10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Listeners and Lookers: Using Pitch Height and Gaze Duration for Inferring Mental States

Juliane T. Zimmermann[1], *Simon Wehrle*[2], *Francesco Cangemi*[2], *Martine Grice*[2], *Kai Vogeley*[1,3]

[1]Department of Psychiatry, Faculty of Medicine and University Hospital Cologne, University of Cologne, Germany; [2]IfL – Phonetics, University of Cologne, Germany; [3]Institute of Neuroscience and Medicine, Cognitive Neuroscience (INM-3), Research Centre Juelich, Germany

juliane.zimmermann@uk-koeln.de

Abstract

Pitch height and gaze duration are used to infer other people's mental states, e.g. their attentional focus, attitudes or emotions. To shed light on the interplay of these two cues we varied pitch height in German utterances and gaze duration in a paradigm including a virtual character and different objects. At a group level, greater pitch height and longer gaze duration on a given object similarly increased participants' ratings of the perceived importance of that object to the virtual character. At the individual level, most participants showed a tendency to be influenced predominantly by only one of the two channels (pitch or gaze). The data suggest a high interindividual variability in the employment of the different, potentially competing nonverbal cues used in estimating the thoughts and judgment of another person.

Index Terms: pitch height, gaze duration, mental state, individual behavior, prominence

1. Introduction

Interaction substantially relies on complex and multimodal nonverbal communication [1]. Meaning can be conveyed and extracted in speech material beyond lexical content on the basis of prosody [2]. In the visual domain, gaze behavior plays a crucial part in conveying and inferring information in social communication [3]. Both of these types of nonverbal cues, prosody and eye gaze, provide us with important information in their respective domain. We can make use of them to infer others' mental states, e.g. their attentional focus, attitudes or emotions. In complex social encounters, most often a combination of more than one channel is involved.

1.1. Intonation as a key to inferring mental states

Prosody can be used to indicate that something is important, new or in the focus of attention, be it an aspect of conversation, or an object in the environment. In German this is achieved through pitch accent placement and type, cued primarily by fundamental frequency, perceived as pitch [4]. For successful communication, speakers take into account what the listener already knows ('givenness') and then apply prosody appropriately [5]. For listeners, pitch is especially relevant to the perception of prosodic prominence [6], indicating how far something is marked as important [7]. Even without speakers intentionally conveying this information, they may inadvertently transmit inferable prosodic cues to what is important in the current situation or to the speakers themselves [8]. Thus, not only 'face-value-importance' is communicated prosodically. Rather, by successfully decoding prosodic information, listeners can infer a speaker's intentions, thoughts and feelings; emotional states can also be encoded and decoded prosodically [9], [10].

1.2. Gaze as a key to inferring mental states

Gaze behavior is a very strong signal by which we express our inner experience. We direct our eyes towards objects we pay attention to and inform others about whether these objects are of general interest or importance in a specific situation [11], [12]. From an early age, our gaze is drawn towards new objects [13], which are likely to be more interesting and informative. We are also able to interpret observed gaze behavior. Another person's directed gaze can lead the observer to attend to the same direction [14]. Gaze directed towards objects can help us understand which object might be especially important in a given situation [15]. Moreover, gaze direction and duration are indicative of preferences [16], [17]: we tend to look longer and more often at preferred stimuli compared to non-preferred stimuli. Crucially, observers are able to interpret gaze duration [18] and direction [19], [20] towards preferred or desired objects.

1.3. Integrating visual and auditory cues

In real life, we are forced to make sense of complex stimuli from many different sources of information and to integrate them into a coherent representation of what is being communicated. A combination of auditory and visual information can be helpful in the interpretation of a message if the incoming information is difficult to understand (e.g. due to noise [21]), but can also be detrimental if both channels provide conflicting information [22]. Likewise, acoustic and visual information are integrated to infer how important a particular object might be for another person. Visual information, such as head nods or eyebrow raises, can increase prosodic prominence perception if it is already present, thus indicating an additive effect of visual and auditory information for prominence ratings [23]. When asked to identify prominent elements of spoken sentences presented in video sequences, the upper half of the head including the eye region is particularly informative [24]. However, it is unclear how exactly prosodic prominence and gaze are used to infer another person's mental states, e.g. importance ratings.

1.4. Study Design and Hypotheses

In the current study, we systematically compare the effect of the two information channels, prosodic prominence and gaze behavior, on the perceived attitude of an agent, i.e. a virtual character towards objects in her environment (importance judgement). More precisely, we manipulate pitch height and gaze duration, both allegedly attributable to the virtual character. Participants are asked to rate how important the object present

10.21437/SpeechProsody.2020-59

in the current situation is to the 'person' represented by the virtual character. We expected participants' perception of the virtual character's mental state to be affected by both a higher pitch excursion on the word referring to an object (suggesting a more prominent pitch accent type) and longer gaze duration of the virtual character towards that object. Specifically, we expected to find an increase of participants' ratings of importance of the object for the virtual character if the object was presented with a more prominent accent type and/or longer gaze duration compared to a less prominent accent type and shorter gaze duration. As it has been reported that less frequent words elicit greater prominence perception [25], [26], we also expected word frequency to have a general influence on ratings.

2. Material and Methods

We tested both the individual and combined influence of pitch height and gaze duration on participants' ratings of the importance of objects for a virtual character. We presented 106 different video sequences of a virtual character's face positioned above an object with a duration of 6.6 s. Depicted objects were different in each trial and each object was presented only once. One female virtual character was presented, corresponding to recordings from one female speaker. The movements performed by the agent were limited to the eyes. The agent's attention towards the object suggesting high importance was operationalized as an auditorily presented utterance with a higher pitch excursion and a longer gaze duration directed towards the object.

2.1. Experimental design

We systematically varied the factors 'pitch height' and 'gaze duration towards the object' on two levels. Pitch height on the accented syllable was either comparatively low or high. Gaze duration towards the object was either comparatively short (0.6 s) or long (1.8 s). Thus, we effectively created four conditions establishing a 2 x 2 experimental design: low pitch and long gaze, low pitch and short gaze, high pitch and short gaze, high pitch and long gaze.

2.2. Selection of objects

We selected 106 different images of objects from a pre-established and well-characterized set of images [27] based on their referential expressions. To reduce any possible influence of the number of syllables on the perception of word prominence, we only selected words with two syllables and penultimate stress. These were most frequent in the set and allowed us to avoid any interference effects due to word boundary effects. Additionally, we partly excluded well-known and often used homonyms.

2.3. Auditory stimulus material

Auditory stimuli comprising the German two-syllable words denoting the 106 different objects including the definite article (e.g. "der Toaster": "the toaster") were created from an H*-accented rendition of each of the 106 target phrases produced by a trained female speaker. An analysis of H* and L+H* on a subset of target words by this speaker indicated that she mainly modulated F0-peak height in differentiating between these two categories. Recordings took place in a soundproof booth, using an AKG C420L headset microphone connected to a computer running Adobe Audition via a USB audio interface (PreSonus AudioBox 22VSL). Stimuli were recorded with a sampling rate of 44100 Hz, 16 bit. We subsequently edited F0-peak height on the target words, so as to obtain a lower and a higher pitch peak (henceforth low and high), with a difference of 45 Hz. As other parts of the pitch contour were unchanged, higher peaks led to greater pitch excursions. Stimuli were tested for 'naturalness' and accent type by six trained phoneticians. Stimuli produced for the 'low' and 'high' condition were rated as sounding natural in 92.14 % and 74.37 % of cases, respectively, and were rated as H* and L+H*, respectively, in 83.65 % and 78.46 % of cases. The resulting speech stimuli were normalized to equal loudness. F0 was edited using smoothing [28], stylisation and resynthesis [29]. Examples are provided in the online multimedia files.

2.4. Visual stimulus material

Video sequences were created by arranging a picture of the female virtual character and an image of one of the 106 different objects in a vertical fashion (Figure 1). At the beginning and the end of the video, the agent exhibited idle gaze behavior, i.e. she performed gaze movements directed towards random locations in the environment. The agent fixated neither the object nor the participant during these phases.

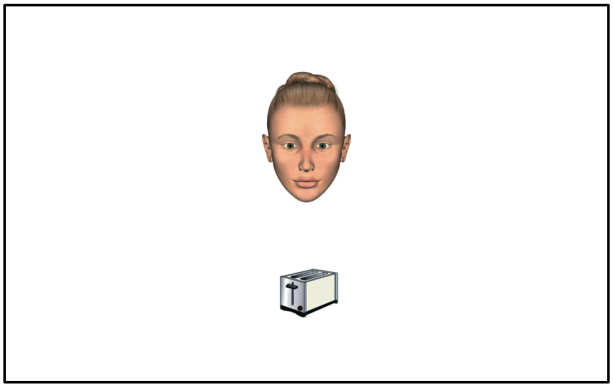

Figure 1: *Still of an example video.*

After 2.0 s, the virtual character looked at the participant for 1.0 s. This gaze was included to induce the experience of mutual interaction between virtual character and participant [30]. Subsequently, the virtual character directed her gaze towards the object (Fig. 1). This fixation lasted either 0.6 s (short gaze conditions) or 1.8 s (long gaze conditions). These durations are based on findings from human-robot interaction regarding different gaze durations and their perception [31]. Afterwards, the agent looked at the participant again. This gaze sequence (participant, object, participant) was preceded and followed by a blink, i.e. the agent's eyes closing for 0.1 s to simulate naturalistic blinking behavior. In addition, some video animations included an additional blink during one or both of the first and second idle, non-communicative phases. After the described gaze sequence, the virtual character continued gazing at random locations until the end of the video, lasting for either 2.0 or 0.8 s depending on short or long gaze conditions, so as to keep the total presentation duration of the object image of 6.6 s constant across videos. All images of the agent's face were taken from a study investigating the perception of gaze direction [32].

Video creation and integration of auditory stimuli was carried out using Python [33] and the FFmpeg module [34]. We created a total of 424 videos (106 per condition). Example videos are provided in the online multimedia files.

2.5. Participants and procedure

We recruited 64 monolingual native German speakers aged between 18 and 65 via an online platform (www.prolific.ac). They were reimbursed with 3.25 Euro for their participation. The study was performed in SoSci Survey [35]. Participants were instructed to imagine that the utterances they perceived were produced by the character on screen. They were informed that the character can convey the importance of the object. Participants were then instructed to answer the same question after each trial: "How important is the object to the virtual character?" (German: "Wie wichtig findet die Figur das abgebildete Objekt?"). Participants were presented with half of the stimuli to keep the task short. Each trial consisted of a video and its subsequent rating. Items were presented in randomized fashion. Each video sequence was followed by a screen asking for the rating on a scale from 1 to 4: 1="not important at all", 2="rather unimportant", 3="rather important", 4="very important".

2.6. Analysis

Data was analysed with R [36] in RStudio [37]. A Bayesian ordinal model (r package 'brms' [38]) was fitted to the data. Fixed effects for participants' ratings were 'gaze duration', 'pitch height', their statistical interaction and the logarithmized and z-transformed values for word frequency of the objects in German [39]. As random effects, we included random intercepts and slopes for the 'subject' effect, and random intercepts for the 'object' effect. A weakly informative prior was used (intercept prior: normal distribution, $M = 2.5$, $SD = 1.5$; slope priors: normal distribution, $M = 0$, $SD = 2$; SD prior: normal distribution, $M = 0$, $SD = 2$). The model ran with four sampling chains of 12,000 iterations each and a warm-up period of 2,000 iterations.

3. Results

The condition characterized by low pitch height and short gaze duration yielded the lowest mean ratings. The condition with both high pitch and long gaze duration yielded the highest mean ratings. The conditions with either increased pitch height or longer gaze duration yielded mean ratings in a middle range between the two aforementioned conditions. Mean ratings within the four conditions corroborated the initial hypotheses (Fig. 2).

Overall, there is strong evidence for our model as opposed to the model not including the factors 'pitch height' and 'gaze duration' ($BF_{10} > 1000$). Higher pitch increased the ratings by 0.56 standard deviations (SD) on the latent rating scale, 95% CI = [0.33, 0.79]. Likewise, longer gaze duration also increased the ratings ($\hat{\beta} = 0.65$, 95% CI = [0.39, 0.91]. In this study, both effects had comparable effect sizes. Their statistical interaction did not affect ratings ($\hat{\beta} = 0.02$, 95% CI = [-0.14, 0.18]). Higher word frequency increased the ratings ($\hat{\beta} = 0.06$, 95% CI = [0.01, 0.12]). The random subject effects were considerable in the model (random intercepts: $\hat{\beta} = 0.57$, 95% CI = [0.47, 0.70]; random effect of pitch height: $\hat{\beta} = 0.87$, 95% CI = [0.71, 1.06]; random effect of gaze duration: $\hat{\beta} = 1.02$, 95% CI = [0.83, 1.24]), except for the random interaction effect ($\hat{\beta} = 0.12$, 95% CI = [0.00, 0.35]), which was not statistically robust. The random object effect, however, was statistically robust ($\hat{\beta} = 0.18$, 95% CI = [0.11, 0.24]).

3.1. Explorative analysis of individual behavior

At the individual level, the effects of pitch height and gaze duration accounted for a change of 0.87 and 1.02 SD on our rating scale, respectively. Therefore, we further investigated to what extent the factors predicted the ratings for each individual participant. Figure 3 shows the individual slope coefficients for the two factors for each subject. Participants' ratings tended to be influenced by either 'pitch height' or 'gaze duration' rather than by both factors in combination. This was mirrored by a negative correlation ($\hat{\beta} = -0.48$, 95% CI = [-0.68, -0.23] of the factors 'pitch height' and 'gaze duration' within the random subject effect.

To identify possible subgroups based on cue 'preference', we applied a hierarchical cluster analysis [40] using Euclidian distance and Ward's method. The resulting classification suggested a two- or three-cluster solution. The clustering is included in Fig. 3, showing the three distinct groups. Due to the degree to which participants took into account pitch height and gaze duration for their ratings, we labelled them 'Listeners', 'Lookers' and 'Neither'. In the two-cluster solution, 'Listeners' and 'Neither' clustered together.

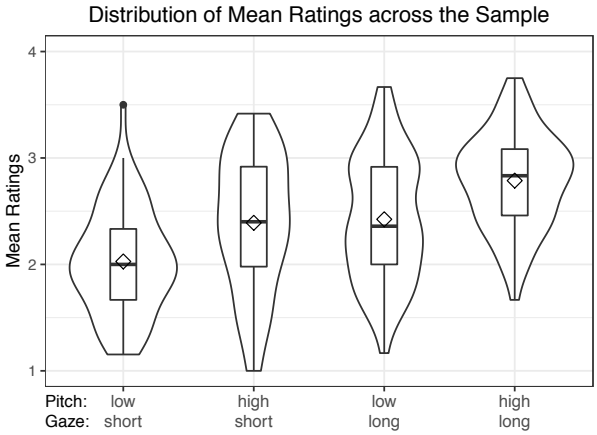

Figure 2: *Distribution of participants' mean ratings of stimuli. The range of the y-axis equals the total rating scale (1-4). Diamonds indicate means across subjects.*

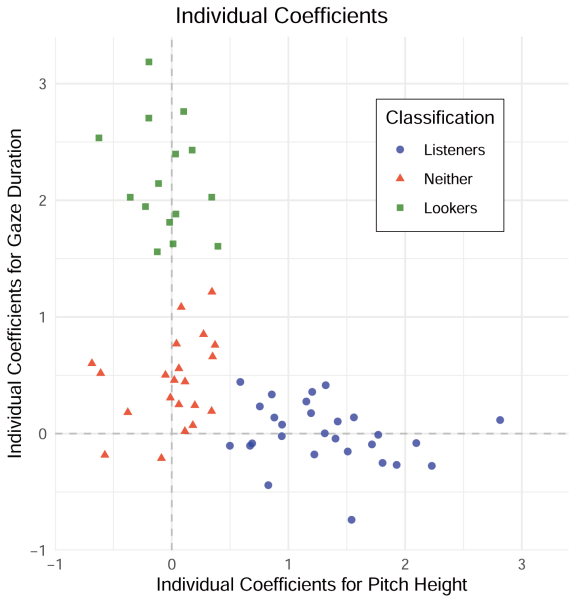

Figure 3: *Individual slope coefficients for pitch height (x-axis) and gaze duration (y-axis) by participant.*

4. Discussion

We investigated how participants' ratings of the ascription of a 'person's' attitude, here the perception of object importance, were affected by pitch height (of utterances referring to target objects) and gaze duration (directed towards target objects by the virtual agent). According to our initial hypothesis, both factors substantially affect the ratings of participants.

The condition characterized by low pitch height and short gaze duration towards the object resulted in the lowest mean ratings of importance. Highest mean ratings were registered for the condition characterized by high pitch and long gaze duration. The mean ratings for the two 'mixed' conditions in which only one of the two signals indicates importance were in the medium range between these two extremes. The effects of pitch height and gaze duration on participants' ratings presented as similar and statistically robust in our analysis. Our findings corroborate previous findings showing that acoustic and visual cues are relevant sources from which the mental states of others can be inferred [8], [9], [18], [19]. Interestingly, most participants were influenced by only one of the two cues, dividing the whole group into either 'Listeners', 'Lookers' or 'Neither'.

Our data do not provide evidence for interactional effects of pitch height and gaze duration in the current task. This is in line with previous studies investigating the interplay of prosody and visual body cues and showing no interaction of the two: Only an additive effect of eyebrow raises and pitch accents was shown for the perception of word prominence [23] whilst no interactional effect of general visual facial information and prosody on prosodic prominence ratings was observed [24]. In our study, participants tended to not take into account both at the same time, so that we cannot conclude that effects in our study add up to contribute to the perception of object importance to the virtual agent.

Higher word frequency was associated with higher importance ratings. At first glance, this seems to contradict the observation that infrequent words elicit higher prominence perception [25], [26]. However, in our dataset, word frequency was intertwined with other properties that also affect the perception of importance: the five most frequent words in our dataset were the German words for 'car', 'key', 'eye', 'plane' and 'finger'. The five least frequent words were the German words for 'spinning wheel', 'doorknob', 'seal', 'spinning top' and 'roller skate'. We assume that relevance for everyday life affected the ratings, so that word frequency and general importance were correlated in our study. Taking a look at the five 'most important'-rated items ('key', 'traffic light', 'spoon', 'brush', 'sun') and the five 'least important'-rated items ('desk', 'peanut', 'church', 'sandwich', 'seal') corroborates this notion.

Variance introduced by individual participants was substantial. Great individual variability has been reported for influence of pitch on the perception of prosodic prominence, i.e. the degree to which words are perceived as highlighted or important [7], [25]. Moreover, the perception of prosodic prominence is not only influenced by speaker and listener characteristics, but also by a combination of both [44]. Interpretation of directional gaze cues also depends on individuals (e.g biological sex [41]).

We found that participants' ratings tended to be influenced by either pitch height or gaze duration or by neither of the two cues, but never by both at the same time. This led us to the identification of groups of 'Listeners', 'Lookers' and 'Neither'. We reject the two-cluster solution because it is theoretically not convincing to cluster participants making use of pitch height with participants making use of neither cue. The existence of a 'neither'-group in our study does simply allow for the conclusion that these people did not take into account either cue. Other possible explanations are suggested in the following paragraph. As for the differentiation of participants into 'listeners' and 'lookers', other studies have provided similar findings. In a production study, it was shown that participants increase speaking efforts and change their gaze behavior to improve communication [42]. However, the authors did not find a strong correlation of both measures. This supports the idea that the majority of people focus more on one channel than the other. Another study reports that people tend to produce pitch accent categories either by altering the shape of the F0-contour peak or its timing [43]. In perception studies, similar results have been reported: Persons rating prosodic prominence tend to concentrate on either prosodic cues or visual facial information [24]. Similarly, studies concerned with the perception of prosodic prominence as well as its reproduction report a division into one group of subjects that relies mainly on pitch and another that relies more on other aspects (such as word frequency) [7], [25], [44].

There are some limitations to the study. First of all, the findings of this reductionistic design cannot be easily transferred to any kind of complex social situation. We created a situation devoid of variation of other cues usually present in a comparable real-life situation. While the voice stimuli were derived from natural speech, the virtual character was not seen to move her mouth along with the presentation of the utterance. Moreover, neither the virtual agent nor the objects were photo-realistic depictions. Second, people were informed that the virtual character is able to convey the importance of the object. This information might have led participants to actively search for a cue to make sense of the otherwise uninformative setting and stop searching once one valid cue (out of two possible cues) was identified as a reliable source of information. This might have led participants to not make use of both cues, which would explain why we did not find an interaction of pitch height and gaze duration. Third, participants were required to indicate how important the character finds the depicted object. Even with the information that the agent can indeed convey importance, the task still is rather vague and relies on subjects' perceptual and mentalizing skills. This could lead to participants having trouble integrating the cues as meaningful in this rather unnatural setting or to them being reluctant to assign mental capacities to a virtual agent in the first place. We did not collect participants' ratings of general object importance.

5. Conclusion

Pitch excursion and gaze duration can be used to infer the mental state of a virtual character. Persons differ in terms of the degree to which they make use of these cues to infer the importance of an object to the virtual character.

The study's limitations might be overcome by using a more life-like experimental environment, including variation of other cues, along with a more engaging task. Future studies could benefit from further investigation into the individual factors accounting for the substantial amount of variability found in the present study.

6. Funding

The study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 281511265 – SFB 1252.

7. References

[1] J. K. Burgoon, L. K. Guerrero, and V. Manusov, "Nonverbal signals," in *The SAGE Handbook of Interpersonal Communication*, 4th ed., M. L. Knapp and J. A. Daly, Eds. Thousand Oaks: SAGE Publications, Inc., 2011.

[2] J. D. O'Connor and G. F. Arnold, *Intonation of Colloquial English*, 2nd ed. London: Longman, 1973.

[3] M. Argyle, R. Ingham, F. Alkema, and M. McCallin, "The Different Functions of Gaze," *Semiotica*, vol. 7, no. 1, pp. 19–32, 1973.

[4] C. Féry and F. Kügler, "Pitch accent scaling on given, new and focused constituents in German," *Journal of Phonetics*, vol. 36, no. 4, pp. 680–703, Oct. 2008.

[5] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, "Acoustic correlates of information structure," *Language and Cognitive Processes*, vol. 25, no. 7–9, pp. 1044–1098, Sep. 2010.

[6] D. Arnold, P. Wagner, and H. Baayen, "Using generalized additive models and random forests to model prosodic prominence in German," 2013.

[7] S. Baumann and B. Winter, "What makes a word prominent? Predicting untrained German listeners' perceptual judgments," *Journal of Phonetics*, vol. 70, pp. 20–38, Sep. 2018.

[8] C. Kaland, E. Krahmer, and M. Swerts, "White Bear Effects in Language Production: Evidence from the Prosodic Realization of Adjectives," *Lang Speech*, vol. 57, no. 4, pp. 470–486, Dec. 2014,

[9] W. Thompson and L.-L. Balkwill, "Decoding speech prosody in five languages," *Semiotica*, vol. 2006, pp. 407–424, Jan. 2006.

[10] K. R. Scherer, "Vocal affect expression: a review and a model for future research," *Psychol Bull*, vol. 99, no. 2, pp. 143–165, Mar. 1986.

[11] A. L. Yarbus, "Eye Movements During Perception of Complex Objects," in *Eye Movements and Vision*, A. L. Yarbus, Ed. Boston, MA: Springer US, 1967, pp. 171–211.

[12] A. Klami, "Inferring task-relevant image regions from gaze data," in *2010 IEEE International Workshop on Machine Learning for Signal Processing*, 2010, pp. 101–106.

[13] R. L. Fantz, "Visual experience in infants: Decreased attention to familiar patterns relative to novel ones," *Science*, vol. 146, no. 3644, pp. 668–670, Oct. 1964.

[14] J. Driver, G. Davis, P. Kidd, E. Maxwell, P. Ricciardelli, and S. Baron-Cohen, "Gaze Perception Triggers Reflexive Visuospatial Orienting," *Visual Cognition*, vol. 6, no. 5, pp. 509–540, Oct. 1999.

[15] A. Freire, M. Eskritt, and K. Lee, "Are Eyes Windows to a Deceiver's Soul? Children's Use of Another's Eye Gaze Cues in a Deceptive Situation," *Dev Psychol*, vol. 40, no. 6, pp. 1093–1104, Nov. 2004.

[16] T. Chuk, A. B. Chan, S. Shimojo, and J. H. Hsiao, "Mind reading: Discovering individual preferences from eye movements using switching hidden Markov models," in *Proceedings of the 38th Annual Conference of the Cognitive Science Society, CogSci 2016*, 2016, pp. 182–187.

[17] S. Shimojo, C. Simion, E. Shimojo, and C. Scheier, "Gaze bias both reflects and influences preference," *Nat Neurosci*, vol. 6, no. 12, pp. 1317–1322, Dec. 2003.

[18] S. Einav and B. M. Hood, "Children's use of the temporal dimension of gaze for inferring preference," *Dev Psychol*, vol. 42, no. 1, pp. 142–152, Jan. 2006.

[19] K. Lee, M. Eskritt, L. A. Symons, and D. Muir, "Children's use of triadic eye gaze information for 'mind reading,'" *Dev Psychol*, vol. 34, no. 3, pp. 525–539, May 1998.

[20] S. Baron- Cohen, R. Campbell, A. Karmiloff- Smith, J. Grant, and J. Walker, "Are children with autism blind to the mentalistic significance of the eyes?," *British Journal of Developmental Psychology*, vol. 13, no. 4, pp. 379–398, 1995.

[21] W. H. Sumby and I. Pollack, "Visual Contribution to Speech Intelligibility in Noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, p. 212, Jun. 2005.

[22] H. McGurk and J. Macdonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, Dec. 1976.

[23] E. Krahmer, Z. Ruttkay, M. Swerts, and W. Wesselink, "Perceptual evaluation of audiovisual cues for prominence," in *INTERSPEECH*, 2002.

[24] M. Swerts and E. Krahmer, "Facial expression and prosodic prominence: Effects of modality and facial area," *Journal of Phonetics*, vol. 36, no. 2, pp. 219–238, Apr. 2008.

[25] J. Roy, J. Cole, and T. Mahrt, "Individual differences and patterns of convergence in prosody perception," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 8, no. 1, p. 22, Sep. 2017.

[26] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *JLP*, vol. 110, pp. 425–452, Jan. 2010.

[27] B. Rossion and G. Pourtois, "Revisiting Snodgrass and Vanderwart's object pictorial set: the role of surface detail in basic-level object recognition," *Perception*, vol. 33, no. 2, pp. 217–236, 2004.

[28] F. Cangemi, *mausmooth [Praat script]*. 2015. Retrieved from http://ifl.phil-fak.uni-koeln.de/sites/linguistik/Phonetik/mitarbeiterdateien/fcangemi/mausmooth.praat

[29] M. Winn, *Fade in, Fade out [Praat script]*. 2014. Retrieved from www.mattwinn.com/praat/RampOnsetAndOrOffset.txt

[30] N. Emery, "The Eyes Have It: The Neuroethology, Function and Evolution of Social Gaze," *Neuroscience and biobehavioral reviews*, vol. 24, pp. 581–604, Sep. 2000.

[31] N. Pfeiffer-Lessmann, T. Pfeiffer, and I. Wachsmuth, "An Operational Model of Joint Attention - Timing of Gaze Patterns in Interactions between Humans and a Virtual Human," in *Proceedings of the 34th annual conference of the Cognitive Science Society*, 2012, pp. 851–856.

[32] H. Eckert, "Erzeugung von Blickreizen virtueller Charaktere mit ambiger kommunikativer Absicht mittels systematischer Variierung zweier Faktoren einer Blickbewegung - Anfangsblick und Blickziel," University of Cologne, 2017.

[33] G. Van Rossum, *Python tutorial. Technical Report CS-R9526.* Amsterdam: Centrum voor Wiskunde en Informatica, 1995.

[34] FFmpeg Developers, *FFmpeg Tool [Software]*. 2018.

[35] D. J. Leiner, *SoSci Survey (Version 3.2.01-i) [Computer Software]*. 2018.

[36] R Core Team, *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2019.

[37] RStudio Team, *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc., 2016.

[38] P.-C. Bürkner, "brms: An R Package for Bayesian Multilevel Models Using Stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, Aug. 2017.

[39] M. Brysbaert, M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte, and A. Böhl, "The Word Frequency Effect," *Experimental Psychology*, vol. 58, no. 5, pp. 412–424, Jan. 2011.

[40] D. Müllner, "fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python," *Journal of Statistical Software*, vol. 53, no. 1, pp. 1–18, May 2013.

[41] A. Bayliss, G. Di Pellegrino, and S. Tipper, "Sex differences in eye gaze and symbolic cuing of attention," *The Quarterly journal of experimental psychology. A, Human experimental psychology*, vol. 58, pp. 631–50, Jun. 2005.

[42] V. Hazan, O. Tuomainen, J. Kim, and C. Davis, "The effect of visual cues on speech characteristics of older and younger adults in an interactive task," in *Proceedings of the 19th International Congress of the Phonetic Sciences*, 2019, pp. 815–819.

[43] O. Niebuhr, M. D'Imperio, B. Gili Fivela, and F. Cangemi, "Are There 'Shapers' and 'Aligners'? Individual Differences in Signalling Pitch Accent Category," 2011, pp. 17–21.

[44] P. Wagner, A. Ćwiek, and B. Samlowski, "Exploiting the speech-gesture link to capture fine-grained prosodic prominence impressions and listening strategies," *Journal of Phonetics*, Jul. 2019.