of emotion [1], posture [6], character coloring [5], and walking pace [3] affect perception of different factors, such as stress and mood. One study found that walking pace seems to correlate with perceived stress in the crowd [3], while another found that character coloring did not seem to impact the perceived mood of the crowd [5].

Mower et. al studied the interaction between emotion communicated via auditory (human voices) and visual(simple animation) stimuli when observing the faces of virtual characters, and found that while both forms of stimuli had an effect on the perceived emotions of the characters, the overall perception seemed to be biased towards the audio. However, there seems to be different opinions on this subject. Åbelin found that *"The visual modality is generally dominant in perception of emotions with conflicting auditory and visual stimuli"* [8]. These contradictory findings might have their cause in the Stimuli used in the different studies; while Mower et. al used virtual characters, Åbelin used video recordings of live humans, and thus it can be argued that the difference in results is due to the fact that virtual characters are perceived differently than actual humans.

Since virtual characters were used in this study, our hypothesis is that the results should be more in line with Mower et. al's than Åbelins; however, the fact that participants are observing multiple virtual characters as opposed to just one might cause the results to differ.

3. METHOD

Participants

The study was conducted with 22 participants (11M:11F). Participants were recruited by by sending out signup link throughout online communication channels, such as Facebook and Slack. The majority of participants were students at the Royal Institute of Technology in Sweden. Some participants had impaired vision, and therefore used corrective methods, such as glasses or contact lenses. No pre-screenings were conducted to figure out if any participants had hearing or vision impairments. Participants were not informed of the purpose of the study in advance, and were only told that they would be watching a few short videos.

Stimuli

The stimuli used for the main experiments were short clips of audio and video combined. To create the clips, assets from

[1] freesound.org

Ramos et. al's [1] study on full-body expression and emotion were used. These assets included 3D mannequin models, animations, pathfinding scripts, and a 3D model of KTH. The assets were rendered using the Unity game engine, and recorded to video. The animations used from [1] were considered to be perceived as happy, based on the results of their pilot study. In total, the scene was populated by 100 mannequins.

Figure 1. Screenshots of the visual stimuli each 2.5 seconds

Androgynous mannequins were chosen, since they eliminated preconceived notions based on gender, age, or race, that might impact participants' perception of the crowd. The mannequins were also finger- and faceless. This decision was taken to get participants to focus on full-body motion, since details such as fingers and faces might divert attention from the full-body movements. The mannequins were a yellow color, this was done for two reasons: the first being that previous studies using the same assets chose to use yellow mannequins [1][3][4], the second being that the color yellow has been suggested to have a low emotional influence when using moderate lightness and hue values [7].

In addition to the visual stimuli, 3 different audio recordings of crowd noises were used, taken from the open audio database freesound.org. Of the three audio recordings, one was perceived as happy, one as neutral, and one as sad. Clip selections were based on user tags on Freesound, as well as the results of the pilot study. These three audio clips was then edited and combined with the video clip from the 3D-scene. Each finished clip used exactly the same video with the three different audio backgrounds and were later uploaded to YouTube[2].

The stimuli was viewed on a 13-inch MacBook Pro with a screen resolution of 2560 x 1600 pixels.

Figure 2. Example of the experiment setup

Participants were seated in a quiet group room at the KTH library, 62cm away from the computer kept in full brightness, and could change the angle of the screen to make sure they could see the stimuli clearly. The participant heard the auditory stimuli from the external speakers of the computer. The volume was controlled to be 75% of the full effect of the speakers.

To contextualize the stimuli, participants were told that they would be watching clips of students waiting in front of the university for a special announcement. This method of contextualization has been used in previous studies [1][4] as well. The participants were told that they would be watching three videos in total, and that they would have to

[2] Happy: https://youtu.be/Uvnf5Z6ij-8
Neutral: https://youtu.be/rvCyqmz5C5A
Sad: https://youtu.be/kHSn3iNPy78

answer a question after viewing each video twice.

Pilot study

To ensure the auditory stimuli of crowd noises used in the main study are perceived as happy, sad or neutral, a pilot study was performed in order to classify different recording of crowd noises.

In the pilot study, thirteen participants listened to 12 different 10-second audio files downloaded from the website Freesound.org. Each clip was pre-placed into one of the three categories, based on tags on the website, as well as our own ratings of the recordings. The 12 audio files used in the pilot study were uploaded to Google Drive[3].

The participants were asked to listen to each clip, ranking them one by one on a 5-point Likert scale (1=sad, 3=neutral, 5=happy). They were not told how the clips were pre-classified before or after participating in the pilot study.

After all participants had listened to and rated the recordings, average ratings were calculated for each recording. The recording with the average rating closest to the value of their pre-classification was chosen for the main study. In the case of two or more recordings having the same average rating, the recording with the least variance in ratings was selected. If two or more of these recordings had the same variance, we made a selection based on our own perception of the recordings. Three participants did not rate one of the audio clips since they were unable to hear the audio and rate it. In the end, the three audio files chosen had the following ratings: happy: avg. = 4.692, s.d. = 0,480; neutral: avg. = 3.077, s.d. = 0.277; sad: avg. = 1.077, s.d. = 0.277. The chosen audio clips are in the Google Drive folder as "Potato.wav", "Garlic.wav" and "Pear.mp3" respectively.

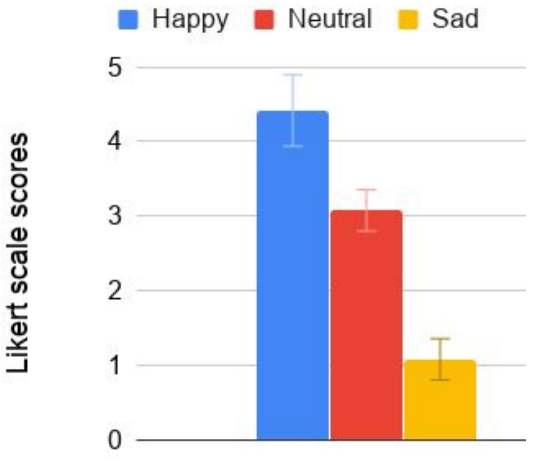


Figure 3. Results of the pilot study: the graph shows the mean ratings for the chosen audio clips, the error bars showing standard deviation.

Main experiment

The main experiment was conducted as follows: participants were seated in front of the computer, and the procedure was explained. The participant was then presented with a Google Forms questionnaire containing three videos, and three prompts to rate the perceived emotion of the crowd on the 5-point Likert scale. After having watched a video twice, the participant ranked the mood of the crowd on the likert scale.

All three clips contained the same video recording, but different audio tracks. The order in which the clips were viewed was randomized; there was one form for each possible ordering of the clips, and which form the participant would fill in was decided by randomly generating a number between one and six on the website random.org, and then presenting the corresponding form to the participant.

After all participants had completed the experiment, the collected data of Likert scale ratings was analyzed using statistical

[3] Audio clips: https://bit.ly/2OKPaPn

methods including calculating means, standard deviations and t-tests, with the purpose of investigating if there were any significant differences in rating between the different auditory stimuli.

4. RESULTS

The ratings for the different clips were as follows:

The clip with happy audio had a mean rating of 4.86 (s.d. = 0.36), the clip with neutral audio received a mean rating of 4.05 (s.d. = 0.59), and the clip with the sad audio received a mean rating of 1.71 (s.d. = 0.78).

It is worth noting the variance of the ratings increased the further the mood in the visual stimulus was from the auditory.

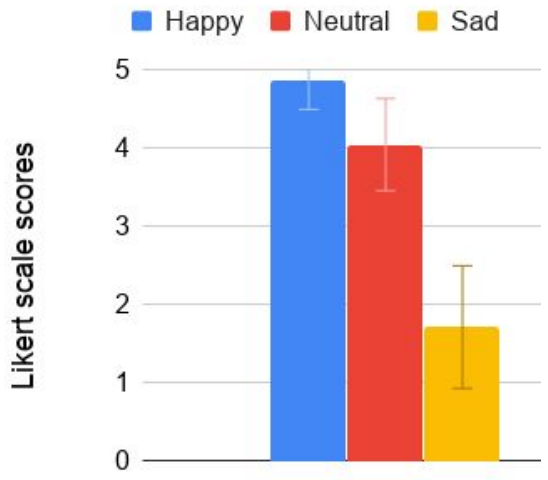


Figure 4. Mean ratings for the different audio clips. Note the increase of the standard deviation as the auditory stimuli is less congruent with the visual.

To investigate whether there were any significant differences in rating between the different clips, a set of paired t-tests were performed on every possible combination of the three data series. Significant differences in rating were found between the happy and neutral clips ($p < 0.01$), the happy and sad clips ($p < 0.01$), as well as the neutral and sad clips ($p < 0.01$).

5. DISCUSSION

As discussed by Mower et al. [2], the perception of emotion in virtual characters is perceived the best when the auditory and visual stimuli are congruent. This can be observed in the results, as the standard deviation appears to become larger for the videos with the auditory stimuli that differs the most from the visual stimuli of synthetic characters displaying a happy body language. (0.359 for happy, 0.590 for neutral and 0.784 for sad auditory stimuli).

As seen in figure 4, the mean ratings for the clips with neutral and sad audio were roughly 1 point higher on the scale than the audio in itself was rated in the pilot study. This might be explained by the fact that the full-body movements of the mannequins in the crowd were known to be perceived as happy, influencing the overall perception of the scene. More neutral animations might have led to mean values closer to the expected numbers, but there is no certain way to tell; conducting the same experiment with a different set of animation might however provide some indicative results. The optimal way would be to test each combination of animation and audio; however, it was decided that the scope of the study would become too large.

Method discussion

While partaking in the experiment, a few participants remarked that the visual stimuli did not change between clips. It is possible that, due to the nature of the stimuli, in combination with the questions asked, these participants were able to deduce the purpose of the study, and therefore might have provided biased answers.

However, the exact number of participants who noticed this is unknown, since only a few mentioned it out loud, and we cannot read minds.

Despite our efforts to avoid the ordering problem we suspect that some results may

still have been affected by it. For example, participants may have perceived a video as happy, only to move on to a later video and perceive it as even happier in comparison. Participants were allowed to go back to change their rating of previous videos after watching all the videos, but this was not communicated well enough by us. Some participants asked if this was allowed, and re-rated the videos after receiving confirmation.

We opted to let participants view a video recording of the crowd for a variety of reasons, the biggest being time constraints, both for us as well as participants. Using video recordings makes the experiment easy to setup and conduct, making it more manageable within the short timeframe we were given. However, it would be interesting to see the results of the same experiment with the modification of participants walking through the crowd in virtual reality, since they could differ from the results of this study; we propose this as a topic of some future study.

6. CONCLUSION

The results of our experiment indicate that background audio has major effect on the perceived mood of a virtual crowd. There are significant differences in perception between the different auditory stimuli, however larger sample sizes may be needed in order to provide a more accurate result.

REFERENCES

[1] Ramos C. Miguel, Qureshi Adam, & Peters Christopher. 2014. Evaluating the perception of group emotion from full body movements in the context of virtual crowds. In *Proceedings of the ACM Symposium on Applied Perception*, 7-14.

[2] Emily Mower, Maja J. Mataric, Shrikanth Narayanan. 2009. Human Perception of Audio-Visual Synthetic Character Emotion Expression in the Presence of Ambiguous and Conflicting Information. *IEEE Transactions on Multimedia* 11, 5: 843-855.

[3] Fredrik Berglund, Ellinor Jutterström, Erik Lindström, Marcus Unander. *Effects of synthetic characters' pace on stress perception*. DT2350 course project. KTH Royal Institute of Technology, Stockholm, Sweden.

[4] Alice Apostoli, Ana Granić, Victor Larsson, Petriina Piuhola. 2015. *Influence of audiovisual stimuli on emotion perception of virtual crowds*. DT2350 course project. KTH Royal Institute of Technology, Stockholm, Sweden.

[5] Joakim Larsso, Robin Tillman, Alex Wennberg. 2015. *Effects of character color changes on the perception of emotion from a crowd of synthetic characters*. DT2350 course project. KTH Royal Institute of Technology, Stockholm, Sweden.

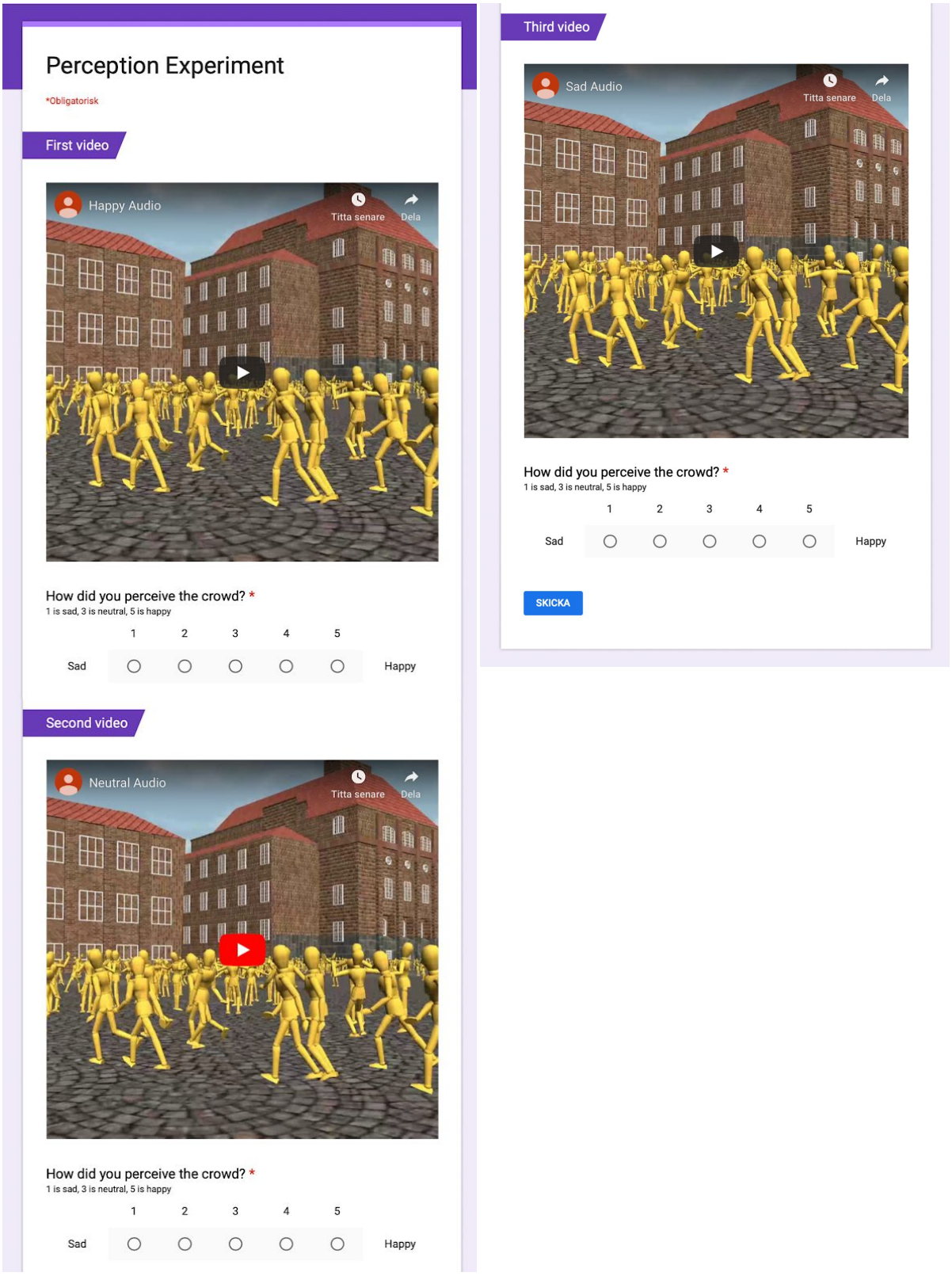
[6] Joanna E. McHugh, Rachel McDonnell, Carol O'Sullivan, Fiona N. Newell. 2010. Perceiving emotion in crowds: the role of dynamic body postures on the perception of emotion in crowded scenes. *Experimental Brain Research* 204, 3: 361-372.

[7] Haifeng Feng, Marie -J. Lesot, Marcin Detyniecki. 2010. Using association rules to discover color-emotion relationships based on social tagging. In *Knowledge-Based and Intelligent Information and Engineering Systems. KES 2010. Lecture Notes in Computer Science* 6276: 544-553.

[8] Åsa Åbelin. 2008. Seeing glee but hearing fear? Emotional McGurk effect in Swedish. In *Proceedings of Speech prosody*, 713-716.

Appendix A: One of the six questionnaires used in the study

NOTE: The titles of the videos were just “Video” during the experiment. The titles “Happy Audio”, “Neutral Audio” and “Sad Audio” were added after the experiment for publishing purposes.



Appendix B: Raw data from the study

Happy	Neutral	Sad
5	4	2
5	4	1
4	3	1
5	4	2
5	3	1
4	3	2
5	4	2
4	4	3
5	5	1
5	4	1
5	4	1
5	4	1
5	4	1
5	5	2
5	4	1
5	5	4
5	5	2
5	4	2
5	4	2
5	4	2
5	4	2

Audio type:	Happy	Neutral	Sad
Mean	4.857142857	4.047619048	1.714285714
Median	5	4	2
Standard deviation	0.3585685828	0.5895922724	0.7837638128

Appendix C: Raw data from the pre-study

Name	Happy 1	Happy 2	Happy 3	Happy 4	Neutral 1	Neutral 2	Neutral 3	Neutral 4	Sad 1	Sad 2	Sad 3	Sad 4
Code	*Potato*	*Tomato*	*Carrot*	*Corn*	*Onion*	*Garlic*	*Chili*	*Lemon*	*Lime*	*Apple*	*Pear*	*Orange*
P1 M	4	4	4	4		3	3	3	1	3	1	2
P2 M	5	4	4	3		3	2	2	1	3	1	1
P3 M	4	4	3	3		3	3	2	3	3	2	2
P4 M	5	5	5	2	3	3	3	3	1	5	1	1
P5 M	4	4	3	3	3	3	3	3	2	3	1	3
P6 M	5	5	3	3	3	3	4	3	1	3	1	1
P7 F	5	5	4	4	2	3	3	3	1	2	1	1
P8 F	5	5	4	3	2	3	3	2	2	4	1	1
P9 M	5	5	4	2	3	3	3	2	1	3	1	1
P10 M	5	5	4	3	2	3	3	3	2	3	1	1
P11 M	5	5	4	4	3	3	3	3	1	3	1	2
P12 F	5	5	4	3	2	3	4	3	2	4	1	1
P13 M	4	5	3	3	2	4	3	3	2	1	1	4

	Happy	Neutral	Sad
Mean	4,417269	3,076923	1,076923
Median	5	3	1
Standar Deviation	0,480384	0,277350	0,277350