TEMPORAL ASPECTS OF TURN-TAKING IN ZOOM CONVERSATIONS

Qiang Xia

Humboldt-Universität zu Berlin
qiang.xia.1@hu-berlin.de

ABSTRACT

The proper timing is crucial for turn transitions. Turn-taking in conversations usually takes place without noticeable gaps or overlaps longer than 200 ms [1, 2]. With the widespread adoption of video-mediated communication during the COVID-19 pandemic, there is a growing need to understand how turn-taking behavior differs in remote situations. The current study compares several temporal aspects of turn-taking in face-to-face and Zoom interactions. Spontaneous dialogues in the Berlin Dialogue Corpus [3] were investigated. Twenty native German speakers conversed in pairs to complete two spot-the-difference tasks in the respective situations. Interlocutors tend to speak more slowly and produce longer gaps and longer overlaps between turns over Zoom. In face-to-face interactions, speaker change happens more frequently, leading to slightly more overlaps. This implies that dialogues over Zoom follow a distinct temporal pattern from those in co-present situations. A threshold higher than 200 ms is necessary to describe the temporal pattern of turn-taking in Zoom conversations.

Keywords: Zoom-conversation, turn-taking, speech corpus, spontaneous speech.

1. INTRODUCTION

Since Sacks, Schegloff and Jefferson [1] first focused on the "speech exchange systems" and concluded that conversations overwhelmingly follow the "one-speaker-at-a-time" pattern where turn transitions with no gap and no overlap are quite common, turn-taking behaviour has been extensively studied for almost fifty years. Empirical studies confirmed that the timing of turn-taking is highly precise, but not as seamless as "zero-gap-zero-overlap" [4, 5]. Transitions with acoustic silences of about 200 ms compose the majority of turn switches. Actually, gap durations within a range of 250 ms are examined to be cross-linguistically valid [6, 7]. Wilson and Wilson [2] explained that smooth transitions are realised by the synchronisation of neural oscillators based on speaker's syllable rate, about 200 ms in informal English dialogues. Thus, listeners' readiness to speak usually counterphases with that of the speaker.

However, it is noteworthy that face-to-face interaction has been the default conversational setting in the body of literature. Still relatively little is known about how turn-taking behaviour is organised in video-mediated conversations, though the COVID-19 pandemic has immensely increased our use of videoconferencing programs such as Zoom in the last few years.

Zoom conversations differ from face-to-face interactions in several ways. Some important paralinguistic cues, for instance, gaze direction, that can facilitate anticipating turn-ends [8], are hardly available in remote conversations [9]. In fact, it has been demonstrated that conversational rhythm is disrupted in Zoom interactions [10], partly because electronic transmission takes time and unavoidable delays of about 30–70 ms are long enough to disturb neutral oscillators that synchronise during conversation [2, 10]. They reported massive differences of turn transition time in face-to-face (135 ms) and Zoom interactions (487 ms). Given the differences between the two conversational situations, Zoom interactions are assumed to include longer gaps. The 200 ms threshold commonly used to describe the turn-taking rhythm in face-to-face conversations may not be suitable for Zoom conversations.

This paper aims to investigate how turn-taking behaviour differs between face-to-face and Zoom interactions in terms of temporal aspects, specifically identifying the threshold that describes the majority of turn transition in Zoom interactions.

2. METHODS

The Berlin Dialogue Corpus [3] was investigated in the study. Twenty native German speakers (mean age = 25.7, SD = 3.8, 10 females, 10 males) who knew each other well prior to the experiment were asked to finish four spot-the-difference tasks (two over Zoom and two in face-to-face situation) in pairs. In the so-called Diapix tasks [11, 12, 13], participants needed to cooperate and identify 13 differences between two similar pictures within 10 minutes. In the co-present situation, participants conversed in a phonetic laboratory. For the Zoom conversation, participants were sitting in different rooms in the same university building and connected via Zoom. To avoid differences arising from recording methods, conversations were recorded by separate microphones in both situations. Zoom acknowledges that there can be latency from a few milliseconds to a few seconds during Zoom calls. Yet, the delay durations vary with various factors and can hardly be specified [14]. Using separate recordings can include the latancies perceived by both sides. Recordings from the separate channels were merged together to reconstruct the dialogue. The collected data and its orthographic transcription were first aligned by WebMaus [15]. Turn-taking relevant elements like speech turns were annotated in Praat [16]. Interval boundaries have been manually corrected. After annotation, the automatic detection tool [17] was used to identify empty and overlapping annotation chunks. In this way, gaps and overlaps between speech turns were created automatically based on the annotations aligned with speech signals. In the end, the data were transformed into an emuR-database [18] for temporal analysis.

The study focused only on between-speaker silences, specifically gaps between speech turns, while silences within a speech turn were not considered. Gaps and overlaps are often treated as two ends of a continuum: when measuring the interval between the end of the first speaker's speech and the starting point of the second speaker's speech, gaps have positive values and overlaps negative [6, 5, 19].

To compare speech rate in the two situations, syllable durations were calculated by subtracting pauses within a turn from the length of the turn before dividing it by the number of syllables produced within it. Syllable were counted by using the R-package SYLLY.DE [20].

In order to investigate the correlation between the position of gaps or overlaps and the task-oriented conversation timeline, the starting point of a gap or overlap was normalised by dividing it by the entire conversation length and rounded to two decimal place. The study excluded the preparing and ending phase of a conversation, where speakers were trying to be heard or finish the recording, corresponding to the time point beyond the range from 0.05 to 0.95. For each normalised time slot, gap and overlap durations were averaged.

3. RESULTS

Extra long gaps are observed at the beginning and near the end of conversations, because participants were still trying to adjust themselves to the new conversational settings or they had difficulties finding more differences. To eliminate the influence of these extra-linguistic factors, gaps that exceeded the upper bound of the data (Q3+1.5IQR, calculated as the upper quartile of the data plus 1.5 times the interquartile range) were removed. Similarly, outliers of overlaps (longer than the lower bound, Q1-1.5IQR) were not taken into account. Previous studies [5, 21] have shown that geometric means provide a more realistic estimate of central tendency than arithmetic means. Therefore, geometric means are presented in the figures.

3.1. Syllable durations

On average, it takes about 217 ms for speakers to produce a syllable over Zoom, significantly longer than in face-to-face interactions, about 206 ms per syllable (t = -5.97, DF = 39, p < .001). In other words, speakers tend to speak more slowly when they talk over Zoom; as illustrated in Fig. 1.

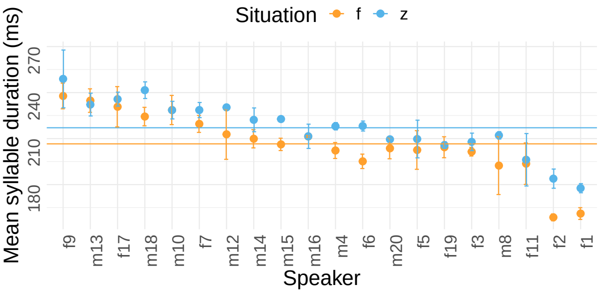

Figure 1: Mean syllable durations in face-to-face (f) and Zoom (f) situations by speaker with error bars. The horizontal lines show the respective mean syllable durations in the two situations.

3.2. Gap durations

As can be seen from Fig. 2, gaps in face-to-face conversations (shown in yellow) are generally shorter than those in Zoom situations. In the co-present scenario, around 50% of gap durations were within the 270 ms range, whereas the median value in Zoom conversations was considerably longer, at approximately 438 ms. To examine the differences further, a linear mixed effect model was computed: using situation and the normalised time point of a gap in the dialogue as the mixed effect, gap duration as the dependent variable and conversation as the random intercept. The results revealed that both situation and normalised time point significantly impact gap durations. Specifically, the Zoom situation increased the duration of gaps. Gaps are expected to be longer nearer to the end of a conversation. The correlations are additionally summarised in Fig. 3.

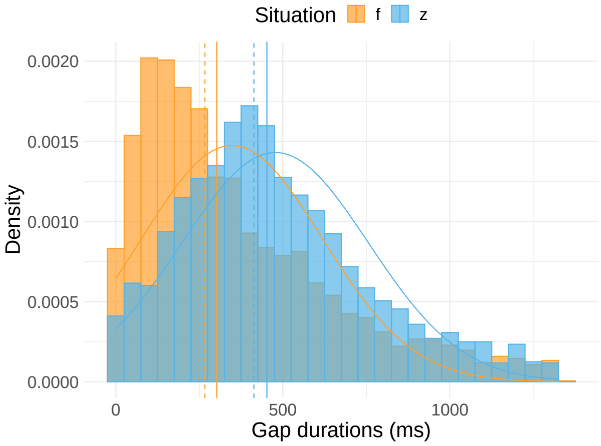

Figure 2: Histogram of gap durations in face-to-face (f) and Zoom (z) situations with the estimated distributions. Bin size 50 ms. Outliers excluded. Dashed lines show the geometric means, solid lines the medians.

Situation	Face-to-face	Zoom
Arithmetic mean	348.04	477.08
Geometric mean	231.60	365.76
Median	269.56	437.49
Mode	152.38	402.38
Standard deviation	270.52	278.80
Without outliers	3088	2722
Total N	3317	2936
Percentage	93.10 %	92.71%

Table 1: Descriptive statistics of gap durations in face-to-face and Zoom situations (in ms).

3.3. Overlap durations and occurrences

Fig. 4 presents that Zoom conversations exhibit longer overlaps compared to face-to-face situation. The median value of overlap durations in Zoom interactions is 416 ms, about 100 ms longer than the 319 ms observed in face-to-face situation; see Tab. 2. A mixed effect linear model was used to determine the effects of situation and normalised time point (mixed effect) on mean overlap durations (dependent variable) with conversation as a random effect. The results showed that only situation significantly affects overlap durations, and there is no correlation between the normalised time point of an overlap and its duration (see Fig. 5).

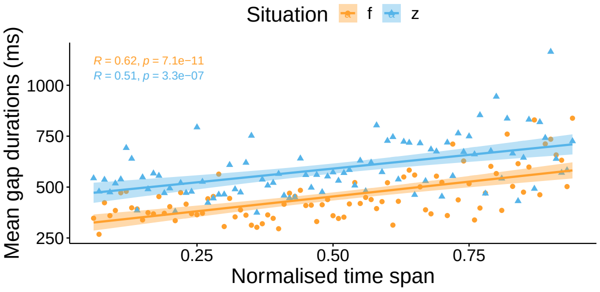

Figure 3: Linear regression between mean gap durations and normalised time point (range: 0.05-0.95) in conversations.

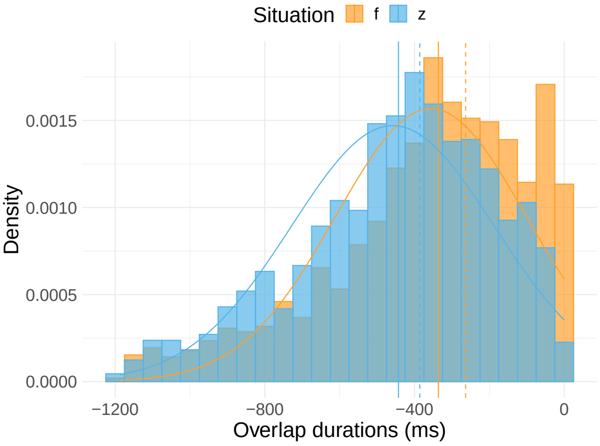

Figure 4: Histogram of overlap durations in face-to-face (f) and Zoom (z) situations with the estimated distributions. Bin size 50 ms. Outliers excluded. Dashed lines show the geometric means, solid lines the medians.

Situation	Face-to-face	Zoom
Arithmetic mean	-356.19	-457.95
Geometric mean	-238.27	-359.05
Median	-319.41	-416.35
Mode	-247.62	-397.62
Standard deviation	254.56	271.35
Without outliers	1935	1790
Total N	2047	1880
Percentage	94.53%	95.21%

Table 2: Descriptive statistics of overlap durations in face-to-face and Zoom situations (in ms).

Although more overlaps are observed in Zoom interactions, the difference is not significant at the 5% level (p = 0.23). Interlocutors switch more

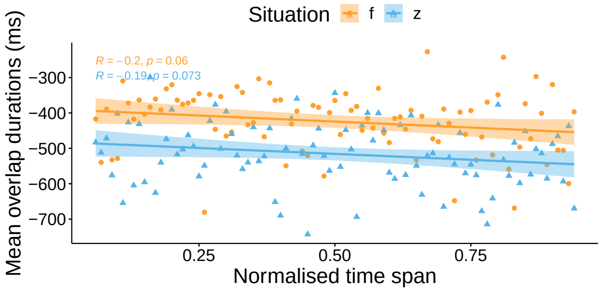

Figure 5: Linear regression between mean overlap durations and normalised time point (range: 0.05-0.95) in conversations.

frequently the speakership when talking face-to-face, approximately 18 times per minute; while in Zoom interactions, the turn transition frequency is about 15 times per minute ($t = 4.25, p < .05$). Yet, the overlap-to-transition ratio in co-present situation is smaller (0.63) than that in Zoom interactions (0.66). However, the difference was not significant.

4. DISCUSSION

It is common to have slight gaps and overlaps during turn-taking in conversations [1, 2, 19], but previous studies have suggested different interval durations for the *zero-gap-zero-overlap* pattern is not realistic [4], such as 120 ms in [5], 200 ms in [2, 22, 19]. The current study found that for face-to-face interactions, a threshold of up to 300 ms could describe the common cases of turn transition intervals, while for Zoom conversations, the threshold was 450 ms.

The longer gaps in face-to-face interaction might be due to the high cognitive loads resulting from the experimental settings. Participants were asked to cooperate to spot all differences between two pictures, which might be cause high cognitive loads while conversing [19]. However, studies suggesting 200 ms as the frequent threshold are primarily based on polar questions [2, 6]. Considering the differences in cognitive loads resulting from experiment settings, the longer gaps found here should come as no wonder.

In Zoom conversations, both gaps and overlaps are longer than those in the co-present situation. Unavoidable transmission delays can explain this phenomenon to some extend. To record conversations from a holistic view, recordings were made separately from both sides of a dialogue. By this means, actual latencies perceived by the two sides were included in the recording. Speakers would stop talking if overlap occurs, so that the *one-speaker-at-a-time* pattern can soon be restored [1]. If speech signal is received with latency, reaction on the signal will also be delayed. In this case, speakers would not be able to relinquish their speech turn *on time*. As a result, overlap continues a bit till it is perceived, causing longer durations. Unfortunately, statistics on electronic transmission delays over Zoom are not available to the public because of commercial interests and legal limitations. For this reason, the true proportion of technical latency in gap and overlap durations can hardly be removed from the current analysis. Even if the assumed 30-70 ms transmission delays [23] were subtracted from the gap and overlap durations in the Zoom data, the thresholds will still be longer than those in face-to-face situations.

Given transmission delays over Zoom and their interference in conversational rhythm, more overlaps had been expected in remote communication. Also [24] confirmed that latencies up to 800 ms increase the number of unintended interruptions. On the contrary, slightly more overlaps were found in face-to-face conversations, which can be traced back to the higher frequency of speaker changes in this situation. Edelsky [25] demonstrated that overlaps can reflect high interactivity and high engagement in a conversation. From this point of view, it can be assumed that participants were more engaged when talking face-to-face, hence overlapped more.

In addition, interlocutors spoke with a lower syllable rate when talking over Zoom. The reason for this difference might be the uncertainty caused by the omnipresent transmission delay. Though no obvious latency was reported during the experiment, speakers could probably still sense the brief delay, which made them unsure if their conversational partner could perceive their speech punctually and completely. It could also explain the the longer gaps in Zoom conversations since the speech rhythm of speakers will synchronise during a conversation [2].

5. CONCLUSION

This study aimed to compare temporal aspects of turn-taking behavior in face-to-face and Zoom conversations. Zoom interactions are found to have longer gaps and overlaps. However, face-to-face interactions had more frequent speaker changes and slightly more overlaps. The impact of transmission delay on turn-taking timing in video-based conversations needs to be explored further, including measuring the exact duration of latency in signal transmission and its interaction with the temporal pattern of turn-takings.

6. ACKNOWLEDGEMENTS

The author would like to thank Christine Mooshammer for valuable input and discussions. And a lot of thanks to Malte Belz for guidance in corpus annotation and helpful comments. Also thanks to reviewers for great suggestions. The work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334.

7. REFERENCES

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[2] M. Wilson and T. P. Wilson, "An oscillator model of the timing of turn-taking," *Psychonomic Bulletin & Review*, vol. 12, no. 6, pp. 957–968, 2005.

[3] M. Belz, A. Zöllner, M. Terada, R. Lange, L. S. Adam, and B. Sell, "Dokumentation und Annotationsrichtlinien für das Korpus BeDiaCo," 2021.

[4] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[5] M. Heldner, "Detection thresholds for gaps, overlaps, and no-gap-no-overlaps," *The Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. 508–513, 2011.

[6] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K. E. Yoon, and S. C. Levinson, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, 2009.

[7] K. Weilhammer and S. Rabold, "Durational aspects in turn taking," in *Proc. 15th ICPhS*, Barcelona, 2003, pp. 2145–2148.

[8] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[9] J. N. Bailenson, "Nonverbal overload: A theoretical argument for the causes of Zoom fatigue," *Technology, Mind, and Behavior*, vol. 2, no. 1, 2021.

[10] J. E. Boland, P. Fonseca, I. Mermelstein, and M. Williamson, "Zoom disrupts the rhythm of conversation," *Journal of Experimental Psychology: General*, pp. 1272–1282, 2021.

[11] K. J. Van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow, "The wildcat corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles," *Language and Speech*, vol. 53, no. 4, pp. 510–540, 2010.

[12] R. Baker and V. Hazan, "DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs," *Behavior Research Methods*, vol. 43, no. 3, pp. 761–770, 2011.

[13] O. M. Bullock and B. Sell, "PDF and PSD files of DiapixGEtv picture materials – German version adapted to elicit tense vowels," 2022.

[14] Zoom, "Video conferencing, web conferencing, webinars, screen sharing," https://zoom.us/, 2023.

[15] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[16] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2022.

[17] T. Mahrt, "A python library for working with praat, textgrids, time aligned audio transcripts, and audio files: PraatIO," https://github.com/timmahrt/praatIO, 2016.

[18] R. Winkelmann, J. Harrington, and K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech & Language*, vol. 45, pp. 392–410, 2017.

[19] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in Psychology*, vol. 6, 2015.

[20] M. Michalke, "Sylly: Hyphenation and syllable counting for text analysis," https://reaktanz.de/R/pckg/sylly/citation.html, 2018.

[21] J. P. de Ruiter, H. Mitterer, and N. J. Enfield, "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation," *Language*, vol. 82, no. 3, pp. 515–535, 2006.

[22] M. B. Walker and C. Trimboli, "Smooth transitions in conversational interactions," *The Journal of Social Psychology*, vol. 117, pp. 305–306, 1982.

[23] J. E. Boland, "Conversation Transition Times: Working Memory & Conversational Alignment," in *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, A. K. Goel, C. M. Seifert, and C. Freksa, Eds. Cognitive Science Society, 2019, pp. 159–165.

[24] S. Egger-Lampl, R. Schatz, and S. Scherer, "It takes two to tango - Assessing the impact of delay on conversational interactivity on perceived speech quality," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, Interspeech 2010*, 2010, pp. 1321–1324.

[25] C. Edelsky, "Who's Got the Floor?" *Language in Society*, vol. 10, no. 3, pp. 383–421, 1981.