Alexa, you are too slow! Invariant turn-transition times and conversational flow in natural human – voice agent interaction

Mathias Barthel
Leibniz-Institute for the German Language
Mannheim, Germany
`barthel@ids-mannheim.de`

Abstract

A corpus of naturally occurring interactions of first-time users with their voice agent is analysed with respect to the voice agent's turn timing and its knock-on effects on the conversational flow and ensuing adaptations of users in their communicative behaviour. Across 1077 analysed turn transitions, turn timing by the voice agent has been found to be atypically slow and largely invariant, stripping human – voice agent interaction of a layer of naturalness. Users are found to quickly adapt to the observed characteristics of voice agents' communication, adjusting both their expectations about the voice agent's output as well as the timing of their own verbal contributions. The consequences for human – voice agent interaction are discussed, showing that major improvements in user experience hinge upon more naturalistic timing abilities in dialogue systems.

1 Introduction

In natural conversation, timing drastically matters. In human-human interaction, interlocutors regularly and swiftly switch the roles of speaker and listener, taking turns at talk that are commonly rather short (Sacks et al., 1974), often less than one second. Transitions between these turns are commonly very well aligned, with next speakers commonly starting their turns when the previous speaker ends their turn. The gap between two adjacent turns is often as short as 200 - 300 ms (Levinson and Torreira, 2015; Heldner and Edlund, 2010). When gaps between turns in focused conversation are generally much longer than that, the interaction will be regarded as problematically unsmooth and the interlocutors producing longer gaps can be perceived as less interested in the conversation and more cold and distant (Pearson et al., 2008), with interlocutors feeling less socially connected (Templeton et al., 2022). This impressive time management between interlocutors is the basis for inferential mechanisms that make the timing of a turn at talk relevant for how it is going to be interpreted, helping the conversational partners to understand their interlocutors' communicative intentions. For example, when one conversational partner makes an assessment like *"This dress is pure fire."* and gets an agreeing *"Yeah."* by their interlocutor after a gap of about 1.2 seconds, this second turn will be interpreted as less agreeing than when it is produced after a short gap of about 300 ms (Blohm and Barthel, 2024; Roberts et al., 2011). Moreover, the respondent will be perceived as more cold and distant when answering after a markedly long gap than when answering promptly (Blohm and Barthel, 2025). Early research in conversation analysis identified a 'standard maximum of unmarked silence' between adjacent turns of about one second before the gap becomes noticeable as being potentially meaningful (Jefferson, 1989). More fine-grained experimental research found a threshold for meaningful gaps to lie around 700 ms of silence between turns (Roberts and Francis, 2013; Henetz, 2017).

Next to their relevance for the interpretation of a speaking turn's meaning, long gaps can have further consequences for the ongoing conversation. For instance, when a first speaker makes an offer and does not get a response before a markedly long delay, they become more likely to re-select themselves again for the the next turn and rephrase or even withdraw their offer before they get a response (Davidson, 1984). Indeed, neuroimaging techniques revealed that a listener is more prepared for a dispreferred response, like a rejection of a request, when the answer is produced after a long gap of about one second than when it is produced after a short gap of about 300 ms. When a dispreferred response, like a refusal of an invitation or a decline of an offer, is given after a very short gap of 300 ms, the listener's EEG signal shows signatures of increased surprisal as compared to a swiftly produced preferred response. This difference in brain responses to a preferred versus a dispreferred response are completely neutralised in responses that come after a longer gap of about one second, showing that dispreferred responses are more expected after longer gaps (Bögels et al., 2015). Put differently, human interlocutors begin to expect a dispreferred response when they did not get a response quickly after the end of their own turn, quite possibly because dispreferred responses are more frequent after long gaps than preferred responses (Kendrick and Torreira, 2014).

These findings show that the timing of speaking turns, especially their delay, can trigger inferential processes that interlocutors generally avoid in conversation by very skilful coordination of their speaking turns. This coordination centrally rests on predictive processing, with interlocutors starting to plan their next contributions already while the current turn is still being produced by their conversational partner (Barthel et al., 2017; Barthel, 2020; Bögels, 2020). In order to be able to start to plan a next turn that is relevant to the current turn by your conversational partner, speakers anticipate the action that is being performed with the current turn very early on during the turn (Gisladottir et al., 2015), so that they can rely on the anticipated meaning of the incoming turn for their response planning (Barthel et al., 2016).

In human-machine interaction many of these fundamental characteristics of turn taking play out differently or are entirely absent (Skantze, 2021). One central observation that has been made is that the timing of turn taking is slow in human-machine interaction. Mostly due to the parsing mechanism that is commonly applied in virtual conversational agents (VAs), turn taking is found to be less dynamic than in human-human interaction. Especially when silence in the input of the automatic speech recognition process is the only cue to a VA that the speaking floor is open, unnaturally long gaps before the VAs' responses are unavoidable. In the absence of predictive processing in the VA, this suboptimal design feature seems to be a necessity, since otherwise intra-turn silences would regularly be mis-categorised as turn-end cues (Ten Bosch et al., 2005), leading to interruptions of unfinished turns by the users. However, if silence is the only cue that is used by the VA to decide whether a turn at talk is complete and requires a reaction, even a long threshold will still lead to occasional undesired interruptions.

As human-VA interactions are getting more prevalent in private settings, with designers advertising the communicative skills of their VA systems, analyses of human-VA interaction become more relevant and can inform due improvements in the available technology. Conversation Analytic research has started to contribute detailed qualitative analyses of case studies of human-VA interaction (Habscheid et al., 2023; Mlynář et al., 2025), focusing on features of the organisation of talk (Pitsch et al., 2009; Fischer et al., 2019) and on strategies of dealing with miscommunication (e.g. Krummheuer, 2008). Where VA systems still fall short of human-like communicative capabilities, users have been found to develop interactional strategies to adapt to the characteristics of the VA (Pelikan and Broth, 2016; Reeves et al., 2018; Porcheron et al., 2018; Barthel et al., 2023).

Many of these qualitative observations yet remain to be quantitatively checked for generalisability with ecologically valid data taken from intrinsically motivated interactions with the device. This paper presents turn timing data from naturally occurring human-VA interactions from private households from a collection of interactions of first time users during their first weeks of using a voice controlled device (Barthel et al., 2023). On the day participants installed their VA, an Amazon Alexa EchoDot, their households were equipped with a conditional voice recorder that recorded 3-minute audio files around command-response sequences (Porcheron et al., 2018), allowing for accurate measures of the timing of turn transitions in these sequences as well as for analyses of turn taking dynamics in interactions between users and their devices. The methods of data collection and analysis are briefly described in Section 2, followed by a presentation and discussion of the study's results in Section 3 before drawing general conclusions in Section 4.

2 Data and Methods

The conditional voice recorder (CVR) data were taken from the corpus presented by Barthel et al. (2023), comprising over 5.000 VA-directed commands recorded in 6 German-speaking households. Participants were recorded in their homes during the first seven to ten weeks of using the device. Using the CVR, recordings of 3-minutes in length around a command containing the wake-word 'Alexa' were obtained, making it possible to measure turn transition times between the commands and the device's responses as well as between the VA's responses and potential follow-up turns in third position by the user. In total, 945 relevant floor transfer offsets by 9 users from 6 households were measured manually in Audacity.[1] Floor transfer offsets were analysed in R (R Core Team, 2025) using the package *lme4* (Bates et al., 2015).

3 Results and Discussion

3.1 Transition times in transitions from user's turn to voice assistant system's turn

An intercept-only model with users as a random effect modelling floor transfer offsets in transitions with the VA taking the second turn shows that modelled average floor transfer offsets were 1366 ms (SE = 30 ms; see Fig. 1, top panel, showing the raw data). With floor transfer offsets in human-human conversation commonly being around 300 ms (Levinson and Torreira, 2015; Stivers et al., 2009; Heldner and Edlund, 2010), we find that VAs reactions are much slower in naturally occurring human-VA interaction than what humans are commonly capable of in conversation. Mean turn transition times produced by the VA are generally in the ball park of what would normally be meaningfully marked gaps in human-human interaction (Roberts and Francis, 2013). While such long gaps are often filled with placeholders like 'uh' or 'um' to indicate that the relevant turn is delayed but coming up (Clark and Fox Tree, 2002; Fox Tree, 2002), no such fillers are present in the analysed VA output. As laid out in Section 1, the slow responses by the VA can have a number of consequences for the unfolding interactions and for users' perception of their interaction with the VA. However, one particular consequence that would have been conceivable given the observations from human-human interactions is not generally present in the data set: Slow turn uptakes by the VA do not frequently cause users to re-select themselves for the next turn. On the contrary, self-selections for a second turn are very rare and appear only after very long gaps after user turns. Take excerpt (1) as an example.[2]

```
(1) [CVR03, day 2]

US: aLEXa:- (1.07) spiele riHANna,
    (1.12) DIAmonds.
    Alexa play Rihanna Diamonds.
    (8.70)
    spiele riHANna; (0.49) DIAmonds.
    (0.80) aLEXa? (1.49) SPIELe, (0.79)
    riHANna, (1.28) DIAmonds. (0.54) das
    LIED.
    play Rihanna Diamonds. Alexa play
    Rihanna Diamonds. the song.
    (6.16)
    halLO, aLEXa?
    (3.10)
    aLEXa:? (1.40) helLO;
    (5.50)
    aLEXa? (0.70) spiel mein HÖRbuch.
    Alexa play my audio book.
    (1.20)
VA: deine auswahl wird von audible
    ABgerufen.
    your selection is retrieved from
    Audible.
```

In this excerpt, the user wants the VA to play a specific song, but the VA does not react to the initial command. After a long 8.6 seconds, the user repeats the command after she realised the VA does not play the song. This repetition is produced without the mandatory command-initial wake word, which the user experienced before to fail to elicit reactions by the VA. Knowing that no reaction by the VA could be expected after this first repetition, the user repeats the command again after only 0.8 seconds, now with the wake word at the beginning of the turn. When this third attempt does not get a reaction, the user produces two turns that were to test whether the VA would react at all, both of which are produced after very long gaps of 6.16 and 3.10 seconds, respectively. After another long gap of 5.5 seconds in which the VA remains unresponsive, the user tries a test command that she knows certainly to have lead to a reaction from the VA in the past. This test command is responded to by the VA after a common gap period of 1.2 seconds. This excerpt illustrates that, at positions where VA reactions are reasonably expected (but not at other positions), users very quickly learn to be prepared for delays that are much longer than what is common in human-human interaction, with the threshold of what is treated as a markedly long gap being shifted beyond 3 seconds of silence, which is about double the attested modal response time of the VA.

---

[1] The CVR did not pick up all commands, as its speech detection model is inferior to the VA's (Porcheron et al., 2018, see also Barthel et al. (2023)). In 165 of the total of 1077 transition relevance places, the VA failed to react. In 45 cases, the floor transfer offset was not measurable.

[2] US = user. Transcripts were created based on GAT2 transcription conventions (Selting et al., 2011).

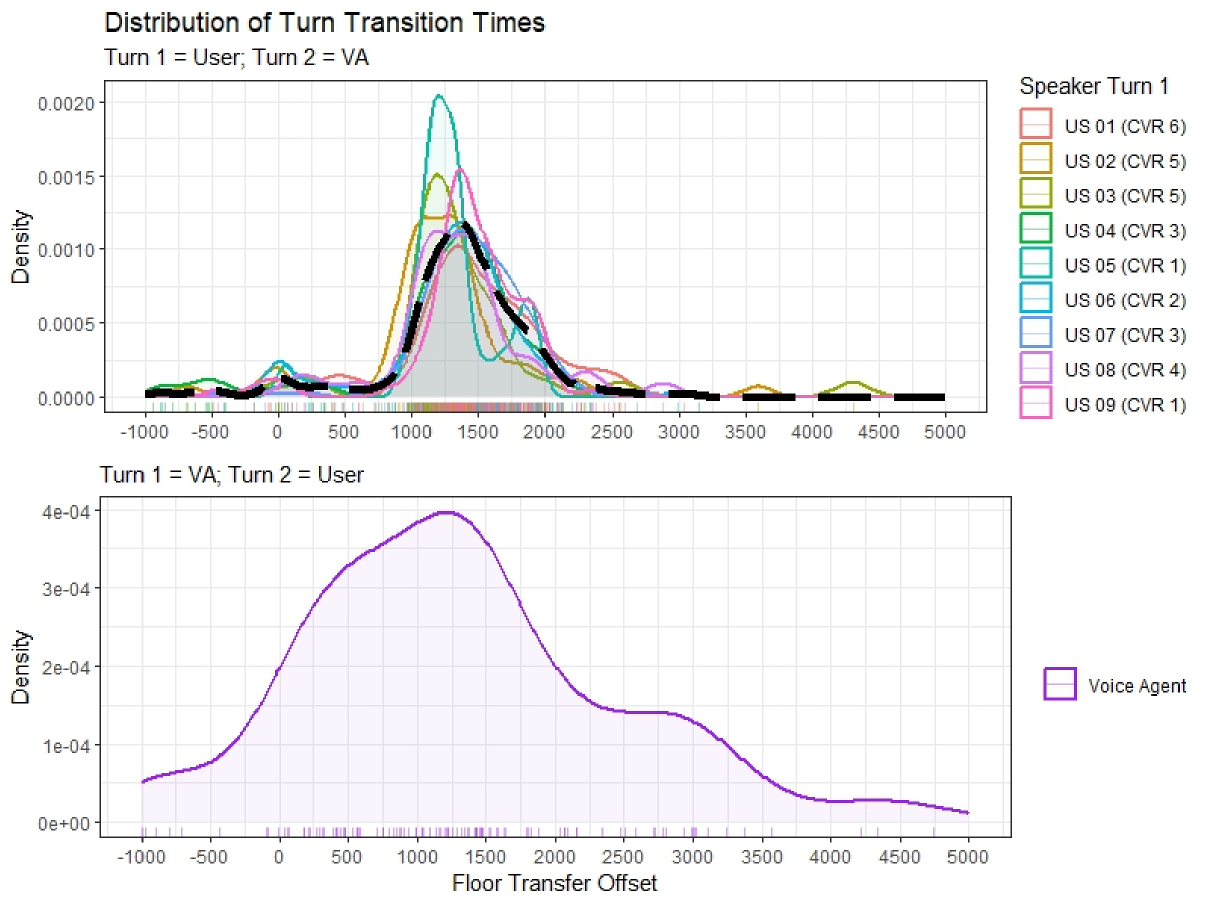

Figure 1: Density plots of floor transfer offsets between a first turn by the user (US) and a second turn by the VA (top panel; $N = 831$) versus a first turn by the VA and a second turn by the user (bottom panel; $N = 114$). Positive values indicate gaps between turns; negative values indicate overlaps between turns. The black, dashed curve indicates the global density distribution of all turn transitions with the VA taking turn 2 pooled together. Coloured tick marks mark the location of each measured turn transition.

3.2 Exceptions to the general slow-response pattern - Overlaps and ‘well-timed’ turn transitions

As can be seen in the top panel of Fig. 1, turn transitions with a very short gap between the user’s turn and the VA’s turn are quite rare in the presented data but do nonetheless exist. However, these rare cases are almost exclusively VA reactions to ‘stop’ commands that are reacted to by terminating the current output. And even in this very common type of command, reactions were only found to be fast when the VA was currently producing verbal output, in which case the command was met with immediate silence. In cases where the VA was currently producing any type of media output, like music or audio books, stop commands were reacted to as slowly as other types of commands. Hence, even though expectably fast reactions to commands do exist, they are not common when a verbal output by the VA is made relevant by the user’s command, as well-timed transitions from a user’s turn to a full verbal turn by the VA are not common in the data.

As can be seen in the top panel of Fig. 1 as well, overlaps of parts of the turn by the user and the turn by the VA are uncommon but do occur. In human-human conversation, overlap between turns is more common but mostly very short (Sacks et al., 1974). Prior work in conversation analysis found differential situations in which overlap occurs and how it is handled by interlocutors (Jefferson, 2004; Schegloff, 2000). Most overlap occurs when the second speaker starts their turn slightly early, but orients towards the transition relevance place at the end of the first speaker’s turn. This sort of overlap is commonly very brief and does not cause any trouble in the interaction, since the first speaker usually ends their turn shortly, naturally resolving the overlap. Another situation in which overlap commonly occurs is during backchannel responses that the second speaker produces during the first speaker's turn (Schegloff, 1982; Yngve, 1970; Gardner, 2001; Knudsen et al., 2020). These usually very short contributions do not claim the right to the speaking floor for a full turn and are used to signal attentiveness or to explicitly yield the rights to a next turn at a transition relevance place. Both of these types of overlap are cooperative in nature and are not commonly treated as problematic by interlocutors. In contrast, a third, comparatively rare kind of overlap is often categorised as 'interruption,' with the second speaker deliberately starting to speak while the first speaker is still mid-turn (Drew, 2009). Some of these mid-turn overlaps serve to initiate repair, for instance to clarify a reference or address problems of hearing (Kendrick, 2015; Dingemanse et al., 2015). In other cases, mid-turn overlaps are produced uncooperatively, deliberately disrupting the principled conversational flow following the one-speaker-at-a-time principle (Sacks et al., 1974). In these cases, interlocutors often adjust their volume and pitch to either compete for the floor or to yield it, indicating an orientation towards the presence of overlapping talk and acknowledging it as a marked interactional situation (Schegloff, 2000).

Notably, the 21 overlaps produced by the VA in the present data set diverge from the patterns observed in human-human interaction in a number of ways. Firstly, there are no overlaps caused by backchannelling by the VA; In fact, there is not a single backchannel turn by the VA system attested in the data. This lack of listener feedback is one reasonable candidate cause for why the interactions with the VA appear undynamic and dysfluent. We will return to this point in Section 3.4 below. Secondly, very few of the recorded overlapping turns by the VA initiated repair. On the contrary, most repair initiating turns by the VA were started only after long gaps of at least about one second, often considerably later. However, since all attested repair initiating turns by the VA were open format repair initiators, the timing of these repair initiating turns is partly comparable to human-human interactions, where open format repair initiation is commonly produced after a considerable gap (Kendrick, 2015). An example is given in excerpt (2), where the VA initiates repair in a very general manner after a long pause of 1.5 seconds, asking for a full repetition of the previous command without offering even a partial understanding of the command.

```
(2) [CVR03, day 8]
US: <<f>aLEXa?> (.) dU musst
    Alexa you must
    (1.50)
VA: ich habe den satz nicht vollständig
    verSTANden.
    I have not understood the sentence
    completely
    (0.64)
    FRAge [mich bitte noch EIn]mal.
    ask me again please
US:       [<<f> alExa?>       ] (1.53)
    <<f> aLEXa?> (1.73) durch die STADT.³
           Alexa (1.53) Alexa (1.73)
    through the city
```

Thirdly, all overlaps produced by the VA can be categorised as 'interrupting,' in the sense that all of them occur mid-turn, at a position where no immediately upcoming transition relevance place was predictably close in the first turn. Consider excerpt (3) as an example:

```
(3) [CVR05, day 3]
US: aLEXa? (0.7) spiel
    AUdible? (0.6) arabisch für
    ANfäng[er kapitel e-      ]
    Alexa (0.7) play Audible (0.6)
    Arabic for beginners chapter o-
VA:       [deine Auswahl wird] von
    audible ABgerufen.
    lernen Sie syrisch-arabisch zu
    SPRECHen (.) band eins wird
    fOrtgesetzt.
           your selection will be
    retrieved from Audible. learn to
    speak Syrian-Arabic (.) volume one
    will be continued
```

In this command, the user addresses the VA with the wake word and produces 'spiel audible' (play audible) with a high rising boundary tone, functioning as a turn-holding cue, indicating more talk to come (Duncan and Niederehe, 1974; Gravano and Hirschberg, 2011). The following 0.6 seconds of intra-turn silence probably triggered the VA system to analyse the command as complete. Even though the user continues her turn with more talk, specifying what she wants the VA to play, the VA's response, once triggered, starts in the middle of the user's turn. Even though the continuation of the turn by the user started long before the turn by the VA, which in human-human interaction orderly leads to the speaker of the first turn keeping the rights to the floor (Schegloff, 2000), the VA

---

[3] 'Durch die Stadt' is the title of a song, which the VA duly plays after the command, making the repair successful (see also Reineke and Helmer, 2024).

continues through with the production of its turn, leading the user to abandon her turn and aborting its production mid-syllable. What is noteworthy in this example, and characteristic for the other cases of overlapping talk by the VA in the data set, is that the long delay before responses by the VA leads to turn starts at unsystematic points during the ongoing turn by the user, often times at positions without any turn-final cues. For that reason, any overlapping talk by the VA is treated as competitive by the users, with a common follow-up in the interaction being a complete or partial repetition of the original command, as in excerpt (4).

(4) [CVR03, day 7]

US: aLEXa? (.) SPIE:L (1.06) unbreak my
[heart.]
Alexa play unbreak my heart
VA: [hier ist] ein RAdiosender, (.) der
dir viellEIcht
gefäll[t; (.) CHART]
here is a radio channel you might
like; chart-
US: [<<f>aLEXa,> SPIEL] unbreak my
heart.
Alexa play unbreak my heart
(1.53)
VA: unbrEAk my heart, radio edit von dee
jay dark von SPOTify,
unbreak my heart radio edit by DJ
dark on Spotify

In both excerpts (3) and (4), which are representative of the whole data set in this respect, the interrupting turns by the VA are in no way marked to be competitive, as they are produced with the same speech rate, pitch, and volume as any other turn by the VA in any other position. Thus, the VA does not portray any understanding of the competitiveness of the conversational situation that was created by the timing of its verbal output, which can be responded to by competition for the speaking floor by the user, as in excerpt (4), where the user repeats their initial command with an upstep in volume and in overlap with the VA.

3.3 Transition times in transitions from voice assistant system's turn to user's turn

While by far the more common adjacency pair in the presented data set is user command - VA response, speaker transitions from a turn by the VA to a turn by the user are also present. These cases are much rarer in the data set ($N = 114$) than user-VA transitions, but what can be reliably observed even with these instances is that turn transitions in VA-user transitions are also far slower on average than what can be expected in human-human conversation, with a mean transition time of 1497 ms (see Fig. 1, bottom panel, for a distribution of turn-transition times).

In the minority of these cases, the VA initiated the sequence, as in excerpt (5).

(5) [CVR09, day 1]

VA: um mit mir zu SPRECHen, (.) nenne
einfach meinen nAmen und stelle
eine FRAge. (0.92) versuche es
MIT, (.) aLEXa; (0.49) wie ist das
WETter?
to talk to me, just say my name and
ask a question. try, Alexa how is
the weather?
(2.86)
US: aLEXa, wie ist das WETter?
Alexa, how is the weather?

These sequences occur exclusively during the setup phase, i.e., during ongoing regular usage, the VA never initiates a sequence without any preceding user command including the wake word. Hence, most cases of VA-user transitions originate from repeated turn taking between user and VA after the user initiated the sequence, as in excerpt (6).

(6) [CVR03, day 6]

US: aLEXa? (1.29) was können wir
SPIE:ln?
Alexa (1.29) what can we play
(2.21)
VA: oKAY; dann SUCHen wir ein
tolles spiel AUs. (0.52) ich
habe emPFOHlene spiele, (0.46)
KINderspiele, (0.43) oder
geograPHIEspiele. (0.48) welche
davon MÖCHtest du? (0.42) du kannst
auch nach wEIteren optionen FRAgen.
okay then we'll choose a great
game. (0.52) I have recommended
games children's games or geography
games. (0.48) which of these do you
want? (0.42) you can also ask for
further options.
(1.43)
US: KINderspiele;
children's games

As can be seen when comparing the two panels in Fig. 1, these two excerpts are representative for the data set in that the timing of users' turns seems to be well matched to the timing of VA turns in terms of a fairly long modal transition time of more than 1.2 seconds. One conceivable driving force behind this difference to human-human conversation would thus be an adaption on the side of the user to the conversational pace set by the VA. However, while users' contributions are fairly late in interaction with the VA, a major remaining difference between VA turn-timing and user turn-timing is that the timing of user contributions is by far more variable than the timing of VA contributions, showing that while users may adapt to the peculiarities of the VA as a conversational partner, they seem to do so in a less mechanistic way. Other, either alternative or additional reasons for the markedly long transition times by users appear reasonable. They concern the conversational flow and sequential complexity of human-VA interactions in the data and will be discussed as part of the following Section 3.4.

3.4 Sequential complexity

In Sections 3.1 to 3.3 above we found that turn transition times in human-VA interaction are generally slow and that examples like the ones above in excerpts (5) and (6), with users taking a second turn in response to a first turn by the VA, are infrequent. One reason for the slow average turn transition times as well as the invariance in turn timing by the VA is the absence of backchannel turns in the data set. Backchannels are known to exhibit short and well-aligned transition times, due to their brevity and their function of yielding a turn at a transition-relevance place (Knudsen et al., 2020). Their use structures and organises conversations and makes turn allocation flexible, creating a dynamic and variable conversation. With the complete absence of backchannels from the data, the recorded turn exchanges appear static, as the available sequential structures in user-VA interactions are mostly limited to single turn by user - reaction by VA sequences, largely excluding multi-turn sequence expansions by users. Moreover, in the rare cases of follow-up turns by the user that seek to expand a sequence, the VA is frequently found to not continue the sequence in a coherently relevant way. Consider excerpt (7) for an illustration.[4]

(7) [CVR03, day 20]

CU: aLEXa? welches ist das grÖßte tier der WELT.
Alexa which is the largest animal in the world.
(1.20)
VA: das gößte und schwerste tier der erde ist der BLAUwal. (0.47) er ist etwa dreißig meter LANG, und wiegt bis zu zweihundert TONnen.
the largest and heaviest animal on earth is the blue whale. (0.47) it is about thirty meters long and weighs up to two hundred tons.
(5.34)
CU: zweiHUNdert? (0.57) dis sogar mehr als der brAchioSAUrus; (0.56) aLEXa. (0.59) aLEXa? dis sogar mehr als der BRAchiosaurus.
two hundred? (0.57) this is even more than the Brachiosaurus; (0.56) Alexa. (0.59) Alexa? this is even more than the Brachiosaurus.
(1.32)
VA: das könnte deine frage beANTworten? (.) laut wikiPEdia, brachiosaurus ist eine gattung sAUropoder dinosaurier aus der oberjura von nordaMErika. (0.50) [brachiosau-]
this might answer your question (.) according to Wikipedia, Brachiosaurus is a genus of sauropsid dinosaur from the Upper Jurassic of North America.
AU: [alExa LAUter.]
Alexa louder.
(1.38)
VA: (low tone beep)
(5.82)
AU: alexa kAnnst du das bitte wiederHOLN;
Alexa can you repeat that please;
(1.26)
VA: es gibt nichts (.) das ich wiederHOLen könnte.
there is nothing I could repeat.

In this excerpt, the user asks an information-seeking question, one of the most frequent use cases in the data set. The VA gives a relevant and informative answer to the question in next position, which is followed by a partial repeat by the user, which, on the surface, takes the form of a repair initiator, mainly due to its rising intonation, but in this position rather functions as a demonstration of information uptake as well as astonishment about the information (Robinson, 2012). The user goes on to expand the sequence by sharing additional information on the topic, at the same time giving an explanation for the displayed astonishment and opening up sequential possibilities for the VA to continue the conversation. In doing so, the user adds the wake word as an increment to the turn, since this has been learned to be a requirement for sucessful communication with the VA in previous interactions. Probably because wake words in

---

[4] CU = child user; AU = adult user.

turn-final positions did not lead to successful turn exchanges with the VA in previous interactions, the user repeats his last turn with the wake word at the beginning. In reaction, the VA treats the last turn by the user as a stand-alone command, again interpreting the turn as a request for information and not for a sharing of knowledge by the user, thereby failing to take into account the sequential history of the ongoing exchange for the first time in this excerpt. After the VA is stopped mid-turn by a command by a second co-present user to lower the output volume, this second user requests the VA to repeat its last turn. In response, the VA again fails to produce a reaction that is relevant in its sequential position, as it replies that there is nothing that can be repeated. Failures like these are common in the data set and are a major reason for reduced conversational complexity in the recorded human-VA interactions, as they reduce the possibly successful use cases that users can pursue to rather flat command-response sequences with highly predictable, slow VA reactions.

4 Conclusion

This paper reviewed interactions of ten first-time users of voice assistants (VAs) from six households taken from a corpus by Barthel et al. (2023), with a focus on the measurable fluency of conversations, in particular floor transfer offsets between turns by the user and the VA. VA reactions to user commands have been found to be considerably slower than what is commonly observable in human-human interaction, with VAs taking about 1.3 seconds on average to verbally respond to a user's turn. Interestingly, users are found to adjust to this timing pattern very quickly, as they display the necessary patience towards the VA that is needed before a response can be expected. During the long gap between a user's turn and the VA's response, users do not typically re-select themselves for a response pursuit, as could be expected to happen in human-human interaction. Instead, long gaps are usually filled with silence, with the user awaiting the VA's reaction. Unfortunately, the audio-only data set analysed here, while having other advantages, does not allow for a multi-modal analysis of users' behaviour during the gap, like gaze, movement, or body orientation, which could allow for more in-depth insights about the effect of (delays in) turn timing on the user experience during interactions with the VA (Hall et al., 2024).

Next to being rather consistently slow in giving verbal (or generally acoustic) responses, the timing of VA reactions has been found to be highly invariant. While interactants use turn timing in human-human interaction to enrich their interpretations of the contents of their interlocutors' turns (Henetz, 2017; Blohm and Barthel, 2024), this proves to be impossible in human-VA interaction. *When* the VA produces a turn with respect to the timing of the preceding turn by the user cannot be expected to be informative about *what* is intended to be communicated with the turn. This invariance in turn timing makes contributions by the VA appear mechanistic and unnatural, and it was found to be spotted very quickly by users during their first interactions with the VA, leading to user adaptations that include atypically long silent waiting times before VA reactions. Obviously, VA systems that depend on silence after user input to detect transition relevance places are bound to show long reaction latencies. However, the fact that the ensuing long gaps are paired with discontinued, one-shot parsing of the user input leads to undesirably long stretches of overlap with potential continuations of user input after the user's turn has been analysed as complete by the VA. The observed lack of display of awareness of the local conversational situations and the absence of interactional micro-management of the VA, especially during overlapping talk, are probable characteristics that can cause users to perceive the VA as being incapable of fluent conversation. If the delay in responses of VAs cannot be shortened, at least response production should be halted when new input by the user is detected.

In the present data set, which provides naturally occurring, uninstructed, intrinsically motivated interactions of users with their VAs, users are found to very rapidly adapt to the atypicalities in VA turn timing. While repeated turn taking is comparatively rare, users' own turns in reaction to preceding turns by the VA are also produced with more delay than commonly observed in human-human interaction, interestingly matching the common response latencies that are produced by the VA. However, users' turn timing is found to still be considerably more variant than VA turn timing, which preserves some naturalness in the recorded interactions. In order to come closer to a dialogue system that is capable of more naturalistic human-VA turn taking in everyday interaction, VAs need to be equipped with a more human-like ability to time their conversational contributions, both in terms of the alignment of their turn beginnings with users' turns' ends as well as in terms of the variation of their turn timing, which is systematically structured as well as interpreted to be meaningful in human-human interaction (Edlund et al., 2008; Roberts et al., 2015; Strombergsson et al., 2013).

References

Mathias Barthel. 2020. *Speech Planning in Dialogue - Psycholinguistic Studies of the Timing of Turn Taking*. PhD Thesis, Radboud University Nijmegen, Nijmegen.

Mathias Barthel, Henrike Helmer, and Silke Reineke. 2023. First users' interactions with voice-controlled virtual assistants: A micro-longitudinal corpus study. In *Proceedings of SemDial*, pages 105–117, University of Maribor.

Mathias Barthel, Antje S. Meyer, and Stephen C. Levinson. 2017. Next Speakers Plan Their Turn Early and Speak after Turn-Final "Go-Signals". *Frontiers in Psychology*, 8:393.

Mathias Barthel, Sebastian Sauppe, Stephen C. Levinson, and Antje S. Meyer. 2016. The Timing of Utterance Planning in Task-Oriented Dialogue: Evidence from a Novel List-Completion Paradigm. *Frontiers in Psychology*, 7:1858.

D. Bates, M. Maechler, B. Bolker, and S. Walker. 2015. lme4: Linear mixed-effects models using Eigen and S4.

Stefan Blohm and Mathias Barthel. 2024. Modulating the temporal threshold of tolerance for silent interturn gaps. In *Proceedings of the 5th Experimental Pragmatics in Italy Conference*, Venice, Italy.

Stefan Blohm and Mathias Barthel. 2025. Why so cold and distant? Effects of inter-turn gap durations on observers' attributions of interpersonal stance. In *Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue*, Bielefeld, Germany.

Sara Bögels. 2020. Neural correlates of turn-taking in the wild: Response planning starts early in free interviews. *Cognition*, 203:104347.

Sara Bögels, Kobin H. Kendrick, and Stephen C. Levinson. 2015. Never Say No … How the Brain Interprets the Pregnant Pause in Conversation. *PLOS ONE*, 10(12):e0145474.

H Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Judy Davidson. 1984. Subsequent versions of invitations, offers, requests, and proposals dealing with potential or actual rejection. In J.M. Atkinson and J. Heritage, editors, *Structures of social action: Studies in conversation analysis*, pages 102–128. Cambridge University Press, Cambridge.

Mark Dingemanse, Seán G. Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S. Gisladottir, Kobin H. Kendrick, Stephen C. Levinson, Elizabeth Manrique, Giovanni Rossi, and N. J. Enfield. 2015. Universal Principles in the Repair of Communication Problems. *PLOS ONE*, 10(9):e0136100.

Paul Drew. 2009. Quit talking while I'm interrupting: A comparison between positions of overlap onset in conversation. In Markku Haakana, Minna Laakso, and Jan Lindström, editors, *Talk in interaction: Comparative dimensions*. SKS Finnish Literature Society.

Starkey Duncan and George Niederehe. 1974. On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10(3):234–247.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8-9):630–645.

Joel E. Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for voice interface design. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–8, Dublin Ireland. ACM.

Jean E. Fox Tree. 2002. Interpreting Pauses and Ums at Turn Exchanges. *Discourse Processes*, 34(1):37–55.

Rod Gardner. 2001. *When listeners talk: response tokens and listener stance*. Number 92 in Pragmatics & beyond. John Benjamins Publishing Company, Amsterdam Philadelphia.

Rosa S. Gisladottir, Dorothee J. Chwilla, and Stephen C. Levinson. 2015. Conversation Electrified: ERP Correlates of Speech Act Recognition in Underspecified Utterances. *PLOS ONE*, 10(3):1–24.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Stephan Habscheid, Tim Hector, and Christine Hrncal. 2023. Human and Non-Human Agency as Practical Accomplishment: Interactional Occasions for Ascription and Withdrawal of (Graduated) Agency in the Use of Smart Speaker Technology. *Social Interaction. Video-Based Studies of Human Sociality*, 6(1).

Lauren Hall, Saul Albert, and Elizabeth Peel. 2024. Doing Virtual Companionship with Alexa. *Social Interaction. Video-Based Studies of Human Sociality*, 7(3).

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Tania Henetz. 2017. *Don't hesitate! The length of inter-turn gaps influences observers' interactional attributions*. PhD Thesis, Stanford University.

Gail Jefferson. 1989. Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In *Conversation: An Interdisciplinary Perspective*, pages 166–196. Multilingual Matters, Philadelphia.

Gail Jefferson. 2004. A sketch of some orderly aspects of overlap in natural conversation. In Gene H. Lerner, editor, *Pragmatics & Beyond New Series*, volume 125, pages 43–59. John Benjamins Publishing Company, Amsterdam.

Kobin H. Kendrick. 2015. The intersection of turn-taking and repair: the timing of other-initiations of repair in conversation. *Frontiers in Psychology*, 6:250.

Kobin H. Kendrick and Francisco Torreira. 2014. The Timing and Construction of Preference: A Quantitative Study. *Discourse Processes*, 52(4):1–35.

Birgit Knudsen, Ava Creemers, and Antje S. Meyer. 2020. Forgotten Little Words: How Backchannels and Particles May Facilitate Speech Planning in Conversation? *Frontiers in Psychology*, 11:593671.

Antonia L. Krummheuer. 2008. Zwischen den Welten: Verstehenssicherung und Problembehandlung in künstlichen Interaktionen von menschlichen Akteuren und personifizierten virtuellen Agenten. In Herbert Willems, editor, *Weltweite Welten*, pages 269–294. VS Verlag für Sozialwissenschaften, Wiesbaden.

Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6(731):10–26.

Jakub Mlynář, Lynn de Rijk, Andreas Liesenfeld, Wyke Stommel, and Saul Albert. 2025. AI in situated action: a scoping review of ethnomethodological and conversation analytic studies. *AI & SOCIETY*, 40(3):1497–1527.

Adam R. Pearson, Tessa V. West, John F. Dovidio, Stacie Renfro Powers, Ross Buck, and Robert Henning. 2008. The Fragility of Intergroup Relations: Divergent Effects of Delayed Audiovisual Feedback in Intergroup and Intragroup Interaction. *Psychological Science*, 19(12):1272–1279.

Hannah R.M. Pelikan and Mathias Broth. 2016. Why That Nao?: How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4921–4932, San Jose California USA. ACM.

Karola Pitsch, Hideaki Kuzuoka, Yuya Suzuki, Luise Sussenbach, Paul Luff, and Christian Heath. 2009. "The first five seconds": Contingent stepwise entry into an interaction as a means to secure sustained engagement in HRI. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 985–991, Toyama, Japan. IEEE.

Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Montreal QC Canada. ACM.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Stuart Reeves, Martin Porcheron, and Joel Fischer. 2018. 'This is not what we wanted': designing for conversation with voice interfaces. *Interactions*, 26(1):46–51.

Silke Reineke and Henrike Helmer. 2024. User practices in dealing with trouble in interactions with virtual assistants in German: Repeating, altering and insisting. *Discourse & Communication*, page 175048132411271494.

Felicia Roberts and Alexander L. Francis. 2013. Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, 133(6):EL471–EL477.

Felicia Roberts, Piera Margutti, and Shoji Takano. 2011. Judgments Concerning the Valence of Inter-Turn Silence Across Speakers of American English, Italian, and Japanese. *Discourse Processes*, 48(5):331–354.

Seán G. Roberts, Francisco Torreira, and Stephen C. Levinson. 2015. The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in Psychology*, 6.

Jeffrey D. Robinson. 2012. Epistemics, action formation, and other-initiation of repair: the case of partial questioning repeats. In Makoto Hayashi, Geoffrey Raymond, and Jack Sidnell, editors, *Conversational Repair and Human Understanding*, pages 261–292. Cambridge University Press, Cambridge.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In Deborah Tannen, editor, *Analyzing discourse: text and talk*, pages 71–93. Georgetown Univ. Pr, Washington, DC.

Emanuel A. Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1):1–63.

Margret Selting, Peter Auer, Sagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, and Thirteen Others. 2011. A system for transcribing talk-in-interaction: GAT 2. *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, 12:1–51.

Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67:101178.

Tanja Stivers, Nick J. Enfield, Penelope Brown, C. Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, J. P. de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Sofia Strombergsson, Anna Hjalmarsson, Jens Edlund, and David House. 2013. Timing Responses to Questions in Dialogue. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 2584–2588, Lyon.

Emma M. Templeton, Luke J. Chang, Elizabeth A. Reynolds, Marie D. Cone LeBeaumont, and Thalia Wheatley. 2022. Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences*, 119(4):e2116915119.

Louis Ten Bosch, Nelleke Oostdijk, and Lou Boves. 2005. On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47(1-2):80–86.

V.H. Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society*, pages 567–577, Chicago.