The Role of Prosody and Gaze in Turn-End Anticipation

Chiara Gambi (cgambi@exseed.ed.ac.uk)
Department of Psychology, 7 George Square
Edinburgh, EH8 9JZ U.K.

Torsten Kai Jachmann (jachmann@coli.uni-saarland.de)
Department of Computational Linguistics, Campus
Saarbruecken, Germany

Maria Staudte (masta@coli.uni-saarland.de)
Department of Computational Linguistics, Campus
Saarbruecken, Germany

Abstract

How do listeners integrate multiple sources of information in order to accurately anticipate turn endings? In two experiments using synthesised speech and a virtual agent we examined the role of verbal and gaze information in a turn-end anticipation task. Listeners were as good at anticipating the synthesised voice as they were with human speakers (Experiment 1). However, the direction and timing of the agent's gaze had little influence on their accuracy (Experiment 2). Overall, these findings support the idea that anticipation of turn ends relies primarily, but not exclusively, on verbal content.

Keywords: turn-taking; prediction; pitch; gaze; virtual agent.

Introduction

Turn-taking is a fundamental component of human interactions (Levinson, 2006; Sacks, Schegloff, & Jefferson, 1974). Yet, the understanding of the cognitive mechanisms underlying turn-taking is in its infancy (Magyari, Bastiaansen, De Ruiter, & Levinson, 2014), and so are the attempts to implement human-like turn-taking in artificial dialogue systems (DeVault, Sagae, & Traum, 2011). Most notably, we lack clear experimental data on the question of how interlocutors combine linguistic information (i.e., syntax and semantics) with suprasegmental information (i.e., prosody) and with visual information (i.e., gaze) to manage turn-taking. To fill this gap, we conducted two turn-end anticipation experiments.

Lexico-Semantic and Prosodic Cues

De Ruiter, Enfield, and Mitterer (2006) pioneered the use of the task we also employed in our study. They asked Dutch listeners to try and anticipate the precise end of turns extracted from a real conversation. Participants did so with great accuracy.

Indeed, the distribution of bias (i.e., the time off-set between the participants' response and the end of the turn) in this task resembled very closely the distribution of inter-turn intervals (i.e., the time off-set between the end of one interlocutor's turn and the start of another interlocutor's turn) in the original conversations.

Crucially, when lexico-semantic content was made unintelligible (via low-pass filtering), participants performed significantly worse. When the pitch contour was flattened, however, participants performed equally well. The importance of lexico-semantic content was confirmed by Magyari and De Ruiter (2012), who found that turns with more positive bias are less predictable. Finally, Riest, Jorschick, and De Ruiter (2015) showed that listeners can still anticipate turn endings when closed class words are made unintelligible, but not when open class words are.

In sum, there is substantial evidence that lexico-semantic information matters for turn end anticipation but De Ruiter et al.'s (2006) claim that prosody is not necessary is still controversial. While the original finding that flattened pitch does not impair turn-end anticipation was confirmed for German adults (Keitel, Prinz, Friederici, von Hofsten, & Daum, 2013), the same study revealed some benefit from the presence of natural intonation for children. Moreover, all of these studies investigated one aspect of prosody, namely pitch. It is not clear whether other suprasegmental features, such as speech rate, location and duration of pauses, and presence of lengthening on some syllables (Gravano & Hirschberg, 2011) are important for turn-end anticipation. In fact, some experimental (e.g., Cutler & Pearson, 1986; Hjalmarsson, 2011) and corpus findings (Gravano & Hirschberg, 2011) suggest a role for pitch, whereas results for phrase-final lengthening are less clear-cut (Gravano & Hirschberg, 2011; Hjalmarssson, 2011). Therefore, the first aim of our study (Experiment 1) was to replicate De Ruiter et al.'s original finding, using the same turn-end anticipation task, but with a different set of turns extracted from spontaneous (German) conversations.

Gaze Cues

In face-to-face interaction, gaze has been identified as an important cue for holding and yielding turns. Kendon (1967) analysed the behaviour of seven pairs of interlocutors and found that the speaker typically looked away from the listener at the beginning of a ‘long’ turn (>5 sec.) and were increasingly more likely to look towards the listener as she approached the end of her turn.

Kendon’s seminal observations received some experimental support in subsequent studies (Torres et al., 1997; Edlund & Beskow, 2009; Bavelas et al., 2000; Hjalmarsson & Oertel, 2012). Moreover, in a recent study (Skantze et al., 2014) a robot used syntactic completeness and filled pauses to invite or avoid feedback from its human interlocutors. Most interestingly, the robot either displayed random gaze behaviour, gaze behaviour consistent with Kendon’s observations, or was hidden behind a paper board. The face-to-face setting enhanced the effect of the other (verbal) turn-taking cues.

To our knowledge, the study by Skantze et al. (2014) is so far the only one to combine lexico-syntactic and gaze information for detecting turn endings. However, they did not vary the timing of the agent’s gaze cues, and they did not report the timing of user’s feedback. While these authors where more interested in whether the user would provide feedback or not, the second aim of our study (Experiment 2) was to test whether varying the timing of the agent’s gaze would induce human listeners to expect the turn to end earlier or later.

Detecting the End of Virtual Agents’ Turns

Clarifying the role of suprasegmental features and gaze in turn-end anticipation is important also with respect to improving an artificial agent’s ability to participate in smooth turn taking with the user (Hjalmarsson, 2011).

Hjalmarsson (2011) used Expros (Gustafson & Edlund, 2008) to produce a synthesised version of real conversational turns and found that participants could accurately decide whether the synthesised speaker was going to yield or keep the turn; indeed, they were as good as when listening to natural turns. This result is striking, but it must be noted that the synthesised versions were very closely matched on all prosodic features to the original recordings, including fundamental frequency, intensity, timing, and even laughter and breathing. The third aim of our study (Experiment 1) was to further test whether a synthesised voice could afford the same accuracy in turn end anticipation as a human voice, despite differences in fundamental frequency and timing, and the absence of extra-linguistic features like laughter and breathing. In this way, we also made sure that our synthesized turns would be processed similarly to natural turns by human participants, which was a pre-condition for adding the gaze manipulation in Experiment 2.

We know that an avatar’s or a robot’s gaze behaviour can affect the turn-taking style of a naïve participant (Edlund, & Beskow, 2009: Skantze et al., 2014). In those studies, however, the authors employed relatively coarse-grained manipulations. Crucially, no study has yet manipulated the *precise timing* of gaze cues with respect to the end of the turn itself. Moreover, and no less importantly, no study has yet investigated whether human participants are able to accurately anticipate the end of turns when they listen to synthesised voice *and* perceive gaze cues produced by a virtual agent. This was the fourth and final aim of our study (Experiment 2).

Experiment 1: Prosody

In this experiment, German speakers attempted to anticipate the precise end of turns. As in De Ruiter et al. (2006), we presented turns either in their original version (NAT) extracted from a corpus of conversational German, with flat intonation (NOPITCH), or with unintelligible content but preserved intonation (NOWORDS). In addition, we presented the same turns in a synthesised voice (SYNTH). The synthesised turns were automatically generated using a text-to-speech synthesiser (MARY TTS; Schröder & Trouvain, 2003).

On the basis of De Ruiter et al. (2006) and Hjalmarsoon (2011) we hypothesised that (1) participants would more accurately anticipate the end of turns in the NAT, NOPITCH, and SYNTH conditions than in the NOWORDS condition, (2) they would be as accurate in the NOPITCH as in the NAT condition, and (3) also as accurate in the SYNTH as in the NAT (and in the NO PITCH) conditions.

Methods

Twenty-four native speakers of German took part. They were each paid 5 euros and reported no language or auditory impairments. Ninety-six turns were extracted from the Kiel Corpus of Spontaneous Speech (Kohler, 1996). This corpus contains 6 conversations amongst pairs of German native speakers who were discussing scenes from the popular TV series Lindenstraße. Speakers could not see one another. Following De Ruiter et al. (2006), we chose only turns that were at least 5 words long, and that were produced by speakers of both genders (female, 57 turns; male, 39 turns). One third of the selected turns was annotated in the corpus as a smooth transition (smooth), one third was followed by a silent pause longer than 100 ms (pause), and the remaining third ended with an overlap longer than 100 ms (overlap) between the interlocutors (see Table 1 for summary statistics on turn duration). These selection criteria served to ensure a good range of predictability (in Expriment 2 we also explored whether turn type interacted with our gaze manipulation; see below).

NAT turns were the original turns extracted from the corpus (interlocutors were recorded on separate channels, so any backchannels were not included). NOPITCH turns were resynthesised using Praat (Boersma & Weenink, 2014) with constant pitch equal to the mean pitch of the original recording (using PSOLA resynthesis). NOWORDS turns were created in Praat by low-pass filtering the original at 500 Hz.

Table 1: Turn Duration by Turn Type and Condition (ms).

NAT, NOPITCH, NOWORDS			
	Turn Type		
Measure	smooth	overlap	pause
Mean duration	3325	4380	4377
Min duration	1007	1261	1173
Max duration	7365	9720	9433
SYNTH			
	Turn Type		
Measure	smooth	overlap	pause
Mean duration	4398	6186	5789
Min duration	1365	2475	1675
Max duration	8780	12650	11650

SYNTH turns had the same content as the originals (including repetitions and disfluencies, but not laughter and breathing). Their prosody was initially produced by MARY, which uses prosodic boundaries (specified in GToBi) to define pitch contours, accents, and phrase-final lengthening. Boundary locations are determined by basic prosody rules (see Schröder and Trouvain, 2003 for details) and were manually adjusted, if needed, to match the original turns' pitch contour or pauses made by the speaker. Individual words were also manually accented if necessary. Fundamental frequency and timing (see Table 1) of SYNTH turns still differed substantially from the originals. The amplitude of all turns was normalised to minimise differences in intensity between conditions (an example is available at http://dx.doi.org/10.7488/ds/234).

We created four lists using a Latin Square design, so that each participant heard any given turn only once, but across participants each turn was presented in every condition (NAT, NOPITCH, NOWORDS, SYNTH). Presentation was blocked by condition and block order counterbalanced across participants. Within a block, presentation order was individually randomised using E-Prime 2.0 (Schneider, Eschman, & Zuccolotto, 2002)

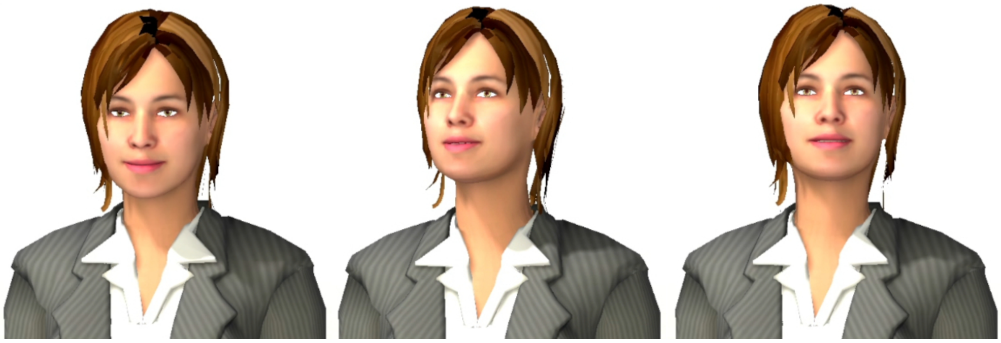

Figure 1: The three main postures of Embr in the AVERT condition. From left to right: Mutual gaze - Gaze to upper left - Averted gaze to the right towards turn end.

Participants listened to the turns over headphones. A 500-ms fixation cross preceded each trial. The screen turned red to mark the start of the turn. Playback was stopped as soon as the participant gave a response to avoid learning effects; after a 1500-ms inter-trial interval (ITI), a new trial began. Instructions emphasised that participants should press the response button as soon as they *expected* the interlocutor to stop, so that the button would be pressed exactly at the turn's end (rather than *waiting* for turn end and *then* pressing the button, cf. De Ruiter et al., 2006). A session lasted about 15 minutes.

Results

Bias is the time offset between the button press and the end of the turn. Positive bias indicates the participant pressed the button after the end of the turn; negative bias indicates they pressed the button before the turn was over. The closer the bias to zero, the more accurate the response. Outliers with a bias higher than 9000 ms in absolute value (0.3 % of the data) were discarded. See Table 2 for results.

Table 2: Mean bias (SE) in ms by Condition in Exp 1.

Condition	Mean	SE
NAT	-315	106
NOPITCH	-146	117
NOWORDS	-460	141
SYNTH	-524	115

Bias is strongly negatively correlated with the duration of a turn (De Ruiter et al., 2006). Since SYNTH turns were considerably longer than the other versions (Table 1), we regressed bias on duration, and looked at the effect of our manipulations on the residuals of this regression to factor out variance purely due to differences in duration between conditions. We used linear-mixed effects models with maximal random structure (Barr et al., 2013) and defined 3 planned contrasts to test our hypotheses. *Lexico-semantic* compared all conditions with intact content against NOWORDS to test the role of lexico-semantic information. *Prosody* compared the NOPITCH to the NAT condition to test the role of pitch information. *Agent* compared the NOPITCH and NAT conditions to the SYNTH condition, to test whether participants would perform as accurately with a synthesised voice as with a human voice. We report estimates, Wald t tests, and profile-likelihood 95% confidence intervals (CI) for fixed effects.[1] We also report likelihood-ratio tests based on model comparison to assess the overall effect of our manipulation.

Overall, Condition did not improve model fit ($\chi2(3)=6.31$, $p<.1$). However, we found that bias was larger in the NOWORDS condition than in all other conditions (B=273, SE=133, t=2.06, [5;546]), indicating that intact lexico-semantic information allows for more accurate (i.e., closer to zero) turn-end anticipation. In addition, there were no differences between NO PITCH and NAT (B=160, SE=110, t=1.45, [-163;497]), nor between these two conditions and the SYNTH condition (B=-154, SE=150, t=-1.03, [-353 ;48]).

Experiment 2: Gaze

In this experiment, the synthesized turns from Experiment 1 were uttered by a virtual agent (EMBR; Heloir & Kipp 2009) and accompanied by mutual and/or averted gaze. Specifically, the agent looked towards the listener as it started the turn. It then either maintained this listener-directed gaze throughout the turn (NEUT), or averted its gaze during speaking before looking back at the listener towards the end of the turn. The timing of this turn-final listener gaze occurred either at what we consider the 'natural' or optimal time point (NAT, about 600ms prior to turn end, as roughly observed in Kendon, 1967) or substantially earlier (EARLY, about 1600ms prior to turn end). Natural gaze was further contrasted with a gaze cue that had the same timing but remained averted (AVERT; Figure 1). Thus, the gaze cue manipulation affected both timing and direction. This served to investigate whether listener-directed speaker gaze needs to occur in a specific time window prior to turn end in order to be a reliable and efficient predictor for listeners, and/or whether gaze needs to be directed towards the listener at all.

In addition, we tested whether the effect of gaze would differ depending on Turn type. In the original conversation, participants could not see one another. We hypothesised the virtual agent's gaze would mostly affect the anticipation of turns that ended with an overlap or a gap, that is turns in which interlocutors did not achieve a smooth transition when information in only one modality (speech) was available to them.

Methods

Forty native speakers of German were paid 5 euros each to take part. None of them had taken part in Experiment 1. They reported no language, auditory or visual impairment.

We selected the subset of 56 turns pre-tested in Experiment 1 that were sufficiently long to allow our gaze manipulation (average bias in Experiment 1: -778 ms for the SYNTH version, -491 ms for the NAT version). Fifteen turns ended with a smooth transition in the original conversation (smooth), twenty-three were followed by a silent gap (pause), and eighteen ended with an overlap between speakers (overlap).

Four video clips were generated for each turn, one for each gaze condition (NEUT,NAT,AVERT,EARLY), using the EMBR framework (see http://dx.doi.org/10.7488/ds/234 for an example). We counterbalanced across turns whether the initial gaze movement away from the listener was to the right or to the left of the screen. We used transcriptions of the turns to obtain appropriate lip movements for the agent. Pauses were added to these transcriptions at those points where they appeared in the SYNTH turns (see Experiment 1). In a second stage, the SYNTH turn was superimposed on the video. Synchronisation with the agent's lip movements was ensured and gaze movements were aligned with the content of the audio file as requested by our gaze manipulation. This two-step procedure was necessary because the SYNTH turns from Experiment 1 were obtained using a different (and more flexible) synthesizer (MARY) than the one integrated within the EMBR framework.

Each movie clip started with the agent looking towards the participant for one second before starting to speak. The clips ended approximately one second after turn end. Bias was computed from the turn end rather than the end of the movie clip. The duration of the clips reached from a minimum of 4200ms to 14400ms with a mean duration of 7904ms. This includes one second of silence in the beginning of each video and the afore-mentioned silence after the turn. Overlap turns were on average longer (8490 ms) than pause (7584 ms) or smooth (7691 ms) turns. NAT and AVERT gaze cues appeared on average 569 ms before turn end (range: 166-1000 ms), while EARLY gaze cues were on average 1593 ms before turn end (range: 1200-2133 ms). Four lists were created using a Latin square design as in Experiment 1. Presentation order was fully randomised, individually for each participant using E-Prime 2.0. Participants watched the videos while listening to the turns presented over speakers. Each trial was preceded by a fixation cross that remained on screen for 2000 ms and followed by a 2000-ms ITI. Video playback stopped as soon as the participant gave a response to avoid learning effects. Every 18 trials participants paused for a short break. In total a session lasted about 20 minutes.

[1] CI were computed using the profile() function in R. When the CI could not be computed for the full random structure, we simplified the random structure. Fixed effects estimates in simplified models were very close to those in the full models. Likelihood ratio tests for individual planned contrasts also confirmed the reported results.

Results

Outliers higher than 5000 ms in absolute value (1.47% of the data) were discarded. Statistical analyses followed the same criteria as in Experiment 1. Turn duration was the same across conditions, so it was not necessary to regress it out. Three orthogonal contrasts were defined for the factor Gaze type. *Gaze shift* compared the NEUT condition against the conditions containing gaze shifts (NAT, EARLY and AVERT). *Gaze direction* compared the AVERT condition, in which the agent kept looking away from the participant, to the NAT and EARLY conditions, in which the agent established mutual gaze with the listener towards the end of the turn. Finally, *gaze timing* compared the NAT to the EARLY condition. For the factor Turn type, we defined two orthogonal contrasts: *overlap* compared overlap turns to the other two types; *gap* compared pause to smooth turns.

Across gaze types, the bias in Experiment 2 was closer to zero than in the pre-test (Table 3). Perhaps the mere presence of visual information about the agent, and the agent's lip movements, improved participants' accuracy. There were no differences between Gaze type conditions (χ2(3)=4.41, p=.22; all |t|<1).

Overall, the Gaze type by Turn type interaction did not contribute to model fit (χ2(6)=10.81, p<.1). However, the presence of a gaze shift influenced the bias for overlap turns more than it did for smooth or pause turns (*gaze shift * overlap* interaction: B=-392, SE=121, t=-3.25, [-665;-151]; see Figure 2). Importantly, this difference was not driven by differences in duration between turn types, as duration did not interact with Gaze type (χ2(3)=3.18, p=.37; all |t|<1.7).

We then analysed each turn type separately, Gaze type had no effect for either smooth or pause (all |t|<1). Instead, Gaze type influenced bias for overlap turns. Particularly, gaze direction (B=-83, SE=134, t=-.62, [-467;203]) and timing (B=-69, SE=142, t=-.48, [-464;206]) did not matter, but the mere presence of a gaze shift brought bias closer to zero than when the agent looked straight at the listener throughout the turn (B=354, SE=127, t=2.78, [111;584]).

Table 3: Mean bias (SE) in ms by Gaze type in Exp. 2.

Gaze type	Mean	SE
NEUT	-211	104
NAT	-147	89
EARLY	-178	90
AVERT	-62	107

General Discussion and Conclusion

In Experiment 1, listeners were more accurate in anticipating turn ends when they had access to the lexico-semantic content of the turns, but their performance was not impaired when the pitch contour was flattened or re-created using a text-to-speech synthesiser. This confirms that linguistic content is more important than intonation.

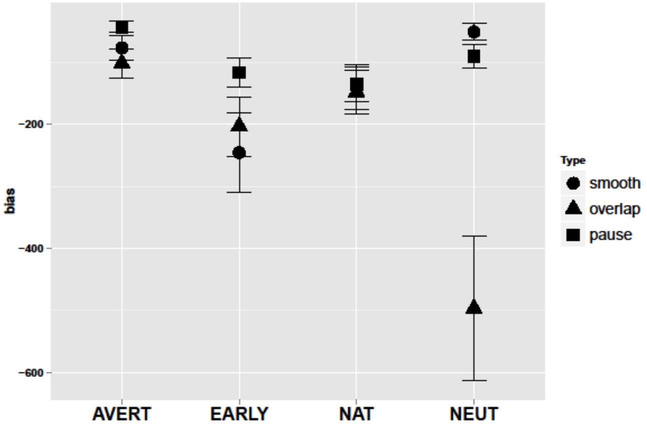

Figure 2: Mean bias (SE) in ms by Gaze and Turn type in Experiment 2

However, it is possible that timing information (such as the location and duration of pauses) is also necessary, in combination with content, for accurate turn end anticipation, since this information was preserved in both the flat pitch and the synthesised turns. Future studies using MARY could test this hypothesis by modifying the location of pauses.

Overall, Experiment 1 replicated De Ruiter et al.'s (2006) findings for our materials, and confirmed and extended Hjalmarsson's (2011) finding that turn end detection is possible for an artificial voice, at least when they are matched on prosodic contour and location and duration of pauses. It also served as a pre-test for the materials used in Experiment 2.

In Experiment 2, we introduced a virtual agent and manipulated the presence of gaze shifts, their direction, and timing. Surprisingly, direction and timing had little effect on the accuracy with which listeners anticipated turn endings, though the trends were in the expected direction: On the basis of Kendon's (1967) observations and Skantze et al.'s (2014) findings, one would have expected listeners to interpret an averted gaze as an indication that the agent was not done speaking, which should have led to a shift towards more positive bias in this condition. In addition, we expected gaze timing to affect anticipation, so that an early listener-directed gaze should have led to a shift towards more negative bias.

It is possible that participants did not interpret gaze cues in this way because they did not engage in a conversation with the agent. However, previous research has established that the turn end anticipation task reflects turn-taking behaviour in real conversations (De Ruiter et al., 2006). In addition, in all conditions (but NEUT) the number of head movements was the same. Participants might have taken the movement itself, or its initiation, as the cue rather than gaze direction.

Alternatively, our findings could be taken to suggest that gaze cues are not so relevant for the accurate (i.e., timely) anticipation of turn endings as they are for the decision whether the current speaker will keep or yield the turn. One possibility is that other sources of information are primarily used to estimate the right time for a potential turn switch, whereas gaze is only used as a secondary cue to help decide whether to take the turn or not when linguistic content indicates that a potential switch is upcoming (see De Ruiter et al., 2006 for similar reasoning about intonation).

Tentative support for this idea comes from our finding that the end of overlapped turns was anticipated more accurately when the agent averted its gaze than when it kept looking at the participant throughout the turn (in the NEUT condition). If overlapped turns contain early potential turn transition points, averted gaze might make participants less likely to think the speaker is about to yield the turn at an early stage.

In sum, our results support the assumption that prosody is less relevant for detecting a turn's end than lexico-semantic information. Timing and direction of speaker gaze as observed by Kendon (1967) do not *per se* improve accuracy.

Acknowledgments

This work was funded by the Cluster of Excellence ``Multimodal Computing and Interaction" of the German Excellence Initiative. We would like to thank Jeanin Jügler for her help with Praat scripting.

References

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68* (3), 255-278.

Bavelas, J., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication, 52* (3), 566-580.

Boersma, P., & Weenink, D. (2014). *Praat: doing phonetics by computer [Computer program]. Version 5.3.73*. Retrieved April 21, 2014, from http://www.praat.org/

Cutler, A., & Pearson, M. (1986). On the analysis of prosodic turn-taking cues. In C. Johns-Lewis (Ed.), *Intonation in discourse*. London: Croom Helm

De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82 (3), 515-535.

DeVault, D., Sagae, K., & Traum, D. (2011). Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse , 2* (1), 143-170.

Edlund, J., & Beskow, J. (2009). Mushypeek: A framework for online investigation of audiovisual dialogue phenomena. *Language and Speech, 52* (2-3), 351-367.

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language,* 25 (3), 601-634.

Gustafson, J., & Edlund, J. (2008). Expros: A toolkit for exploratory experimentation with prosody in customized diphone voices. In *Proceedings of Perception and Interactive Technologies for Speech-Based Systems* (pp. 293-296). Berlin/Heidelberg: Springer.

Heloir, A., & Kipp, M. (2009). EMBR - A realtime animation engine for interactive embodied agents. *In Proc. of the 9th Int'l Conference on Intelligent Virtual Agents.* Berlin Heidelberg: Springer.

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication,* 53 (1), 23-35.

Hjalmarsson, A., & Oertel, C. (2012). Gaze direction as a back-channel inviting cue in dialogue. *In IVA 2012 Workshop on Realtime Conversational Virtual Agents.*

Keitel, A., Prinz, W., Friederici, A., Hofsten, C., & Daum, M. (2013). Perception of conversations: The importance of semantics and intonation in children's development. *Journal of Experimental Child Psychology, 116* (2), 264-277.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica,* 26, 22-63.

Kohler, K. (1996). Labelled data bank of spoken standard German: The Kiel corpus of read/spontaneous speech. In K. J. Kohler (Ed.), *Proceedings of the Fourth International Conference on Spoken Language Processing* (pp. 1938-1941).

Levinson, S. (2006). Roots of human sociality. In N. Enfield, & S. Levinson (Eds.), *Culture, cognition and human interaction*. Oxford: Berg Publishers.

Magyari, L., & De Ruiter, J. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in psychology*, 3, doi: 10.3389/fpsyg.2012.00376.

Magyari, L., Bastiaansen, M., De Ruiter, J., & Levinson, S. (2014). Early anticipation lies behind the speed of response in conversation. *Journal of Cognitive Neuroscience*, 26 (11), 2530-2539.

Riest, C., Jorschick, A., & De Ruiter, J. (2015). Anticipation in turn-taking: Mechanisms and information sources. *Frontiers in Psychology, 6*, doi: 10.3389/fpsyg.2015.00089.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 4 (1), 696-735.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime User's Guide.* Pittsburgh: Psychology Software Tools Inc.

Schröder, M., & Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *Int. Journal of Speech Technology, 6*, 365-377.

Skantze, G., Hjalmarsson, A., & Oertel, C. (2014). Turn-taking, feedback and joint attention in situated human-robot interaction. *Speech Communication,* 65, 50-66 .

Torres, O., Cassell, J., & Prevost, S. (1997). Modeling gaze behavior as a function of discourse structure. *First International Workshop on Human-Computer Conversations.*