On speech-hand synchrony during conversations in a virtual underground station

Ľuboš Hládek, Bernhard U. Seeber

Audio Information Processing, Technical University of Munich, 80333 Munich
E-mail:ge67kip@mytum.de

Introduction

When people communicate in acoustically adverse environments, acoustic speech signals degrade due to noise or reverberation, making spoken communication difficult. However, the human auditory system exploits linguistic cues, cognitive ability, prosody cues, or contextual cues, to help to resolve the missing acoustic information. Nevertheless, in many cases purely auditory mechanisms are insufficient, and the listener needs to rely on additional visual information, for instance from lip movements of the collocutors [1]. While the visual prosody of lip reading and head movements significantly contributes to speech understanding [2], [3], the relationship between hand movements and speech perception, and between hand movements and speech production is less understood.

The hand gestures carry complementary or redundant information to speech production [4]. For instance, in gesticulation, when beat gestures accompany speech production, the hand movements do not add specific meaning to the speech. However, these beat movements may stress the essential parts of the speech, such as the stress of the word given by the speech prosody, which is vital for the phrase's meaning. On the other hand, iconic gestures (or deictic or metaphoric) carry additional information, and therefore, the cognitive process is different than in the case of beat gestures. Previous work pointed out that gesticulation and speech production might be governed by the same or coupled cognitive processes evidenced by the decoupling of the hand-gesture synchrony in cognitively demanding tasks or acoustically adverse conditions [5], [6]. Thus, in the current work, we hypothesize that when people have a free, unstructured conversation in a noisy environment, speech-hand gesture synchrony will be affected when the communication becomes more effortful due to the increased background noise.

In a previous experiment with a communication activity [7], the total duration of hand gesticulation, or the number of sub-movements, was increased at higher noise levels. Nevertheless, peak velocity, vertical amplitude, or gesture size were not positively correlated with the noise level. However, the situation might be different when people perform a free, unobstructed conversation in an environment with precise control of the visual and acoustic properties. Hence, in the present experiment, we re-examined the movement kinematics and hand-speech synchrony of two freely conversing participants in an acoustically controlled environment in immersive virtual reality. We expected that synchrony and gesture kinematics are influenced when communication becomes more challenging due to increased background noise.

Methods

Three pairs of people without known hearing problems and with pure-tone thresholds in the range of normal hearing took part in the *Conversation in Noise* experiment. Each pair consisted of one male and one female. The participants were students or colleagues from the Technical University of Munich, whose English was their second language. Each participant provided written informed consent, and the procedures were approved by the Ethical Committee of the Technical University of Munich (65/18S).

The task of the experiment was to have free, unscripted conversation in the English language for 30 minutes in the free-field virtual audio-visual simulation of the underground station [8]. Short breaks split the 30 minutes into six blocks with varying acoustic conditions: No Noise, 70 dB Noise, 80 dB Noise. The noise was amplitude-modulated Fastl noise [9] coming from four positions distributed around the participants at azimuths: 0°, 90°, -90°, 180°. Each of the three conditions was repeated twice, the condition order was pseudorandomized, and the condition was fixed within the 5-minute block. The noise sources were generated independently; the noise level was calibrated relative to the center of the loudspeaker array to a sample signal without reverberation. Reverberation added 2.3 dB; thus, the presentation levels of the noise conditions were 72.3 dB SPL and 82.3 dB SPL.

Figure 1: Two participants are conversing in the virtual underground station, simulated using rtSOFE inside an anechoic chamber.

The simulation was done using the real-time Simulated Open Field Environment [10], [11], which is a comprehensive set of freely available tools for creating acoustic simulations, which was interfaced with the Unreal Engine (v 5.2). The acoustic stimuli were delivered over 61 calibrated loudspeakers surrounding the space (4.2 m x 4.2 m) where the participants could move freely. The visual surroundings of the underground station were projected on four screens encompassing the experimental space using four low-noise projectors (Figure 1). The participants had headset

microphones calibrated before the experiment for live-reverberating and recording their speech. Participants wore motion-tracking suits for recording their body motion with a high spatial and temporal accuracy (360 Hz). The specialized hardware synchronized the motion tracking and the sound presentation system using the word-clock signal of the sound card. The voices of the participants and the noise included reverberation (T_{60} = 1.68 s, DRR = 2.7 dB, EDT = 0.31 s), which was computed in real-time based on the position of the participants within the acoustic model of the underground station.

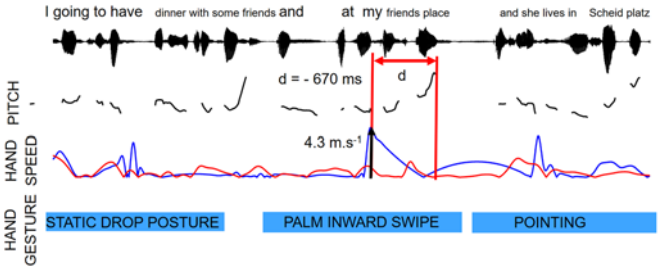

Figure 2: A sample of data from one participant's microphone and motion tracking and the analyzed statistics. The top row shows the transcript of the audio track. The pitch data were extracted using Praat [12]. The hand speed shows data from the left (blue) and right (red) hand (10 Hz low-pass filtered). The bottom panel shows an example of gesture categorization. The red arrow shows delay d, and the black arrow indicates the peak speed of the left hand.

The first step of the analysis was to manually annotate [13] hand movements into categories describing different types of hand movements; the categories were defined in the previous work. The speech-hand synchrony was analyzed in terms of the hand-speech delay

$$d = s_{max} - F0_{max} \quad \text{[ms]} \quad (1)$$

where $F0_{max}$ corresponds to the time of peak pitch, which was a simple maximum of the F0 trace within the gesture duration, and s_{max} corresponds to the time of peak speed, which was defined as the time of the highest non-directional speed, either from the left hand or from the right hand within 1 second relative to the peak pitch and gesture duration. Thus, a negative value indicates that the gesture peaked before the pitch peak (Figure 2). Gestures within the first 10 seconds of the block were not considered for the analysis. The statistics were pooled across all hand gestures within the respective conditions; for each participant, this period was 10 minutes of conversation.

Results

Figure 3 shows that the statistics d tends to increase with increasing noise level, but given the size of the error bars, the increase is evident mainly between the No Noise and 80 dB condition by 81 milliseconds. The analysis of the median absolute deviation (MAD) (Figure 3B), which reflects the spread of the distribution of the d statistics, is rather constant across conditions. Therefore, the distribution of d statistics shifts towards more negative values but does not increase variance.

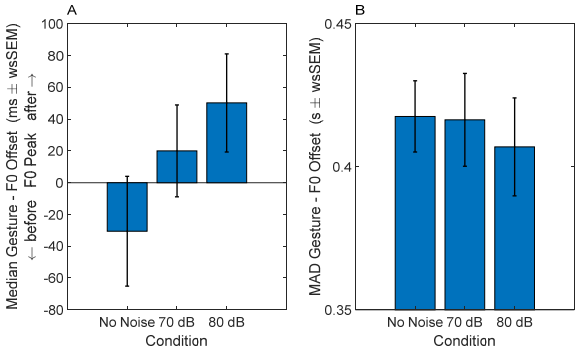

Figure 3: (A) Across-subject mean of medians of d (B) across-subject mean of median absolute deviation (MAD) of the statistics d. Error bars show within-subject SEM.

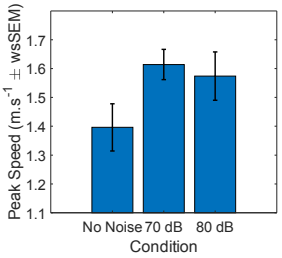

Figure 4: Across-subject mean of peak speed of hand gestures during conversation.

The average peak speed (Figure 4) has a magnitude of 1.39 $m.s^{-1}$ in the No Noise condition and tends to increase by approximately 0.19 $m.s^{-1}$ from No Noise to 70 dB condition but does not increase further in the 80 dB condition.

Discussion

Here, we present a preliminary analysis of data from three participant pairs of the ongoing study. The results suggest that hand-speech synchrony in free, unscripted conversation of two standing participants is sensitive to increasing background noise levels. The observed shift of the statistics d towards the positive values means that the gestures have a higher tendency to start after the pitch maximum in the 80 dB condition than in the No Noise condition. The shift may reflect the speakers sensitivity for increased cognitive demands of the listener for communication in a noisy environment. However, it does not indicate the loss of synchrony, which would be visible as an increase in variance of the distributions. Thus, the shift towards more positive values may reflect a processing delay. It is difficult to compare these results with the previous experiments [5], [6] due to the methodological differences; however, the general assumption that the statistics d would be sensitive to cognitively demanding tasks was confirmed.

The increase in gesture peak velocity was not observed in Trujillo et al. (2021) [7], which can be attributed to the methodological differences between the two studies. While in the current study, participants had a free, unobstructed conversation in clearly defined acoustic conditions, the levels in the previous study varied for each participant, and the task included various restrictions to the movement. The increase in gesture speed thus indicates that the gestures became more pronounced in higher noise, meaning that

either the gesturing became qualitatively different or that people used different types of gestures, which are faster, which should be further analyzed in the following steps.

In the current analysis, the gestures were analyzed across all possible hand gesture categories irrespective of whether the gestures were beat gestures or iconic gestures, though most of the gestures were beat gestures based on preliminary data observations. However, previous literature suggested that different gestures may react differently to increasing cognitive demands. Additionally, the current analysis pools across male and female participants due to low sample size, but future analysis should investigate possible differences.

References

[1] L. V. Hadley, W. O. Brimijoin, and W. M. Whitmer, "Speech, movement, and gaze behaviours during dyadic conversation in noise," *Sci. Rep.*, vol. 9, no. 1, p. 10451, Dec. 2019.

[2] A. MacLeod and Q. Summerfield, "Quantifying the contribution of vision to speech perception in noise.," *Br. J. Audiol.*, vol. 21, no. 2, pp. 131–41, May 1987.

[3] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual Prosody and Speech Intelligibility," *Psychol. Sci.*, vol. 15, no. 2, pp. 133–137, Feb. 2004.

[4] K. Bergmann, V. Aksu, and S. Kopp, "The relation of speech and gestures: Temporal synchrony follows semantic synchrony," in *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011)*, 2011.

[5] L. De Jonge□Hoekstra, R. F. A. Cox, S. Van der Steen, and J. A. Dixon, "Easier Said Than Done? Task Difficulty's Influence on Temporal Alignment, Semantic Similarity, and Complexity Matching Between Gestures and Speech," *Cogn. Sci.*, vol. 45, no. 6, Jun. 2021.

[6] W. Pouw and J. A. Dixon, "Entrainment and Modulation of Gesture–Speech Synchrony Under Delayed Auditory Feedback," *Cogn. Sci.*, vol. 43, no. 3, Mar. 2019.

[7] J. P. Trujillo and J. Holler, "The Kinematics of Social Action: Visual Signals Provide Cues for What Interlocutors Do in Conversation," *Brain Sci.*, vol. 11, no. 8, p. 996, Jul. 2021.

[8] Ľ. Hládek, S. D. Ewert, and B. U. Seeber, "Communication Conditions in Virtual Acoustic Scenes in an Underground Station," in *2021 Immersive and 3D Audio: From Architecture to Automotive, I3DA 2021*, 2021, pp. 1–8.

[9] H. Fastl, "A background noise for speech audiometry," *Audiol. Acoust.*, no. 26, pp. 2–13, 1987.

[10] B. U. Seeber and T. Wang, "real-time Simulated Open Field Environment (rtSOFE) software package." Zenodo, 2021.

[11] B. U. Seeber, S. Kerber, and E. R. Hafter, "A system to simulate and reproduce audio–visual environments for spatial hearing research," *Hear. Res.*, vol. 260, no. 1–2, pp. 1–10, Feb. 2010.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2024.

[13] "ELAN." Max Planck Institute for Psycholinguistics, The Language Archive., Nijmegen.