SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies
13 September 2016, San Francisco, USA

Towards graceful turn management in human-agent interaction for people with cognitive impairments

Ramin Yaghoubzadeh, Stefan Kopp

Social Cognitive Systems Group, CITEC, Bielefeld University, Germany
ryaghoubzadeh@uni-bielefeld.de, skopp@uni-bielefeld.de

Abstract

A conversational approach to spoken human-machine interaction, the primary and most stable mode of interaction for many people with cognitive impairments, can require proactive control of the interactive flow from the system side. While spoken technology has primarily focused on unimodal spoken interruptions to this end, we propose a multimodal embodied approach with a virtual agent, incorporating an increasingly salient superposition of gestural, facial and paraverbal cues, in order to more gracefully signal turn taking. We implemented and evaluated this in a pilot study with five people with cognitive impairments. We present initial statistical results and promising insights from qualitative analysis which indicate that the basic approach works.

Index Terms: human-computer interaction, virtual assistant, interruption, turn taking, gesture, cognitive impairment

1. Introduction

Spoken human-machine interaction has become a widely adopted paradigm in recent years. In addition to being a helpful technology to keep one's hands free in a variety of everyday contexts, spoken interaction also opens access to modern technology as a whole for certain groups of people, specifically those that cannot readily understand, learn, read, or manipulate interfaces employing other modalities. While graphical interfaces with flat hierarchies counteract some of these usability problems, the presentation and negotiation of information there does not always correspond well to those in spoken human-human interaction, which many of those people will be quite familiar with. However, off-the-shelf spoken language technology, which has become very good at recognizing words even when uttered by new users and answering common sets of – well-formed – questions, usually also lacks many of the aspects of this human-human mode of interaction, namely its conversational, incremental and reciprocal nature. Humans in interaction constantly exchange back-channel information relating to – possibly preliminary – evaluations of the unfolding stream of information. By attending to the other party, a speaker is aware of the back-channel feedback (paraverbal, facial, gestural) of a listener and will incrementally and smoothly incorporate it into their content selection and presentation. Likewise, the listener can tell when and where the speaker encounters a problem, and can intervene in a timely manner, if necessary. They might also be aware of possible points of misunderstanding, and either address them immediately – by barging in, often in a cooperative fashion – or queue them for later implicit or explicit resolution.

To achieve these capabilities, one crucial function in dialog systems is interrupting the user and taking the floor – but doing so in a cooperative manner that is consistently acceptable for users even over longer time spans and many repeated instances.

In the following sections, we will first provide an overview of the theoretical and analytical background relating to multimodal turn management signals, and look at related work on turn taking control in interactive systems. We will then present the scenario and user groups for our pilot study, and present the autonomous interruption controller that was run alongside a Wizard-of-Oz controlled main dialogue. After a description of the procedure and the interview structure for the assessment of subjective ratings, we will present initial statistical data, followed by a detailed analysis of one particularly informative interaction fragment, before concluding our presentation.

2. Background and related work

The different manifestations of turn-keeping and turn-grabbing signals were early described by Duncan and Fiske [1]; Bohle [2] provides a comprehensive overview and discussion. According to the latter source, one single characteristic, unimodal signal generally constructs a clear meaning in these situations, but the intensity can be increased by employing multimodal presentation. Addressing those behaviors in the listener role that do not generally have the effect of signaling a desire to obtain the floor, they list minimal acknowledgements, clarification requests, other-completions, short paraphrase, and head gestures – one should also add paraverbal back-channel feedback to the list [3]. As for the floor-asserting behaviors, more specifically floor grabs by the listener, they list head or gaze aversion from the speaker, as well as the initiation of gesticulation. Kaartinen [4] analyzed gestural behavior as turn-taking signals in news interviews, noting the role of adaptation of gestures in forming multifunctional constructs encompassing turn-taking information; and particularly highlighting (quasi-)deictic handshapes, first and foremost extended fingers.

The efficacy and acceptability of interruptive behavior on the part of dialog systems has been well researched.

Ter Maat et al. [5] investigated the effect of interruptive turn taking by an agent in a Wizard-of-Oz setup, comparing (unimodal spoken) early turn grabs, i.e. interruption-causing overlaps, to turn taking immediately at and slightly after appropriate points. They found that early barge-ins were perceived as more assertive, but also as significantly more disagreeable, rude and of lower conversational aptitude.

Cafaro et al. [6] examined ratings of simulated agent-agent interactions with comparable interruption types, but additionally manipulating the cooperativity of the interrupter's content selection strategy (e.g. elaboration vs. topic jump as a reply to a question). Strategy changes towards cooperativity in particular led to increased perceived friendliness and reduced dominance – but less so that the selection of interruption type, corroborating the findings of ter Maat et al.

10.21437/SLPAT.2016-5

In our own work with autonomous dialog systems for older adults and people with cognitive impairments, we previously found that the spoken, conversational, paradigm of task-related interaction with a system transfers to both groups, both in terms of feasibility and acceptance [7]. In those studies, we strictly let the users control the pacing of the dialogue and the amount of transferred content, while priming for specific information presentation only when subjects yielded their turns spontaneously. While the performance and perception of the autonomous system was comparable to an earlier Wizard-of-Oz prototype [8] for most people, we found that a noticeable minority of participants from both groups were prone to excessively verbose or tangential presentation even after repeated instruction to the contrary (cf. Fig. 1) – which caused ASR and NLU to drastically decrease in performance, thus necessitating proactive, interruptive, system-side floor governance.

3. Pilot study

Considering this requirement for proactive floor management, and the aforementioned work on the reduced perceived cooperativity caused by pure verbal barge-ins, we constructed an autonomous prototype interruption controller (`flow_controller`) based on the research on the multimodal construction of turn-grabbing behavior. We employed it in a pilot study with five participants with cognitive impairments, engaging in a spoken human-agent interaction in a Wizard-of-Oz-controlled discussion game. These sessions were embedded in a larger study exploring the effects of agent body language on the persuasiveness and reception of system-generated argumentation for older adults as well as younger controls (n=40 each; analysis in progress), for which younger subjects with cognitive impairments were also recruited by our corporate partner, the large health and social care provider v. Bodelschwinghsche Stiftungen Bethel.

Since the participants with cognitive impairments were not expected to be able to fill out the required 90+-DOF questionnaire for the experiment proper, the interruption condition was piloted instead. Participants from all user groups were presented with the same scenario and task, described below.

3.1. Setup and participants

From the point of view of the participants, the setup consisted of a 27" touch screen, a microphone and an eye tracker, as well as cameras recording two angles (Fig. 2 depicts the view from the rear camera). The screen was able to show the game scene, showing the animated virtual agent, "Billie", as well as lists representing the game state. The agent was controlled by the ASAPRealizer software for behavior realization [9]; text-to-speech was provided by a CereVoice [10] component controlled by the realizer, which was able to provide some realizations of paraverbal signals. A directional microphone and a low-cost eyetracker were mounted below the screen. The system was primarily controlled by a Wizard-of-Oz console that interacted with a component managing the game state and graphical presentation. However, the agent's nonverbal and paraverbal behavior was controlled autonomously by the `flow_controller`, described below. This was contingent on the audio state as reported by a simple audio level detector, which was in turn inhibited by ongoing agent utterances (see Fig. 3 for an overview of components). The eye tracker was, in this incarnation of the system, employed as a source of data for qualitative analysis and as a basic functionality test for our user groups, although incorporation into the interruption controller is planned for the future.

Figure 2: *Overview of the setup (as seen from the rear camera). Touchscreen PC with eye tracker mounted below. High-fps face camera recording from below the screen. The physical item list and the items to allocate on it in preparation for the game are visible on the desk. The microphone is occluded by the participant (anonymized). Start of interaction.*

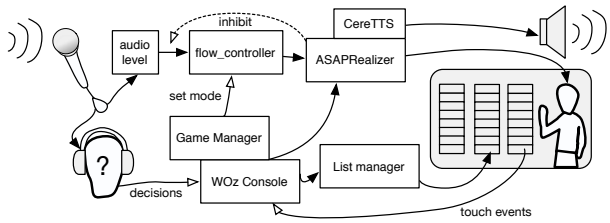

Figure 3: *Overview of components. "Audio level" and "Flow controller" constituted the autonomously acting subsystem. Discourse progress and contents were controlled by the Wizard (wearing headphones).*

Participants (n=5, 2 male, 3 female, ages 29–48) were recruited from a care institution providing support to people with cognitive impairments, both in communal and individual assisted living arrangements. Exact clinical diagnoses were not able to be divulged by the care providers. As with previous experiments, we asked that only subjects be recruited whose articulation was clear enough to be generally comprehensible by untrained, unfamiliar listeners.

3.2. Interruption controller

Animations for signaling turn grabs were first recorded using full-body motion capturing, then reduced to spinal and arm movements and preprocessed to obtain a chainable, smooth database for procedural animation. In lieu of a speech recognizer's voice activity detection module, a simple audio-level activated trigger was implemented using pyaudio that reported durations of ongoing and finished audio events. It was inhibited by open agent utterances, effectively removing cross-talk at the cost of ignoring overlapping speech. We were only interested in the duration of the user turn proper, and surmised from previous experiments that prolonged periods of overlap were unlikely to be frequently caused by the user group.

The `flow_controller` component, responsible for interruption generation, was able to be configured in four modes (0–3): modes 1, 2 and 3 started a slowly progressing three-state

```
a1  AGNT  Do you have another appointment?
    SUBJ                                  Yes. Then, I have yet another appointment ... on Friday

a2  AGNT  So, on Friday, right?        OK. At what time does it start?
    SUBJ                       Right.                                  Then I'll pick 3 PM again,

a3  AGNT                  So, at 3 PM, right?                 So, at 3 [interrupt]   Good.
    SUBJ  have ice cream.                      [hoarsely] Yah            Yes.

a4  AGNT  So, at that time, there is "Have ice cream", right?         Okay. Then I'll enter it as follows...
    SUBJ                                                    Right.


c1  AGNT Then tell me the next appointment, please.
    SUBJ                                          I have uhm (-) today shopping *thr 3 PM 3 PM *appoin

c2  AGNT
    SUBJ  appointment with <Name> (.) and then I also(?) later go shopping later *thr 3 PM with <Name>

c3  AGNT
    SUBJ  (.) and (-) then I also go shopping (-) later
```

Figure 1: Transcripts of interaction segments with different interaction styles observed in previous studies with an autonomous prototype system. The non-verbatim translation from German attempts to represent dysfluencies and errors intuitively. **Top:** older adult, brief but casual style; **bottom:** person with noticeable cognitive impairment, verbose turns, exacerbated by dysfluent and unclear articulation; this led to considerable ASR processing delays (the participant eventually entered the shopping appointment successfully).

Figure 4: *Four stages of nonverbal interruptive behavior. From left to right: Idle; first signal (reached after about 4s, shown with mouth half-open); second signal (reached after about 7s); final stage (held until user ends utterance or Wizard barges in).*

cascade of interrupting behaviors (cf. Table 1), varying slightly in surface form by mode, for all utterances above a threshold duration (set to 2s+). Behaviors included hand raises (half-open hand or pointing shape), gaze aversion, open mouth and paraverbals ("ah" and throat clearing). For short utterances, the agent would provide positive feedback by nodding. Mode 0 allowed for a non-interrupting state: the agent would nod at the defined transition points and then remain static in the idle position for the remainder of the user's turn.

3.3. Task and procedure

The task for the participants was a discussion game in the "desert survival" scenario. The premise was that the agent and the subject were stranded in a remote location, with their airplane destroyed and only a set of twelve items still intact. The task of participants was to order them, ranked by their perceived usefulness, and then engage in a discussion with the agent to find a consensus order.

A brief principal instruction was provided by the experimenter, then the interaction started. The Wizard controlling the agent would first greet the user and ask their name, then explain all stages of the game. The users were then asked to rank the list of items, for which we prepared a paper list with twelve empty item slots, and paper slips corresponding to the items to be placed on the list. All items were individually explained to the user, as was the meaning of list items "high" or "low" on the list to account for possible problems with abstract numbers. Subjects then had two minutes to decide on a preferred ranking. The agent would enter "hidden" rankings on his list on the screen. When the subject's list had been finished, the agent asked them to read them off in order. The agent entered those rankings in the leftmost (user) list. Thereafter, the agent "showed" its list – but instead silently generating a conflicting ranking according to a pre-defined permutation scheme.

The subsequent crucial discussion phase started, the agent presented the user's ranking along with its own and a relative statement (like: "You placed the Lighter on 1, I have it on 7. So, you rated it as more important than I did. Could you explain why you placed it there?") At this point in the discussion, the interruption controller, that was set to mode 0 (do not interrupt) in all other contexts, was set based on the index of the currently

Table 1: *Interruption controller: modes and phases, with respective generated actions. Modes were set according to the index of the discussed items (see text). Phases were successively entered during user turns that proceeded for long enough. *Note: all hand positions other than the idle position were combined with a slight gaze aversion (constant angle).*

Phase \ Mode	0	1	2	3
– (short)	nod	nod	nod	nod
1	nod	raise to mid*, index extended	raise to mid*, mouth open	raise to mid*, index ext'd, utter "ah"
2	nod	raise to high*, mouth open	raise to high*, mouth open	raise to high*, mouth open
3	nod	keep raising*, hand open, utter "ah"	keep raising*, hand open, clear throat	keep raising*, hand open, clear throat

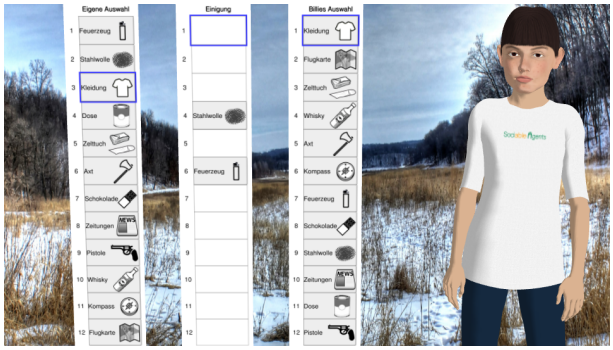

Figure 5: *Crashed in the taiga! Scene setup: left list: initial user choice; right list: agent 'choice'; center list: ranking made by user after the exchange of arguments. In this instance, the discussion has just taken place for the user's third most important item ("clothes", highlighted), and the user just selected a final slot for it, in this case following the agent's suggestion.*

discussed item (from first to twelfth: modes 0, 2, 0, 1, 2, 0, 1, 2, 3, 0, 2, 3). Therefore, at items 1, 3, 6, and 10 (a third of all items), the system would remain in non-interrupting mode as a reference for comparison.

After subjects presented their opinion about an item, the agent would invariably utter an argument from a precompiled list that contained one supportive and one dismissive argument for each item, selected depending on relative ranking. Then, the user was always given the choice to fix a position for the item on the common, central, list. Selections in the lists could be made either by speaking the rank number, or by touching the corresponding field on the screen (cf. Fig. 5). When all 12 items had been discussed and agreed upon, the user could modify the list one final time if they so wished, after which thanks and valedictions were presented by the agent, and the iteraction was over.

After the experiments, a simplified structured interview was conducted for each subject. A visual 5-point Likert rating aid (definitely yes – ... – definitely no) was employed to gain quantifiable ratings to ten questions, although the primary aim was to gather comments and qualitative information. The questions were (approximate correspondences in Simple English): Q1: Did the game with Billie go well? Q2: Was Billie nice to you? Q3: Did you understand what Billie said? Q4: Could Billie understand you as well? Q5: Did Billie listen when you wanted to say something? Q6: Did you find the game easy enough? Q7: Was the length of the game okay? Q8: Did you call the shots in the game? Q9: Did Billie butt in or interrupt you? Q10: Did you have fun playing the game? In particular, Q5 and Q9 were inserted as a pair of opposing valence, contingent on the experimental manipulation and its effectiveness and perceived

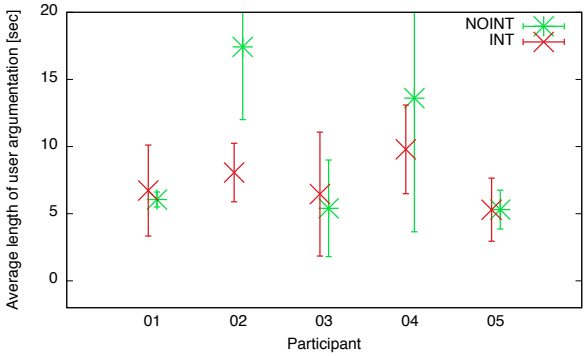

Figure 7: *Length statistics for discussion turns for each participant, non-interrupting (mode 0) vs. interrupting (mode 1–3) conditions. (Participant 03 was in the audio-only interaction.)*

intrusiveness. Subjects were also asked to report prior technical experience and answer the more general question "Do you enjoy talking to other people?".

3.4. Results and discussion

All five subjects were able to complete the whole task. Subject 2 mostly opted for the touch-based rank selection in the course of the discussion, while the other participants interacted using speech only. For subject 3, a technical problem led to the loss of the video signal on their screen after the introductory explanation; the subject accepted this silently and concentrated mostly on her physical list from then on. Since the task was completed successfully, we used this as an audio-only reference. The eye-tracker only reliably worked for subjects 1 and 4, thus final analysis of the gaze behavior can only be made after video-based annotation for the other subjects.

The scenario was not consistently conducive to long elaborations by all users, although two participants did produce them. While the sample size is much too small for robust statistical results, interesting trends can be gleaned from the graph in Fig. 7: for subject 2 and 4, who produced the most elaborate argumentations in the turn we focused on, a noticeable difference between the non-interrupting and interrupting items can be seen, indicating that the interruption strategy might have had an effect. This was most valid for the first three or four items, where all participants had a quite clear idea of their motivation to rank the items highest (note that subject 4, during item 6, where they were free to talk, just coughed, sighed and uttered "tough question", staring at the agent until the Wizard continued.) In the videos, no clear, hard 'breaking points' can be observed in any of these cases, indicating that the progressive nature of the signal might have progressively steered them to a smooth turn completion.


```
01  SUBJ       ja-aso meine überlegung war weil ö:   (0.6)    den ö:    (0.4)     ö:m  (0.3)
               yes-well my idea was since uh                   the uh              uhm
    SUBJ-gaze  @GUI                                              @Table           @GUI
    AGNT-gest  | hold IDLE ------------------------------------------------------| raise to MID ---

02  SUBJ       <<click>> der whiskey          ja   ess (.) essma so nich wichdich is ne den kam-man (0.8)
                         the whiskey          well fir...  firstly isn't that important right you can
    SUBJ-gaze                FLICK@Agent @GUI                          @Agent          @GUI 0.6s   @Agent
    AGNT-gest  ---------------| hold MID ------------------------------| raise to HIGH, extend index
                                                                          Img#1     Img#2     Img#3

03  SUBJ       den |kamman weil wemman schmerzen hat oder so zum   (0.85)    <<click>>    (-)
               it   you can since when you are in pain or something to
    SUBJ-gaze  @GUI                                          @Agent       @GUI
    AGNT           (ah)
                     ah
    AGNT-gest  ----| raise to APEX, open palm ----------------------| hold APEX ---------------

04  SUBJ        brauchen aba-ansonsten   (1.5)
                use but otherwise
    SUBJ-gaze           @Table        @Agent
    AGNT                                      bei mir is-der whiskey an stelle fünf
                                              I've got the whiskey at position five
    AGNT-gest  | relax to IDLE -----------| hold IDLE ----------------------------------
```

Figure 6: Transcript and anonymized snapshots from an item argumentation by subject 1 (see 3.5 for discussion). Pause lengths in parentheses, short pauses given as (.), (-) [11]. Times of the three frames indicated by Img#x, in blue.

We surmise that an attenuation effect contributed to an observably reduced verbosity in later items: participants apparently ranked the most important and least important items with a clear idea of their merits or downsides, with the rest of ranks (around #7–#11) possibly filled rather indifferently with the unclear remainder.

The ratings of the interview questions relating to the experimental manipulation were rated equally by all participants (Q5 was rated "decidedly yes", Q9 as "decidedly no"). Agent niceness and enjoyment of the game were likewise rated with the most affirmative option by all subjects. Subject 2 noted in the free-form interview comments that she noticed the agent's gestures but found them slightly odd; she thought the agent looked like it "might want to say something".

3.5. Qualitative analysis

Even for those participants where no noticeable effect on utterance length could be observed, detailed analysis still indicates that the agent behavior modulated the subject's pacing and continued attention to the agent, indicating that timely and contingent content presentation, as afforded by an autonomous dialogue system as opposed to WOz, could allow for a cooperative takeover at these points.

Fig. 6 highlights one such situation. The participant is mainly focused on the GUI part of the screen (the agent's list) and his paper list, but regularly checks back with the agent while he is explaining his decision. The section highlighted in red shows a typical fragment of interaction where the system managed to capture the attention of the user. The user looks at the agent during his utterance – the agent has already performed the weakest interruption signal and is just about to generate the second animation (raise hand further with index finger extended and mouth slightly open). The user looks back to the left, but his gaze returns to the agent after merely 0.6 seconds. He then hestitates mid-sentence for 0.8 seconds, directing his attention immediately at the agent.

We argue that the interruption subsystem managed to construct a possible transition point here (which the Wizard did however not utilize before the user resumed) – with the short gaze shift to the lists either due to a delay in the user's reaction time, or else having seen the continuation of the signal in peripheral vision.

4. Conclusions

Our preliminary results indicate that the nonverbal agent behavior generated by the interruption controller did lead to graceful (self-)interruption in some of our participants with cognitive impairments, while others that did not noticeably vary in presentation length still reacted noticeably to the emitted signals. As for the ratings of intrusiveness and cooperativity, none of the participants judged this as interruptive behavior per se, and agent ratings were all maximally favorable. The practical efficacy of these interruption signals is certainly also dependent on proper contextual content selection at the very time of a generated transition opportunity or floor yield – which was not trivial to realize for the human Wizard. Depending on the scenario, this could work better with a spoken dialog system, which we will explore next. Another issue is the exact surface realization of the signals of increasing intensity. While we hand-crafted our signals based on literature on the topic, an evaluation of different versions of the signals (possibly using crowdsourcing with unimpaired users) could also be helpful.

Overall, we deem the further exploration of nonverbal control signals to govern the floor in conversational, spoken dialog systems a promising endeavor.

5. Acknowledgements

This research was partially supported by the German Federal Ministry of Education and Research (BMBF) in the project 'KOMPASS' (FKZ 16SV7271K) and by the Deutsche Forschungsgemeinschaft (DFG) in the Cluster of Excellence 'Cognitive Interaction Technology' (CITEC).

6. References

[1] S. Duncan and D. Fiske, *Face-to-face interaction: Research, methods, and theory*. Hillsdale, NJ: Erlbaum, 1977.

[2] U. Bohle, *Das Wort ergreifen – das Wort übergeben: Explorative Studie zur Rolle redebegleitender Gesten in der Organisation des Sprecherwechsels*. Berlin: Weidler, 2007.

[3] J. Allwood, J. Nivre, and E. Ahlsen, "On the semantics and pragmatics of linguistic feedback," *J Semantics*, vol. 9, pp. 1—-26, 1992.

[4] S. Kaartinen, "Multi-functional gestures in interruptions a look at news interview situations," Master's thesis, Oulu, Finland, 2013.

[5] M. ter Maat, K. P. Truong, and D. Heylen, *How Turn-Taking Strategies Influence Users' Impressions of an Agent*. Berlin, Heidelberg: Springer, 2010, pp. 441–453.

[6] A. Cafaro, N. Glas, and C. Pelachaud, "The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, ser. AAMAS '16. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2016, pp. 911–920.

[7] R. Yaghoubzadeh, K. Pitsch, and S. Kopp, "Adaptive grounding and dialogue management for autonomous conversational assistants for elderly users," in *Proceedings of the 15th International Conference on Intelligent Virtual Agents*, ser. LNCS (LNAI), vol. 9238, 2015, pp. 28–38.

[8] R. Yaghoubzadeh, M. Kramer, K. Pitsch, and S. Kopp, "Virtual agents as daily assistants for elderly or cognitively impaired people," in *Proceedings of the 13th International Conference on Intelligent Virtual Agents*, ser. LNCS (LNAI), vol. 8108, 2013, pp. 79–91.

[9] H. van Welbergen, D. Reidsma, and S. Kopp, "An incremental multimodal realizer for behavior co-articulation and coordination," in *Proceedings of the 12th International Conference on Intelligent Virtual Agents*, ser. LNCS (LNAI), vol. 7502, 2012, pp. 175–188.

[10] M. P. Aylett and C. J. Pidcock, *The CereVoice Characterful Speech Synthesiser SDK*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 413–414.

[11] M. Selting et al., "Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)," *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, vol. 10, pp. 353–402, 2009.