The influence of likeability ratings of audio-visual stimuli on cortical speech tracking with mobile EEG in virtual environments

E.Wiedenmann[1*], M. Daeglau[1*], J. Otten[2], B. Mirkovic[1], G. Grimm[2], V. Hohmann[2], S. Debener[1]
[1] *Department of Psychology, Carl von Ossietzky University of Oldenburg, Germany*
[2] *Department of Medical Physics and Acoustics, Carl von Ossietzky University of Oldenburg, Germany*
Shared first authorship

Abstract

Most studies investigating auditory attention decoding (AAD) rely on audiobooks, mimicking the case of listening to someone unknown without any visual information. However, in every-day life, congruent additional visual information, especially in adverse listening situations, can facilitate auditory attention. It is also known that expressively presented auditory content is followed more attentively than neutrally presented content. Further, listening to someone we like is less effortful than to someone we rate neutrally or even negatively. Whether these findings generalize to AAD is currently unknown.

Virtual reality environments (VEs) provide a flexible opportunity to combine the reproducibility of laboratory settings with the complexity of every-day life to investigate different listening situations and factors contributing to auditory attention.

In this study, we focus on the relationship between likeability ratings of six different speakers and auditory attention depicted as cortical speech tracking via mobile electroencephalography (EEG). A total of 20 participants were presented with audio-visual scenes comprising one of the speakers at a time telling stories either with babble noise or no additional background noise. In addition, virtual speakers were animated based on the real speakers. Likeability ratings were obtained for the real speakers via a 5-point Likert scale.

We hypothesize a difference in the reconstruction accuracy between characters with a high likeability ratings compared to characters with a low likeability ratings for the videos showing the real speakers, as well as for the videos showing the real speakers without sound. We further explored the benefits off visual cues on attention. Our results will shed light on the ecological validity of VE paradigms and the role of the speaker-listener-bonding for cortical speech tracking studies.

Introduction

We typically interact with other people in multisensory, primarily audio-visual intensive environment. Every time we want to interact with another person, we pay attention to what they are saying. To better understand attention to speech, it is interesting to explore, what encourages us to pay attention and what discourages us. One major aspect of our interaction with another person is how much we like our counterpart. We make this judgment within milliseconds [1].

Attention and likeability have not been in the focus of previous audio-visual speech research. Yet there is clear evidence that emotional stimuli capture more attention than neutral ones [2, 3]. These studies focus on either emotionally valent pictures, or faces with emotional expressions and therefore do not represent the everyday experience of most people.

Studies demonstrating the effect of emotions on attention using videos are very limited. However, there has been a model for listening engagement (MoLE) developed [4]. This model states that the emotionally-colored experience, such as enjoyment, of the listener affects the motivation to engage in a story. It has been shown that the enjoyment of stories is positively correlated with the absorption of the story [5]. This study also demonstrated that expressively presented stories are followed more than neutral ones.

Virtual reality environments (VEs) have been used to successfully illustrate a realistic audio-visual environment in which participants task performance is comparable with task performance in the real environment. This has also been proven for virtual speakers in VEs [6]. There have been separate studies showing that participants can recognize emotions in virtual reality of animated characters [7]. The VE used in this experiment has been tested before and was used to enable simulating realistic audio-visual environments [8].

With natural speaker paradigms, cortical tracking of speech with mobile EEG has become increasingly popular. An increase in attention has been associated with an enhanced cortical tracking of the speech envelope [9, 10, 11].

This experiment investigates the influence of likeability ratings on attention in a VE through cortical speech tracking with mobile EEG [12]. Additionally, we explored how speech perception can benefit from visual cues. As such the study will shed light on the ecological validity of audio-visual VEs and the role of the speaker-listener-bonding for cortical speech tracking.

Methods

Participants

In this study 20 healthy participants were recruited. Data from two participants were incomplete and therefore excluded from further processing. The age ranged from 22 to 35 years and had a median of 26 years [22.07., 28.93]. 13 females and 5 males participated in this study. The subjects' native language was German, or German was learned in early childhood. Furthermore, they had self-reported normal hearing and normal or corrected-to-normal vision.

Stimuli

There were 18 different videos with six different speakers. Each video lasted between 180 and 600 seconds. The order of the videos was randomized across participants. Half of

the videos had background noise included. In videos with background noise, the speaker had also listened to noise while the videos were recorded.

The stories told by the speakers were not scripted and were cut down from a longer recording session.

Audio-visual scenes comprised one out of six speakers at a time telling stories either with babble noise or no additional background noise. Videos showed real speakers or their virtual avatars with visible lip movement or a masked mouth. Additionally, unisensory auditory and visual modalities were included.

Modalities changed in a pseudo-randomized order every 30 s while the stories' content was ongoing.

For this study, we compared the following modalities for the less likeable character with the other five very likeable characters:

- audio-only
- video-only showing the real speakers or the virtual speakers
- audiovisual showing the real speakers or the virtual speakers either with an established lip simulation algorithm based on a simple vocal tract model [13], or a new image-based DNN algorithm

Questionnaires

The participants filled out a questionnaire about the content of each video, their level of exhaustion, their level of tiredness and a questionnaire about the likeability of each speaker. The likability rating consisted of 3 different questions ("Wie sympatisch war die Person?", "Wie natürlich kam die Person rüber?", "Wie gut konntest du den Stories folgen?", "How likeable was the person?", "How natural did the person come across?", "How well could you follow the stories?") with a 5-Point Likert Scale.

Virtual environment

The VE in the lab was created with the open-source software toolbox TASCAR. It can implement a real-time low-delay high-quality interactive audio rendering environment [14].

The participants were seated approximately 174 cm in front of a 300-degree projection screen. The distance was measured from the center of the head and was kept constant over all participants. The distance was the same towards each part of the projection screens. The VE was dark except for the light reflection of the screen. The speech signals were produced by a single loudspeaker at the position of the face of the stimuli, behind the acoustically transparent screen. The background noise sound was produced by 16 full-range loudspeakers arranged in a circular array and positioned behind the screen.

EEG recording and analysis

The EEG data were recorded with a mobile EEG cap system (SMARTING, mBrainTrain, Belgrade, Serbia) which recorded from 24 scalp sites using sintered Ag/AgCl electrodes with FCz as ground and AFz as reference (Easycap, Herrsching, Germany) and mounted with a mobile EEG amplifier. The EEG data and the motion sensor signals were transmitted via Bluetooth to a recording PC positioned outside of the VE.

The EEG data were preprocessed with the EEGLAB toolbox Version 14.1.1 [12] for MATLAB (Version 9.3; MathWorks, Natick, MA, USA). The preprocessing steps followed the study by Daeglau et al. [15].

For the reconstruction of the speech envelope from the EEG data the mTRF toolbox [16] was used. The envelope reconstruction followed the steps from the study by Puschmann et al. [17].

Results

Character Ratings

Figure 1 shows the average character ratings over all the participants for the question "Wie sympatisch war die Person?" (transl.: "How likeable was the person?") on a 5-Point Likert Scale.

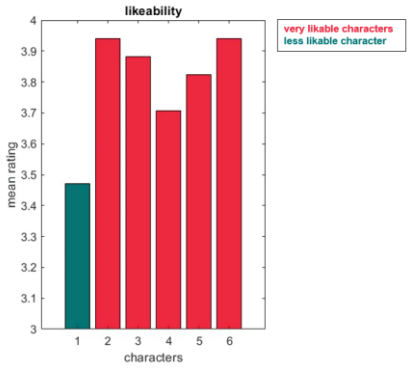

Figure 1 Average rating of the characters on a 5-Point Likert Scale, in green the less likable character and in red the very likable characters

Audiovisual envelope tracking results

A one-tailed paired t-test was performed to compare the mean speech envelope reconstruction accuracy of the less likable character and the very likable characters for the audiovisual condition (Figure 2). There was a significant difference r_z between the less likable character (M = 0.0054, SD = 0.0018) and the very likable characters (M = 0.0105, SD = 0.0007) (t(3) = -8.29, p = .004, Bonferroni corrected) for the first time window (140-187ms). For the second time window (250-296ms) there was a significant difference r_z between the character the less likable character (M = 0.0003, SD = 0.0016) and the very likable characters (M = 0.0207, SD = 0.0046) (t(3) = -10.65, p = .002, Bonferroni corrected).

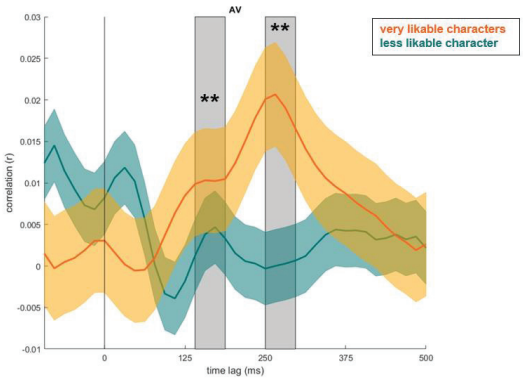

Figure 2: The mean speech envelope reconstruction accuracy r_z (± standard error of the mean) for the less likable character and the very likable characters for the audiovisual condition (**; p < .01).

Video only envelope tracking results

For the video only condition, a one-tailed paired t-test was performed to compare the mean speech envelope reconstruction accuracy of the less likable character the very likable characters for the video only condition (Figure 3). There was no significant difference r_z between the less likable character (M = -0.0036, SD = 0.0063) and the very likable characters (M = 0.0012, SD = 0.0008) (t(3) = -1.50, p = .231) for the first time window (140-187ms, Bonferroni corrected). For the second time window (250-296ms) there was no significant difference r_z between the less likable character mean likeability (M = -0.0068, SD = 0.0070) and the very likable characters (M = 0.0020, SD = 0.0012) (t(3) = -2.84, p = .066, Bonferroni corrected).

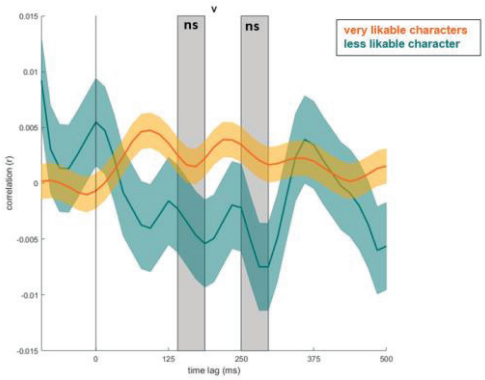

Figure 3: The mean speech envelope reconstruction accuracy r_z (± standard error of the mean) for the lowest and highest rated character for the visual only condition (ns; p > .05).

All modalities envelope tracking results

A one-way ANOVA was performed to compare the effect of the visual conditions on r_z (Figure 4). For the first time window (140-187ms) a one-way ANOVA revealed that there was a statistically significant difference for r_z between at least two groups (F(2, 9) = [102.8], p < .001). A Tukey's HSD test for multiple comparisons indicated that r_z was significantly different between the video only and audio only conditions (p < .001, 95% C.I. = [-0.019 -0.011]) as well as between the video only and audiovisual conditions (p < .001, 95% C.I. = [-0.023, -0.015]). There was no statistically significant difference between audiovisual and audio only conditions (p = .053). For the second time window (250-296ms) a one-way ANOVA revealed that there was a statistically significant difference for r_z between at least two groups (F(2, 9) = [56.41], p < .001). Tukey's HSD Test for multiple comparisons indicated that r_z was significantly different between the video only and audio only (p < .001, 95% C.I. = [-0.028 -0.015]) as well as between the video only and audiovisual (p < .001, 95% C.I. = [-0.030, -0.016]). There was no statistically significant difference between audiovisual and audio only conditions (p = .861).

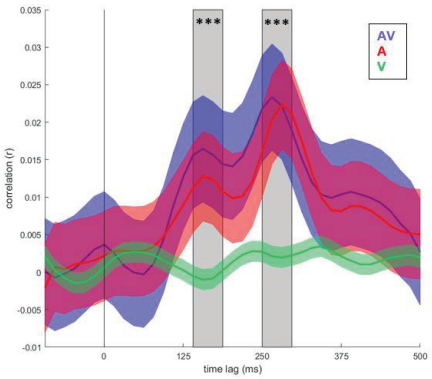

Figure 4: The mean speech envelope reconstruction accuracy r_z (± standard error of the mean) for the audiovisual (AV), audio only (A) and video only (V) condition (***; p < .001).

Discussion

The study investigated the influence of likeability ratings on attention in a VE through cortical speech tracking with mobile EEG. An influence was confirmed for the audiovisual condition but not for the video only condition. The study also investigated whether visual cues are beneficial for attention to speech. As predicted, we found better AAD for audiovisual and audio only stimuli in comparison to video only stimuli.

There was a significant effect of likeability on attention though speech envelope tracking for the audiovisual condition. It could have been that the effect was driven more by the content of the stories or the voices of the speakers rather than the likeability of the characters. More likely this finding indicates that the evaluation of our counterpart and how much we like them, influences how much attention we pay to them in everyday life. This is consistent with previous research [2, 5], which demonstrated that emotional as well as expressively presented stories should capture attention more.

For the video only condition, we did not observe an influence of likeability on speech envelope tracking. This is in contrast to previous research [2], which indicated that emotional stimuli, even without sound should capture more attention than neutral ones. The findings may imply that speech has a major influence on the likeability ratings of the characters or even the stories themselves. Nevertheless every character told two different stories to try and get a mixture of different stories and to not get the listeners biased for the stories.

It should be noted that the likeability ratings for the six characters used in this study were rather similar. Five out of the six characters were rated as very likeable and only one character was rated as much less likeable. Future research should use stimulus material covering a larger range of likeability ratings to better address the influence of character likeability on speech envelope tracking.

There was a significant effect of the video conditions on attention through speech envelope tracking. However it was only shown that there was a significant difference between the audiovisual and video only as well as between the audio only and video only condition. This partially supports the findings from previous research [17] which found that there was a significant increase for the audiovisual condition

compared to the audio only and visual only condition. This suggests that the additional visual input did not aid young, healthy participants paying more attention to the stimuli as it was the case in a previous study with elderly hearing impaired [17].

Our results add to the claim [8] that complex experiments regarding audiovisual attention can be conducted in VE. Future studies should use different characters representing a larger range of likeability to further our understanding of the role of engagement on speech processing.

Acknowledgments

This research is funded by a research grant from the German Research Foundation (Deutsche Forschungsgemeinschaft; SPP 2236 project 444761144).

The authors thank Ina Cera and Focke Schröder for their valuable contributions to the set up and Jennifer Decker for helping with the data collection.

Literature

[1] Bar, Moshe; Neta, Maital; Linz, Heather (2006): Very first impressions. In *Emotion (Washington, D.C.)* 6 (2), pp. 269–278. DOI: 10.1037/1528-3542.6.2.269.

[2] Schupp, Harald T.; Junghöfer, Markus; Weike, Almut I.; Hamm, Alfons O. (2003): Attention and emotion: an ERP analysis of facilitated emotional stimulus processing. In *Neuroreport* 14 (8), pp. 1107–1110. DOI: 10.1097/00001756-200306110-00002.

[3] Anderson, Adam K. (2005): Affective influences on the attentional dynamics supporting awareness. In *Journal of experimental psychology. General* 134 (2), pp. 258–281. DOI: 10.1037/0096-3445.134.2.258.

[4] Herrmann, Björn; Johnsrude, Ingrid S. (2020): A model of listening engagement (MoLE). In: *Hearing research* 397, S. 108016. DOI: 10.1016/j.heares.2020.108016.

[5] Herrmann, Björn; Johnsrude, Ingrid S. (2020): Absorption and Enjoyment During Listening to Acoustically Masked Stories. In: *Trends in hearing* 24, 2331216520967850. DOI: 10.1177/2331216520967850.

[6] Hendrikse, Maartje M.E.; Llorach, Gerard; Grimm, Giso; Hohmann, Volker (2018): Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters. In: *Speech Communication* 101, S. 70–84. DOI: 10.1016/j.specom.2018.05.008.

[7] Geraets, C. N. W.; Klein Tuente, S.; Lestestuiver, B. P.; van Beilen, M.; Nijman, S. A.; Marsman, J. B. C.; Veling, W. (2021): Virtual reality facial emotion recognition in social environments: An eye-tracking study. In: *Internet interventions* 25, S. 100432. DOI: 10.1016/j.invent.2021.100432.

[8] Fuglsang, Søren Asp; Dau, Torsten; Hjortkjær, Jens (2017): Noise-robust cortical tracking of attended speech in real-world acoustic scenes. In: *NeuroImage* 156, S. 435–444. DOI: 10.1016/j.neuroimage.2017.04.026.

[9] Holtze, Björn; Jaeger, Manuela; Debener, Stefan; Adiloğlu, Kamil; Mirkovic, Bojana (2021): Are They Calling My Name? Attention Capture Is Reflected in the Neural Tracking of Attended and Ignored Speech. In: *Frontiers in neuroscience* 15, S. 643705. DOI: 10.3389/fnins.2021.643705.

[10] Kurthen, Ira; Galbier, Jolanda; Jagoda, Laura; Neuschwander, Pia; Giroud, Nathalie; Meyer, Martin (2021): Selective attention modulates neural envelope tracking of informationally masked speech in healthy older adults. In: *Human brain mapping* 42 (10), S. 3042–3057. DOI: 10.1002/hbm.25415.

[11] Crosse, Michael J.; Di Liberto, Giovanni M.; Bednar, Adam; Lalor, Edmund C. (2016): The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. In: *Frontiers in human neuroscience* 10, S. 604. DOI: 10.3389/fnhum.2016.00604.

[12] Delorme, Arnaud; Makeig, Scott (2004): EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. In: *Journal of neuroscience methods* 134 (1), S. 9–21. DOI: 10.1016/j.jneumeth.2003.10.009.

[13] Llorach, G., Evans, A., Blat, J., Grimm, G., & Hohmann, V. (2016, September). Web-based live speech-driven lip-sync. In 2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES) (pp. 1-4). IEEE.

[14] Grimm, G., Luberadzka, J., & Hohmann, V. (2019). A toolbox for rendering virtual acoustic environments in the context of audiology. *Acta acustica united with acustica*, *105*(3), 566-578.

[15] Daeglau, Mareike; Zich, Catharina; Welzel, Julius; Saak, Samira Kristina; Scheffels, Jannik Florian; Kranczioch, Cornelia (2021): Event-related desynchronization in motor imagery with EEG neurofeedback in the context of declarative interference and sleep. In: *2666-9560* 1 (4), S. 100058. DOI: 10.1016/j.ynirp.2021.100058.

[16] Crosse, Michael J.; Di Liberto, Giovanni M.; Bednar, Adam; Lalor, Edmund C. (2016): The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. In: *Frontiers in human neuroscience* 10, S. 604. DOI: 10.3389/fnhum.2016.00604.

[17] Puschmann, Sebastian; Daeglau, Mareike; Stropahl, Maren; Mirkovic, Bojana; Rosemann, Stephanie; Thiel, Christiane M.; Debener, Stefan (2019): Hearing-impaired listeners show increased audiovisual benefit when listening to speech in noise. In: *NeuroImage* 196, S. 261–268. DOI: 10.1016/j.neuroimage.2019.04.017.