Atten Percept Psychophys (2018) 80:27–41
https://doi.org/10.3758/s13414-017-1428-0

Knowing when to respond: the role of visual information in conversational turn exchanges

Nida Latif[1] · Agnès Alsius[2] · K. G. Munhall[2]

Published online: 27 October 2017

Abstract When engaging in conversation, we efficiently go back and forth with our partner, organizing our contributions in reciprocal turn-taking behavior. Using multiple auditory and visual cues, we make online decisions about when it is the appropriate time to take our turn. In two experiments, we demonstrated, for the first time, that auditory and visual information serve complementary roles when making such turn-taking decisions. We presented clips of single utterances spoken by individuals engaged in conversations in audiovisual, auditory-only or visual-only modalities. These utterances occurred either right before a turn exchange (i.e., 'Turn-Ends') or right before the next sentence spoken by the same talker (i.e., 'Turn-Continuations'). In Experiment 1, participants discriminated between Turn-Ends and Turn-Continuations in order to synchronize a button-press response to the moment the talker would stop speaking. We showed that participants were best at discriminating between Turn-Ends and Turn-Continuations in the audiovisual condition. However, in terms of response synchronization, participants were equally precise at timing their responses to a Turn-End in the audiovisual and auditory-only conditions, showing no advantage of visual information. In Experiment 2, we used a gating paradigm, where increasing segments of Turns-Ends and Turn-Continuations were presented, and participants predicted if a turn exchange would occur at the end of the sentence. We found an audiovisual advantage in detecting an upcoming turn *early* in the perception of a turn exchange. Together, these results suggest that visual information functions as an early signal indicating an upcoming turn exchange while auditory information is used to precisely time a response to the turn end.

Keywords Conversational turn-taking · Multisensory processing · Perception and action · Visual perception

✉ Nida Latif
nida.latif@mcgill.ca

[1] Department of Psychology, McGill University, Stewart Biology Building N6/7, 1205 Dr. Penfield Avenue, Montreal, QC H3A 1B1, Canada

[2] Department of Psychology, Queen's University, Humphrey Hall 307, 62 Arch Street, Kingston, ON K7L 3N6, Canada

Introduction

Every day, we engage in conversations that proceed effortlessly; we hardly notice how seamlessly we switch roles from talker to listener with our conversational partner. The most fascinating property of this organized system of turn-taking behavior is how quickly and efficiently it proceeds. This speed and effectiveness is even more remarkable when we consider the sheer volume of information processing that occurs during conversation. During face-to-face conversation, conversational partners both produce and perceive a myriad of auditory and visual informational sources about communication that indicate whether and when a turn is about to end (for review, Ford & Thompson, 1996; Fox Tree, 2000). The rapid use of multiple cues indicating an upcoming turn exchange is a critical part of communicative interaction; it is still unclear, however, how the different sources of information contribute to the perceptual and planning processes involved in turn-taking behavior.

During conversation, both speakers and listeners work collaboratively to ensure successful turn-taking behavior. Within the auditory domain, speakers produce several cues to demonstrate that they are ready to give up the conversational floor (for review, Levinson, 2016; Thórisson, 2002). For example, they use specific lexical items and syntactic constructions to indicate their contribution is coming to an end (for review, Ford & Thompson, 1996). This lexico-syntactic information indicating a turn-end is emphasized by accompanying

prosodic cues, such as a drop in pitch and intensity and the lengthening of the final words of a sentence (Duncan, 1972; Gravano & Hirschberg, 2011). In return, listeners are able to perceive these cues and use them efficiently to accurately determine an upcoming turn exchange. For instance, de Ruiter, Mitterer & Enfield (2006) showed that perceivers are more accurate at determining when a turn will end using auditory information that has been unmodified compared to when it has been low-pass filtered to minimize lexico-syntactic cues. Further, prosodic information such a falling pitch contour has been shown to enhance a perceiver's turn exchange detection compared to when the pitch has been flattened (Riest, Jorschick & de Ruiter, 2015).

Considering the visual domain, there are two categories of visual cues that comprise communicative interactions. On the one hand, there is 'linguistic' visual information that directly reflects the content of speech (e.g., lip movements; Grant & Seitz, 2000). On the other hand, there are nonlinguistic visual cues that are supplemental to information communicated through the speech channel (e.g., gestures, gaze; Cassell, McNeill & McCullough, 1998; Kennington, Kousidis & Schlangen, 2013). There is much evidence to support that the presence of visual information within both these categories significantly benefits communication. For example, having access to visual information from a speaker's face (including lip movements) enhances intelligibility (Sumby & Pollack, 1954), while gestural information aids in the completion of joint tasks with a partner (Kraut, Fussell, and Siegel, 2003). In the specific context of conversational turn exchanges, both speakers and listeners produce nonlinguistic visual cues indicating that a turn is coming to a close. Speakers reduce their overall body motion, position themselves towards the listener and establish joint gaze (Bavelas, Chovil, Coates & Roe, 1995; Kendon, 1972; Thomas & Bull, 1981; Ho, Foulsham & Kingstone, 2015). Conversely, listeners make preparatory head movements and maintain direct eye contact to demonstrate their desire to contribute to the conversation (Kendon, 1967; Hadar, Steiner, Grant & Rose, 1984). Several studies have demonstrated that visual information is certainly important for maintaining the conversational turn exchange structure. For example, it has been shown that speakers can utilize direct gaze to induce a response from the listener (Kendon, 1967).

Specifically considering the use of these cues for the *perception* of turn exchanges, only two known studies explicitly examining natural conversation have provided support for the importance of visual information in turn perception (Valtersson & Torreira, 2015; Latif, Alsius & Munhall, 2017). Valtersson & Torreira (2015) trained a statistical model to classify turn-ends and turn-continuations by including different auditory (i.e., intensity, pitch and speech rate) and visual (i.e., gaze, hand gestures, and head movements) cues. A subset of the turn-ends and turn-continuations were first used to train the model to identify turn type. Then, the model classified a new set of turn-ends and turn-continuations based on that training. Their findings showed that performance in identifying turn-ends versus turn-continuations could be (modestly) improved by including both visual and auditory information when training the model. In another study conducted in our laboratory, we provided further experimental support for the finding that visual information can improve turn exchange perception (Latif, Alsius & Munhall, 2017). We examined the role of audiovisual information in turn perception accuracy by presenting participants with clips of turn-ends and turn-continuations in auditory-only, visual-only or audiovisual information contexts. It should be noted that the full range of auditory and visual information was examined (i.e., complete auditory and visual channels) rather than selecting specific cues within each modality to examine the influence of information availability in general. We found that participants were more accurate at identifying a turn-end in the audiovisual condition compared to both unimodal conditions, suggesting the addition of visual information enhances the accuracy of turn-taking perception.

Together, these studies suggest that the visual channel provides the perceiver with a range of visual cues that can be effectively used—in combination with auditory cues—to detect upcoming turn exchanges. However, efficient interaction is not determined simply by knowing that a turn exchange will occur (i.e., accuracy) but also *when* it will occur so as to respond at the appropriate moment.

The ability of conversational partners to maintain appropriate timing during conversation ensures that interactions are well-coordinated and proceed successfully, with minimal interruptions or long silences. Previous research has shown that, on average, inter-turn gaps are less than 200 ms with a mode of 0 ms. This strong tendency to minimize the length of time between turns during conversations is consistent across several languages (Stivers, Enfield, Brown, Englert, Hayashi, Heinemann et al., 2009), and results from the listener's ability to anticipate the occurrence of an upcoming turn event. In a task where listeners were presented with sentences before a turn exchange and were asked to make a button-press response at the exact moment that the turn would end, they did so more than 200 ms earlier than the turn-end (i.e., when the current talker would have stopped speaking; de Ruiter et al., 2006), indicating that they were anticipating the appropriate moment to make their button-press response and not simply reacting to the cessation of the talker's speech.

The ability to anticipate events to plan appropriately timed responses is not specific to conversational interaction. Predictive mechanisms based on perception-action coupling (i.e., the reciprocal relationship between perceiving sensory information and producing an appropriate response; Warren, 1990) have previously been used to explain our general ability

to estimate the timing of events. Such a mechanism uses available sensory information to allow for planning of motor responses as perceptual events unfold (Schubotz, 2007). Our ability to anticipate upcoming events is essential for ensuring that we respond appropriately in all kinds of real-world situations: from avoiding collisions (Tresilian, 1995) and performing sports-related activities such as hitting a ball with a bat (Raganathan & Carlton, 2007) to facilitating social activities such as joint collaborative tasks (e.g., carrying objects with another person) and orchestrating musical performances (Pfordresher, 2006). The association of conversational anticipation and general perceptual-motor prediction is supported by clinical evidence. Individuals with social disorders characterized by deficits in conversational skills have been shown to exhibit impairments in general prediction abilities, suggesting a common underlying mechanism (Sinha, Kjelgaard, Gandhi, Tsourides, Cardinaux, Pantazis et al., 2014).

Little is known about how the availability of auditory and visual information during conversation influences the production of appropriately timed responses. Certainly, a number of studies have shown that auditory and visual cues also are important for tracking and defining the temporal structure of turn-taking behavior. However, most previous studies have only focused on the influence of auditory cues on response timing. For example, participants were less precise at timing a button-press response to the end of a turn using auditory degraded information (de Ruiter et al., 2006) compared to when the auditory signal was not manipulated, demonstrating that the availability of lexico-syntactic information influences response timing when perceiving turn exchanges. Similarly, some studies have shown that visual cues also are important for the temporal organization of turn-taking behavior. For example, listeners anticipate the end of a turn by shifting their gaze to the upcoming speaker before they have finished speaking (Tice & Henetz, 2011; Holler & Kendrick, 2015). Further, Stivers et al. (2009) demonstrated that a speaker's direct gaze during a turn exchange is associated with faster responses from a listener. The importance of visual information in turn exchange timing is especially evident when we consider hearing-impaired perceivers who communicate solely through visual means; fluent users of sign language demonstrate incredible similarity to those who communicate audiovisually in the distribution of response times during turn exchanges (de Vos, Torreira & Levinson, 2015). This suggests that the mechanism involved in the timing of turn-taking behavior generalizes to communication in general. Given such evidence, it is surprising that no studies have explicitly examined how visual conversation cues influence timing a response to the end of a turn.

In the current studies, we examined the contribution of visual and auditory information when anticipating an upcoming turn exchange (i.e., when a talker is about to finish their turn). It should be noted that in these studies (like Latif et al., 2017), we examined the visual and auditory channels in natural contexts. In other words, we did not isolate specific linguistic or nonlinguistic visual and auditory cues; rather, we were interested in examining the general influence of having the visual and auditory channels available when making predictions about turn exchanges during conversation. In Experiment 1, we used a synchronization task to explore, for the first time, whether the presence of visual information influences our ability to accurately time a response to the end of a turn. In Experiment 2, we examined the relative timing of the use of visual and auditory information.

Experiment 1

In this experiment, we used a synchronization task to examine the timing accuracy of predicting the end of a turn when the stimulus was presented in different modalities. Participants were required to use a button-press to respond precisely to when they believed that the current talker (Talker A) would finish their turn. Previous studies have shown that synchronizing a response to a non-speech perceptual event (e.g., the endpoint of a moving object) requires anticipatory timing processes, and participants vary in how accurately they are able to respond depending on whether auditory or visual information is present (Aschersleben & Prinz, 1995; Chen, Repp & Patel, 2002; Miyake, Onishi & Pöppel, 2004; Repp & Su, 2013). In the context of turn exchange perception, it has been shown, using auditory-only information, that participants indicate the end of a turn prior to when the current speaker actually stops speaking (de Ruiter, et al., 2006). However, no known studies to date have examined the online perception of visual information and how it might compare to auditory-only perception when responding to turn exchange behavior. It is known that the presence of visual cues leads to faster responses by a listener during a turn exchange, suggesting that visual information influences the temporal structure of conversation (Stivers et al., 2009). Here, we utilized a similar paradigm to that used by de Ruiter et al. (2006), who asked participants to use a button-press response to indicate when turns would end. Their stimuli consisted of different kinds of turn-ends including some that contained multiple syntactically possible points of completion. These were stimuli that contained a point at which the turn exchange occurred, however, prior to the end of the turn, there were several possible points that an observer might consider a turn-end (e.g., 'I have it officially, is it still mine?' In this example, the turn could syntactically end either after the word 'officially' or after the word 'mine'; de Ruiter et al., 2006). Participants were required to anticipate the point at which the turn would truly end. In contrast, in our experiment, we included trials that were explicitly identified as Turn-Continuations (i.e., a single sentence after which the same talker would continue speaking) with a single possible

point of completion. Using this method, participants were required to make a choice of whether or not to respond, resulting in a single explicit correct or incorrect response. In addition, we controlled for the types of turns that were included by imposing specific criteria on stimulus selection.

Method

Stimuli

In this experiment, 24 pairs of same-gender friends (mean age: 20.86 years; 18 female pairs; all native English speakers with no hearing/speech impairments) were video-recorded while engaging in an unstructured conversation for 10 min. From each conversation, five turn exchanges (i.e., Turn-Ends) were selected such that no exchange contained interruptions or verbal backchanneling behavior and no exchange was a question-answer sequence. These criteria were selected to minimize the influence of unmanipulated factors such as processing differences in various categories of turns and thus reduce unwanted noise. First, it is known that responses to questions vary in timing depending on the complexity of the upcoming response (Casillas, Bobb & Clark, 2016; Stivers et al., 2009). Therefore, we eliminated question-answer sequences in our stimuli. Second, previous studies have shown that backchannelling behavior is considered a signal to the speaker to continue speaking (Duncan, 1972; Cassell, Torres & Prevost, 1999) and thus would potentially introduce cues indicating a 'Turn-Continuation' within our 'Turn-End' stimuli, creating unwanted noise in our stimuli. We controlled for such noise by excluding instances of backchannelling responses in *both* our turn types.

A total of 120 Turns (24 pairs × 5 Turn-Ends) were selected. Here, we were interested primarily in the temporal accuracy with which participants timed their responses to coincide with the end of a turn. However, we wanted to ensure that participants were attempting to discriminate between Turn-Ends and Turn-Continuations, as would be the case in natural conversation. Therefore, we included 24 Turn-Continuation stimuli (one from each pair) as catch trials using the same criteria used to identify Turn-Ends. Auditory-only (AO) and visual-only (VO) versions were created from the audiovisual (AV) version of each stimulus by removing the visual and auditory information, respectively (Fig. 1).

Once the Turn-Ends and Turn-Continuations were identified, Turn-End stimuli were edited to 501 ms [15 frames at 29.97 frames per second (fps)] past the start of the second talker's (i.e., Talker B) speech. Turn-Continuation stimuli were edited to 501 ms past the start of the next sentence spoken by the same talker (i.e., Talker A). All edits were determined using acoustic events indicating the start of speech[1]. The length of the resulting clips ranged from 901 ms to 5100 ms. Limiting the clip's end to 501 ms past the acoustic onset of Talker B's speech was appropriate because we were interested in looking at predictive behavior (i.e., how well people were able to anticipate an upcoming turn and synchronize their response), not reactive responses. Further, previous studies examining anticipatory responses to auditory turn exchanges using a similar paradigm have shown that participants made button-press responses significantly earlier than the end of the turn (e.g., de Ruiter et al., 2006).

Experimental equipment

The experiment was conducted in a single-walled sound booth. All video stimuli (Resolution: 800 × 600 at 29.97 fps) were displayed using DMDX software (Forster & Forster, 2003) and the audio signal was played from speakers (Paradigm Reference Studio/20) positioned on either side of the monitor. Participants were seated approximately 57 cm away from a flat CRT monitor (48.26 cm; Daewoo CMC901D). Due to the time-sensitive nature of this synchronization task, participants responded using a mechanical keyboard to ensure timing accuracy[2] (Corsair STRAFE w/ Cherry MX Red Mechanical Switches).

Participants

Twenty-four undergraduates (mean age: 21.48 years; 22 females) participated in this experiment for monetary compensation or course credit. All participants were native English speakers with no hearing or speech impairments and with normal or corrected-to-normal vision.

Procedure

Prior to the experiment, participants were given 12 practice trials (two Turn-Ends and two Turn-Continuations in each of the three modalities, AV, AO and VO) with feedback, to get familiarized with the task. In the experiment, they were then presented with 120 Turn-Ends trials and 24 Turn-Continuation trials in three modality-specific blocks of 48 trials each (i.e., one block for each of the AO, AV and VO conditions with 40 Turn-Ends and 8 Turn-Continuation trials in each block). Assignment of stimulus clips to each modality

[1] The auditory component was selected because it is more precise at indicating the end of a turn. However, an examination of how the auditory end compared to the visual end showed only a small difference of 0.67 frames (or 22.36 ms at 29.97 frames per second (fps) on average.

[2] Mechanical switches ensure response rate of 1 ms compared to 20 ms for other generic keyboards (Corsair, 2016; Plant, Hammond & Whitehouse, 2003).

Fig. 1 Schematic of stimuli presented in the three modality conditions

condition was counterbalanced such that all clips were presented in each modality equally often across all participants and no clip was repeated for a single participant. Block order was also counterbalanced across participants and stimulus presentation within blocks was randomized.

A fixation cross presented for 500 ms began each trial, followed by the presentation of a stimulus. Participants were asked to indicate by pressing a button the exact moment at which they believed the first talker would finish their turn. They were encouraged to make a button-press response as close to the real end of the turn as possible. Participants were also informed that some trials would not contain a turn-end and instead, the current talker would continue speaking. In this case, participants were instructed not to make a button-press response. It should be noted that participants were not informed that there was an uneven number of Turn-Ends and Turn-Continuations. Further, participants were not provided with any feedback regarding their accuracy in identifying the two turn types.

Stimuli were presented until either the participant made a response or until the end of the clip (i.e., 501 ms past the start of Talker B's speech, or Talker A's second sentence). Any instance where a button-press response was not made within the stimulus presentation time was a considered a 'Turn-Continuation' response. How well participants could respond to the end of a turn was defined by a measure of response time offset (RTO), which was calculated as participants' button-press response times subtracted from the time corresponding to the end of Talker A's speech (as determined by the auditory signal) for each item.

Results and discussion

Although we were most interested in button-press response timing behavior to Turn-Ends, we still wanted to measure overall turn identification accuracy to ensure that participants could indeed discriminate between the turn types (i.e., Turn-Ends versus Turn-Continuations) as we demonstrated in our previous study (Latif et al., 2017). To do this, we performed an initial analysis of participants' overall ability to discriminate between Turn-Ends and Turn-Continuations by calculating d'. d' is a signal-detection method that analyzes a perceiver's ability to discriminate between 'signal' and 'noise' by calculating a ratio of the perceiver's Hit Rate (i.e., the proportion of time participants reported perceiving the signal when the signal was indeed presented) and False Alarm rate (i.e., the proportion of time participants reported perceiving the signal when, in fact, they were presented with noise). Here, 'Turn-Ends' were arbitrarily assigned as the 'signal' and 'Turn-Continuations as noise though, because d' is a ratio, the same values would be found if this designation was reversed. This analysis is especially useful since it is robust against unequal number of 'signal' and 'noise' trials, as was the case in our experiment.

On average, 20% (SE = .014; AO = 22% of trials, AV = 13% of trials, VO = 25% of trials) of Turn-Ends were not responded to, resulting in the analysis being conducted on 80% of all Turn-End trials[3]. We compared participants' ability to discriminate Turn-Ends from Turn-Continuations by comparing average d' scores against 0 (i.e., chance performance). Participants were able to distinguish Turn-Ends from Turn-Continuations significantly more than chance in all three modalities (AO d' = 1.73, $t(23) = 15.76$; AV d' = 2.57, $t(23) = 20.50$; VO d' = 1.63, $t(23) = 19.35$; all $P < .001$). Further, a repeated measures analysis of variance (ANOVA) showed that participants were better at distinguishing Turn-Ends from Turn-Continuations in the audiovisual condition compared to both unimodal conditions [main effect of modality: $F(2,69) = 12.50$, $P < .001$; AV–AO mean difference = 0.84, SE = .24, $P = .006$; AV–VO mean difference = .94 , SE = .20 $P < .001$; both P-values were Bonferroni corrected[4]]. No significant differences were found between the AO and VO conditions.

To examine our main variable of interest (i.e., button-press response timing to the end of a turn), we preformed our main analysis on RTOs for the Turn-Ends trials only. False alarm responses to the Turn-Continuations were discarded. Since RTOs were the difference between participants' button-press response and the end of the current turn, negative RTOs

[3] A similar ratio of false alarms occurred on the Turn-Continuation trials. Participants on average pressed the button in 21% of Turn-Continuation trials (AO = 24%; AV = 15%; VO = 23%).

[4] Significance were determined by assessing whether our raw P-value passed the Bonferroni-corrected criterion.

indicated that participants responded before the end of Talker A's contribution, while positive RTOs indicated that the button-press response was made after Talker A had finished their turn.

The distributions for RTOs in the three modalities are presented in Fig. 2 (AO: Median = 144.00 ms; AV: Median = 147.78 ms; VO: Median = 175.41 ms). Overall, the results show that participants were able to synchronize their button-press responses to the end of the turn with peaks around 0 ms.

To analyze how modality contributed to participants' ability to time a button-press response, we performed a statistical analysis that allowed us to take both item and subject variance into account (Clark, 1973; Baayen, Davidson & Bates, 2007; Brysbaert, 2007). We fitted a regression model to analyze how participants' RTOs were predicted by the modality in which information was presented. A linear mixed model (LMM) was implemented using SPSS's Mixed procedure (LMM: SPSS 24). The model was fitted using restricted maximum likelihood estimation (REML). Modality was included as a fixed factor with participants and items included as random factors. All possible slopes were included and an unstructured variance structure was specified. The dependent variable was RTO[5]. Results revealed that Modality significantly predicted RTOs [$F(2) = 9.89$, $P < .001$]. When participants responded to AV stimuli and AO stimuli, their RTO was significantly shorter than when they responded to the VO condition (AV–VO: $t = 4.45$, $SE = 23.59$, $P = < .001$; AO–VO: $t = 2.35$, $SE = 24.50$; $P = .01$). No significant differences between the AO and AV condition were found ($P = .13$). In other words, participants more frequently responded closer to the point of turn exchange in the AV and AO conditions compared to the VO condition [M(VO) = –79.12 ms, SE = 27.63; M(AV) = 23.06 ms, SE = 17.49, M(AO) = –6.78, SE = 19.70; Fig. 3). Further, we compared RTOs in each condition to determine how well participants could synchronize their responses with the end of the turn. To do so, we ran one-sample t-tests for each modality condition comparing means against an RTO of 0 (i.e., completely synchronous response). We found no significant differences from 0 for the AV and AO conditions (AO: $P = .73$; AV: $P = .19$). Button-press responses in the VO condition were significantly different from 0 ($t = –2.86$, $P = .004$). Taken together, it was evident that button-press responses to visual information alone were less accurate in timing, and deviated further from the end of the turn.

In addition to examining how accurately participants could make a button-press response to the end of a turn, we were interested in how consistently they were able to make that response. To do so, we examined participants' standard deviations (i.e., variability in responses) as a measure of response precision between modality conditions. We conducted a repeated-measures ANOVA on the participants' standard deviations for their RTOs. Results showed a significant main effect of modality [$F(1,23) = 15.92$, $P < .001$). Bonferroni-corrected pairwise comparisons revealed significant differences between both the AV and the AO conditions compared to the VO condition (AV–VO: mean difference = 201.80, SE = 36.72, $P < .001$; AO – VO: Mean difference = 174.82, SE = 38.06, $P < .001$). No significant differences were observed between the AV and the AO conditions ($P = .99$; Fig. 4). This pattern of findings directly reflects and verifies the findings reported for our main regression analysis. These results suggest that, without auditory information, participants are much less precise in their button-press responses and that audiovisual information provides no advantage to response timing precision.

In this experiment, we demonstrated that participants responded anticipatorily to the end of a turn, regardless of modality. However, participants' accuracy in timing a button-press response to the end of a turn was influenced by the modality in which they received turn exchange information. In general, participants were more frequently accurate at timing their response when they had auditory information available (i.e., both auditory-only and audiovisual conditions) than with visual information only.

In addition to timing accuracy, we included a measure of response precision by examining standard deviations to address the nature of our button-press response distributions. As is evident in Fig. 2, the distributions of responses were negatively skewed. Thus, using multiple measures, including the spread of responses, would better characterize button-press response times as a function of modality in our task. We found that participants' response precision reflected the same pattern of findings as their accuracy; participants were more precise in their button-press responses (i.e., smaller standard deviations) in the audiovisual and the auditory-only conditions and much less precise in the visual-only condition. In general, participants responding to visual-only stimuli were prone to making more timing errors indicated by the greater frequency of early responses. It is possible that early button-press responses in the visual-only condition were a result of being misled by visual cues that are available early during a turn. For example, it has been shown that although overall postural changes mark the end of a talker's contribution, *upper body* postural changes, specifically, mark the beginning of a turn (Cassell, Nakano, Bickmore, Sidner & Rich, 2001).

[5] The length of the clip was not included as a factor in the model due to the difficulty of controlling for length of utterances when obtaining stimuli from natural conversation. However, participants' RTO was correlated with length of clip using Pearson's *r* in order to examine whether clip length had a significant effect on performance. All three modality conditions presented a significant negative correlation between RTO and length of clip (AO: r = –.42; AV: r = –.46; VO: r = .56); that is, the longer the presented turn, the quicker the speed of the response. Importantly, no significant difference in correlations were observed between modality conditions. It should be noted that negative correlation between stimulus presentation and response time is a common finding in perception response time studies (e.g., Niemi & Näättänen, 1981).

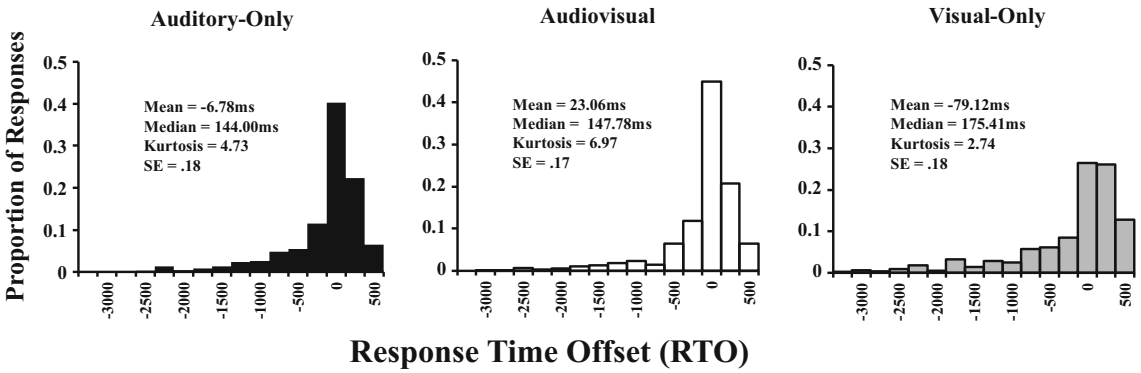

Fig. 2 Distributions for response time offsets (RTOs) for the three modality conditions across all participants. In general, participants were able to time their response to the end of the turn; however, the distribution for RTOs for the visual-only (VO) condition had a flatter peak as indicated by the kurtosis value

Without auditory information supplementing the visual information, some subtly different early cues may have been misattributed to Turn-End cues in the visual-only condition leading to more frequent erroneously timed button-press responses.

Further, greater variability in responses to the visual-only condition is, in fact, consistent with results of studies investigating general prediction abilities using sensorimotor synchronization paradigms. It has been demonstrated that synchronizing a button press with a visually presented stimulus (i.e., a flash) results in greater variability in anticipatory responses compared to auditory stimuli (Chen et al., 2002). Further, participants are more likely to respond too early when making a synchronization response to a visual stimulus versus an auditory stimulus (Rosenblum, Gordon & Wuestefeld, 2000; McLeod & Ross, 1983). As can be seen in the distributions shown in Fig. 2, the visual condition shows greater spread and a thicker negative tail indicating higher probability of large anticipation responses. It is possible that responding early to visual information is adaptive; failure to respond to a visual cue on time in the real-world can have potentially harmful consequences (e.g., timing the arrival of a moving object to avoid harm). The fact that similar behavioral responses to perceptual information are found even in conversations, where priorities are not primarily visual, is suggestive of the involvement of a common underlying predictive mechanism.

It should be noted that although we show anticipatory button-press responses to stimuli, our results differ from previous turn exchange perception studies. Most studies investigating turn timing report anticipatory responses to auditory turn exchanges that occur significantly prior to the end of the turn using tasks similar to the one used here (e.g., 200 ms in advance; de Ruiter et al., 2006, Magyari & de Ruiter, 2012). That is, although our results show a similar response distribution to previous work with many early responses that occur prior to the turn end, on average, our button-press responses are not made as early as those reported in these studies. One potential reason that might

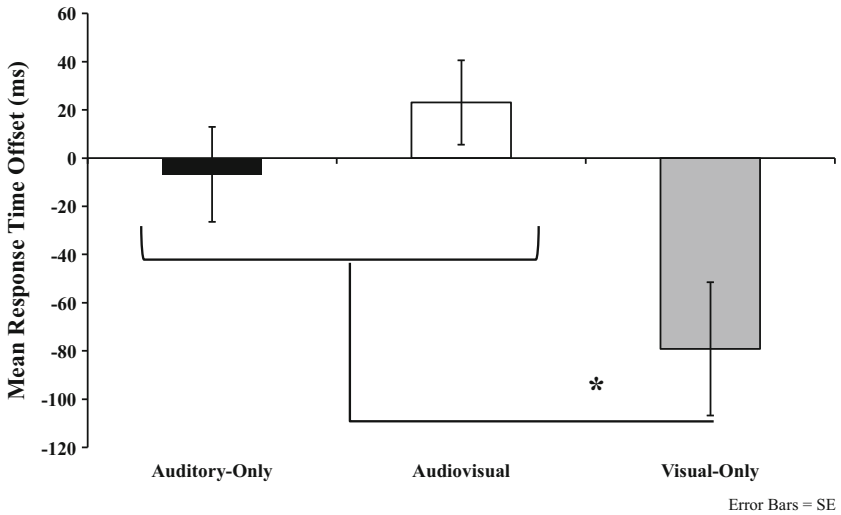

Fig. 3 Average RTO. Participants responded significantly earlier in the visual-only condition compared to both the Auditory-Only (AO) and the Audiovisual (AV) conditions. No differences between the AO and AV conditions were found

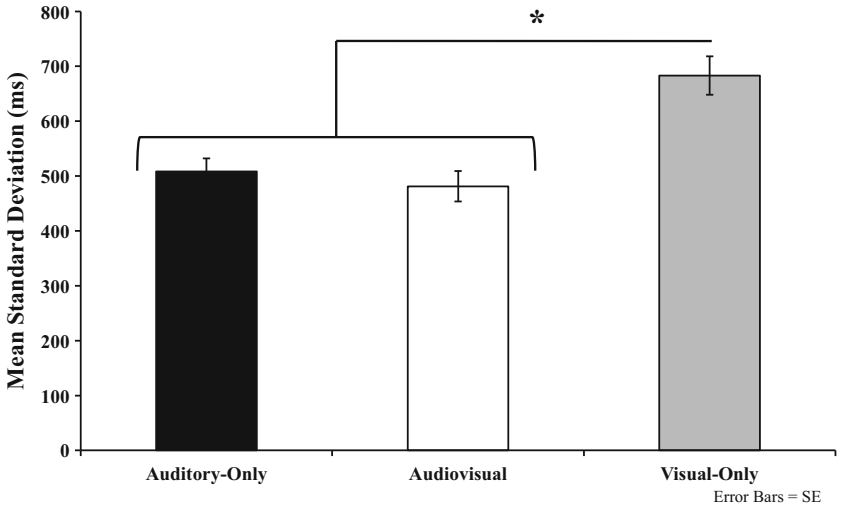

Fig. 4 Average standard deviation of RTOs. A greater variability in RTOs was observed in the VO condition compared to both the AO and AV conditions

explain the varying outcomes in our study compared to others is that participants had different specific task demands. In our study, participants were presented with two types of trials (i.e., Turn-Ends and Turn-Continuations) and were required to either respond or withhold a response. In other studies, participants were only presented with Turn-Ends that contained several possible points of completion. Thus, in the previous studies, participants were aware they needed to respond to every trial and only needed to track the timing of that response. In our study, however, participants needed to engage in an additional task by tracking *both* whether to respond and when to do so. It is well established that distributing attentional resources in a dual-task paradigm leads to slower response timing than a single task (for review, see Pashler 1994). It is, therefore, possible that the difference in task demands between our study and previous studies resulted in slower responses times for our participants.

Further, there is a large variability in inter-turn timing that is often not emphasized in the literature. Although talkers, on average, do not leave gaps between turns, the distributions, in fact, range from –600 ms to 1000 ms (Stivers et al., 2009). This is likely due to the fact that the types of turns and other contextual factors influence inter-turn times. For instance, it has been shown that the time to respond to complex versus simple questions and the emotional valence of the upcoming response influence how quickly turn exchanges occur (Casillas et al., 2016; Kendrick & Torreira, 2014). Thus, combining many types of turns (e.g., question-answer sequences, Yes-No responses, overlapping responses; de Ruiter et al., 2006), influences how the perceivers' button-press response times can be interpreted. Rather than average across all these other sources of variance, we imposed a stricter criterion on selecting our Turn-Ends allowing us to better control for variability due to turn type.

The results of Experiment 1 show that the modality in which information is perceived influences the accuracy of participants' responses to the end of a turn with the auditory and audiovisual modalities showing similar response precision. However, the fact that we observe an audiovisual advantage in our preliminary check of participants' ability to distinguish between Turn-Ends and Turn-Continuations is noteworthy. This audiovisual advantage in distinguishing turn types has been shown previously (Latif et al., 2017). The results of current experiment show that the audiovisual advantage in accuracy was not reflected in participants' ability to time their button-press response to the end of a turn, since participants performed equally in terms of timing in the audiovisual and auditory-only conditions. This raises the possibility that auditory and visual information may serve complementary roles; perhaps visual information contributes to early (and more accurate) signaling of an upcoming turn exchange while auditory information is needed to precisely time a response.

The idea that different cues (e.g., linguistic, non-linguistic, visual, auditory) might make distinct yet complementary contributions to turn perception as the message from the speaker unfolds is not new. Studies examining the contribution of auditory cues in turn anticipation have shown, for instance, that syntactic and lexical information allow for a listener to anticipate the end of a turn earlier than prosodic cues (de Ruiter et al., 2006). While prosodic cues, such as falling pitch, can provide information that a turn is about to end, it only does so in the final syllables (Bögels & Torreira, 2015). Outside the verbal domain, it has been shown that several visual cues denote different points of a turn. For example, prominent gestures (Rickel & Johnson, 2000), upper body postural change (Cassell, et al., 2001) and head turns (Cassell, Bickmore, Billinghurst, Campbel, Chang, Vilhjalmsson & Yan, 1999) are often associated with the beginning of a talker's contribution, while whole body postural changes (Cassell et al., 2001),

eyebrow raises and head nods (Cassell et al., 1999) are associated with Turn-Ends. How perception of turn exchanges may be influenced by the distribution of visual information has yet to be determined.

One method to study the timing of the use of perceptual information (both auditory and visual) is the gating paradigm, which has been used successfully in speech perception studies (e.g., Grosjean, 1980; Munhall & Tohkura, 1998). In this methodology, participants are presented with successively longer segments of an utterance in order to determine the durational threshold at which information is sufficient to accurately identify the stimulus before its completion (Grosjean, 1980). Using this paradigm, it has been shown that visual information plays an important role *early* in speech perception; when single words are presented audiovisually, perceivers are able to identify them much earlier in the gating procedure than with unimodal information (Jesse & Massaro, 2010). Overall, these studies demonstrate that auditory and visual cues contribute to the perception of an audiovisual stimulus in a complementary manner as it unfolds. Perhaps even in the case of turn exchange perception, there are differences in when conversational auditory and visual cues can reliably indicate an upcoming turn end.

In Experiment 2, we investigated whether auditory and visual information varied in when, over the course of a turn, they could reliably indicate an upcoming turn exchange.

Experiment 2

In this experiment, we examined the individual and shared contributions of visual and auditory cues over the course of an utterance leading up to a conversational turn exchange. A gating paradigm was used where participants were presented with different amounts of an utterance prior to the point of turn exchange. Classic studies using gating procedures presented participants with increasingly larger successive segments of a word and showed that information about a stimulus is accrued incrementally until a threshold is reached where it can be identified (Grosjean, 1980). Here, we presented different lengths of a sentence prior to the point of turn exchange (or prior to the start of the next sentence for Turn-Continuations), in order to examine at which point in time participants could reliably identify an upcoming turn exchange as a function of the available information. Turns were presented in audiovisual, auditory-only or visual-only modalities.

In addition to investigating when participants would be able to accurately indicate an upcoming Turn-End, we were also interested in investigating at which point participants could accurately identify when a turn would continue (i.e., the same talker would continue speaking). It is possible that the information and the point at which that information can be used differs depending on whether a perceiver is identifying a Turn-End versus a Turn-Continuation, and thus both these turn exchange events were included in our gating paradigm.

Method

Stimuli

The Turn stimuli used in this experiment are the same as those used in Experiment 1 (and Latif et al., 2017). However, here we were interested in determining when over the course of a turn perceivers could distinguish between a Turn-End and a Turn-Continuation. Therefore, we included an equal number of Turn-End and Turn-Continuation trials for a total of 120 Turn-Ends and 120 Turn-Continuations (24 pairs × 5 Turn-Ends + 24 pairs × 5 Turns-Continuations). Just as in Experiment 1, AO and VO versions were created from the AV version of each stimulus.

Once the Turn-Ends and Turn-Continuations were identified, the stimuli were edited for use in our gating procedure. We edited each Turn-End at four different gates: 600 ms, 400 ms, and 200 ms[6] before the final gate at the point of turn exchange (T-0) (Fig. 5). The 'point of turn exchange' is defined here as the point right when the second talker (Talker B) starts speaking following the first talker's (Talker A) contribution. The point of turn exchange was redefined for this experiment due to the difference in the task. In Experiment 1, participants were explicitly instructed to respond to the *end* of Talker A's speech. We were interested in the distribution of participants' response times and included part of Talker B's contribution to accommodate for any potential late responses. In this experiment, however, participants were asked to predict whether the talker would finish their turn before Talker B began. Since Talker A still holds the floor until Talker B begins speaking and a response at any time prior to the start of Talker B's speech would indicate a correct predictive response, T-0 was modified. For the Turn-Continuations, sentences were edited 600 ms, 400 ms, and 200 ms before the final gate at T-0. For Turn-Continuations 'T-0' is defined as the start of the next sentence spoken by the same speaker. This resulted in a total of 960 clips (24 pairs × 5 Turn-Ends × 4 gates + 24 pairs × 5 Turn-Continuations × 4 gates).

Experimental equipment

The video and audio presentation software and equipment were identical to Experiment 1. Participants responded using a standard keyboard.

[6] Note that these values are approximations of the actual values (i.e., –600.60 ms, –400.40 ms, –200.20 ms). Videos were edited using number of frames. Gates were created at 6, 12 and 18 frames using a frame rate of 29.97 fps (33.37 ms/frame).

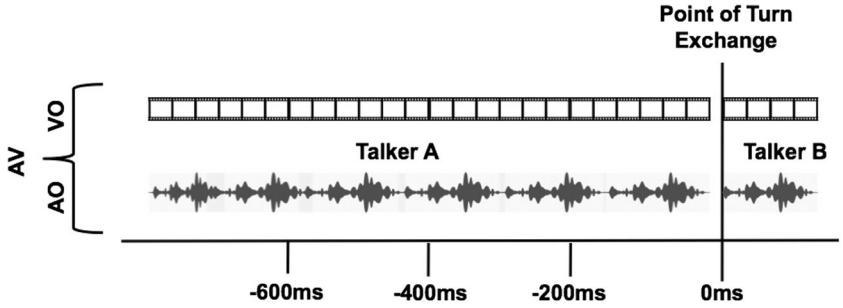

Fig. 5 Schematic of the modified gating paradigm. Participants were presented clips in three modality-specific blocks. Clips within each block were edited such that participants received Turn-Ends and Turn-Continuations at 0 ms, –200 ms, –400 ms and –600 ms prior to the point of turn exchange or the next sentence. It should be noted that participants did not hear Talker B speak in any trials

Participants

Forty-eight undergraduates participated in this experiment (mean age: 23.23 years; 43 females). All participants were native English speakers with no hearing or speech difficulties and normal or corrected-to-normal vision.

Procedure

Stimuli were presented to participants in three modality-specific blocks (AO, AV or VO). Prior to the start of each modality block, participants were given eight practice trials (two Turn-Ends and Turn-Continuations at each gate) with feedback so they could familiarize themselves with the task. Participants were then presented 120 Turn-Ends and 120 Turn-Continuation trials in the three modalities (AO, AV or VO; 80 trials per block). The assignment of clips to each gate and modality was counterbalanced so that all clips occurred in each modality equally often across all participants, and each clip was presented only once to each participant throughout the experiment. That is, participants were presented with 10 Turn-Ends and 10 Turn-Continuations at each gate (T-0, –200 ms, –400 ms and –600 ms) in each modality-specific block. Stimulus presentation was randomized and the order of modality blocks was counterbalanced across subjects. Participants viewed a fixation cross for 500 ms followed by presentation of a stimulus. Participants were provided with the following instructions: 'In this experiment, you will be presented with clips of two people engaged in conversation. After the clip is finished, you will be asked whether you believe the talker who is speaking has finished their turn. You will use the 'Yes' and 'No' buttons to make your response. Sometimes, the clip will cut off in the middle of a sentence. In that case, try to anticipate whether you think the talker will finish their turn once their sentence is complete'.

Results and discussion

Two types of analyses were conducted on our data. In the first preliminary analysis, we analyzed accuracy by employing methods traditionally used in studies involving gating paradigms. We compared the proportions of Turn-End and Turn-Continuation responses to chance performance (50%) to determine the gate at which information in the different conditions was sufficient to recognize a turn exchange. This was analyzed using average participant performance across all items and multiple, Bonferroni-corrected one-sample t-tests. In the main analysis examining turn perception accuracy over the course of turn, we performed a statistical analysis that took both item and subject variance into account, just as in Experiment 1 (Clark, 1973; Baayen, Davidson & Bates, 2007; Brysbaert, 2007). Here, we fitted a model to analyze how participants' turn exchange perception was predicted by the amount of available information (i.e., gate), the modality in which it was presented (i.e., AV, AO and VO) and the turn type (i.e., Turn-End vs. Turn-Continuation).

For our preliminary accuracy analysis, one-sample t-tests comparing the proportion of Turn-End responses to chance showed that in the AV condition, participants were able to predict an upcoming turn exchange (at least) 600 ms before the end of the turn [–600: $t(47) = 2.71$, $P = .009$; $t(47) = –400$ ms: $t(47) = 4.86$, $P < .001$; –200 ms: $t(47) = 8.08$, $P < .001$; T-0: $t(47) = 11.25$, $P < .001$]. However, in the AO and VO conditions, participants only performed better than chance 200 ms before the end [AO: $t(47) = 10.17$, $P < .001$; VO: $t(47) = 4.02$, $P > .001$] and at T-0 [AO: $t(47) = 11.25$, $P < .001$; VO: $t(47) = 9.50$, $P < .001$]. That is, with audiovisual information participants could identify an upcoming turn exchange earlier than with unimodal information. It should be noted, however, that participants could identify Turn-Continuations significantly greater than chance in all modalities at all gates, suggesting that the cues for two turn types may not be processed in the same manner.

For our main analysis that focused on how turn perception varied over time, we examined the influence of Modality and Gate on participants' accuracy at identifying Turn-Ends and Turn-Continuations. This was analyzed using a mixed-effects binary logistic regression. This analysis allows us to identify the effect of predictors (i.e., Gate, Modality and Turn Type on a binary response variable; i.e. correct/Incorrect identification

of stimuli). This analysis was implemented using SPSS's generalized linear mixed models specifying a binomial distribution and a logit link function (GLMM: SPSS 24). We included Turn type (Turn-End vs. Turn-Continuation), Modality (AO vs. AV vs. VO) and Gate (continuous) as fixed factors and used sequential Bonferroni corrections for any follow-up t-tests. Items and Participants were included as random factors. Participants' accuracy for Turn-Ends and Turn-Continuations was coded as a binary outcome variable (Correct or Incorrect).

The results of our mixed-effects analysis revealed significant main effects of Modality [$F(2) = 5.67$, $P = .003$], Gate [$F(1) = 56.15$, $P < .001$] and Turn Type [$F(1) = 74.48$, $P < .001$]. Further, significant Modality × Turn Type [$F(2) = 4.42$, $P = .01$] and Modality × Gate × Turn Type [$F(3) = 46.86$, $P < .001$] interactions were found. Participants were generally more likely to be accurate in the AV condition compared to the AO condition ($b = .15$, $SE = .01$, $t = 3.28$, $P = .002$). Participants were also more likely to be accurate in the AV condition compared to the VO condition ($b = .12$, $t = 3.72$, $SE = .01$ $P = .001$).

Across all modalities, performance in identifying Turn-Ends was better than performance in identifying Turn-Continuations ($b = .59$, $SE = .11$, $t = 5.10$, $P < .001$). Pairwise comparisons following up on the significant Modality × Gate × Turn interaction were conducted. By fixing the continuous Gate predictor at our selected gates (i.e., –600 ms, –400 ms, –200 ms and T-0), we found that participants did not rely on visual information closest to the point of turn exchange when predicting Turn-Ends; participants performed equally well in the AV and AO conditions and better in both conditions compared to the VO condition [Gate at T-0: AV–VO ($t = 3.78$, $SE = .024$; $P < .001$), AO–VO ($t = 3.37$, $SE = .024$, $P = .002$); Gate at –200: AV–VO($t = 4.14$, $SE = .02$, $P < .001$); AO–VO ($t = 2.44$; .02; $P = .002$)]. However, participants benefited from the availability of visual information earlier over the course of a turn and were more accurate in identifying Turn-Ends in the AV condition compared to both AO (Gate at –600: $t = 3.05$, $SE = .03$, $P = .002$) and VO conditions (Gate at –600: $t = 4.54$, $SE = .03$, $P < .001$; Fig. 6a).

Unlike Turn-Ends, in the Turn-Continuation condition participants did not experience a benefit when identifying audiovisual compared to the unimodal stimuli within the earlier gates; that is, auditory information was more informative compared to visual information when identifying that a turn would not end [Gate at –600: AO–VO ($t = 2.30$, $SE = .026$, $P = .04$); AV–VO ($t = 3.13$, $SE = .026$, $P = .005$)]. Participants performed equally well in all three modality conditions when identifying Turn-Continuations closer to the end of the turn (Fig. 6b).

In this experiment, we demonstrated that, in general, availability of auditory information was always more effective in the identification of both Turn-Ends and Turn-Continuations compared to visual information alone. However, the availability of both auditory and visual information (i.e., AV condition) provided a significant benefit early, specifically in the anticipation of an upcoming Turn-End, compared to either modality individually.

A similar pattern of results has been found in previous speech gating studies where an audiovisual benefit was found early in the perception of a word (Jesse & Massaro, 2010). At least in part, this difference may be due to the nature of audiovisual speech stimuli. It has been shown that, as speech unfolds, the availability of visual information precedes that of auditory information (i.e., the onset of the visual mouth movements occurs earlier than the onset of the voice; Chandrasekaran, Trubanova, Stillittano, Caplier & Ghazanfar, 2009). This may result in cross-modal processing where the early visual speech stimulus plays a priming role to enhance auditory speech recognition (Munhall & Tohkura, 1998). Just as specific visual cues may provide earlier indication of the unfolding auditory content of speech, unique visual cues that precede the auditory component of a turn, may mark an approaching turn exchange. For example, several studies have identified speech-preparatory repositioning of the head by the listener (McClave, 2000) and direct gaze behavior from the speaker close to the end of a turn (Cassell et al., 1999). It is possible that such visual cues may provide early enhancement of turn recognition, which then ultimately is decided based on auditory information.

A secondary reason for an early audiovisual advantage may be that the two modalities share information that is both complementary and redundant, as has been suggested in studies of speech perception (Massaro, 1998). As auditory and visual information accumulates over the course of perception, the audiovisual benefit is no longer observable due to the redundancy in the visual and auditory channels (Jesse & Massaro, 2010). A similar explanation might also be applied to the audiovisual advantage observed here; as a turn progresses, certainty in the auditory-only channels increases thus eliminating the audiovisual advantage. Indeed, previous studies have demonstrated that although visual information during turn exchanges is largely redundant with the auditory information, combining both sources of information provides a modest advantage when classifying the end of a turn (Torreira & Valtersson, 2015). Here we demonstrate that the additional benefit of visual information is only apparent early in turn exchange perception.

It is important to note that the early audiovisual advantage only applied to the perception of Turn-Ends. The same audiovisual advantage was not observed for Turn-Continuation perception (i.e., both AO and AV were different from VO) suggesting that auditory information was most important in improving the ability to discern whether a turn would not end. This is, in fact, consistent with our previous findings (Latif et al., 2017), where we used the same stimuli. Here, taking the

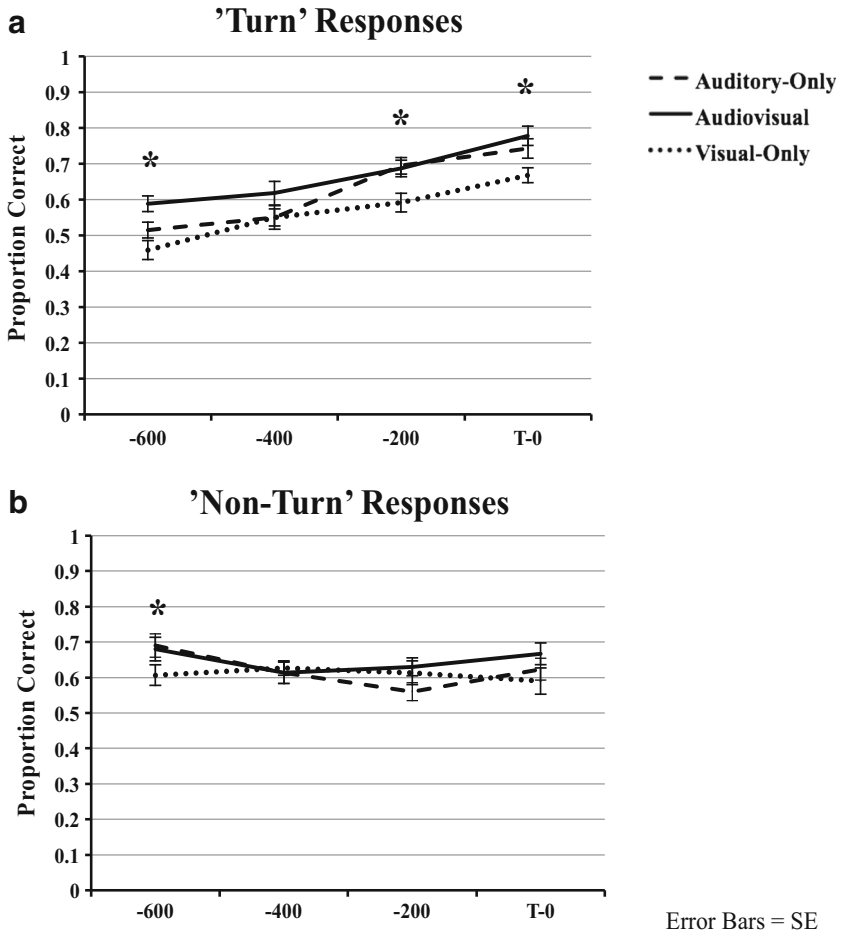

Fig. 6 **a** Proportion of correct responses for the 'Turn-End' trials. Participants showed a significant audiovisual advantage early over the course of turn exchange perception (at –600 ms). As the turn progresses, the audiovisual advantage is not observed, as indicated by the *asterisks*. **b** Proportion correct response for the 'Turn-Continuation' trials

time course of perceptual cue availability into account, we again show that only auditory information is necessary for perceiving a turn-continuation. In fact, at no point along the course of this condition did the visual cues provide information above and beyond the auditory cues. This suggests that although auditory information is necessary for perceiving turn exchange behavior in general, the manner in which cues are used to determine the end of a turn are not the same as those used to detect a Turn-Continuation. Barkhuysen, Krahmer & Swerts (2008) addressed the perception of 'end' versus 'continuation cues by examining responses to complete or partial lists of items relayed by a single talker. They suggested that when perceivers must determine whether an utterance has ended (i.e., whether a spoken list will end), they do so based on local cues that occur near the end of the utterance. In contrast, to determine whether a speaker will continue, perceivers rely on the *absence* of Turn-End cues to then base their decision on global cues from the entire utterance (Barkhuysen, Krahmer & Swerts, 2008). It is possible that in natural, back and forth conversations (compared to spoken lists), Turn-End and Turn-Continuation decisions are more complex and are not made by simply determining the presence or absence of specific 'Turn-End' cues; rather, independent cues may denote the two turn types. Since our ability to make both these decisions is important for successful communication, further work should examine whether the decision that the talker will continue speaking is made based on the absence of Turn-End cues, the presence of specific Turn-Continuation cues, or a combination of the two cue types.

General discussion

In order to achieve the well-coordinated turn exchange behavior observed in conversation, we must use available information to not only decide whether it is appropriate to speak, but to also decide the most appropriate time to do so. In these experiments, we investigated whether auditory and visual information influenced how well participants could time a button-press response to the end of a turn and when, over the course of a turn exchange, participants could reliably use these sources of information to

identify an upcoming Turn-End. In Experiment 1, we showed that although participants were better at discriminating between a Turn-End and a Turn-Continuation in the audiovisual condition compared to the two unimodal conditions, auditory information (i.e., the auditory-only and audiovisual conditions) was needed to synchronize a response with the end of a turn. Similar to Experiment 1, in Experiment 2, we showed that auditory information was generally important for anticipating both an upcoming Turn-End and predicting that a turn would continue. However, the presence of both auditory and visual information (i.e., audiovisual condition) provided a significant advantage above and beyond that of auditory-only and visual-only information early (i.e., in the early gate of the Turn-End trials) in anticipation of an upcoming turn exchange.

Our results showing (1) an early audiovisual advantage but (2) no timing difference between the audiovisual and auditory-only conditions when perceiving Turn-Ends suggest that visual and auditory information may serve complementary roles in turn-taking behavior. Perhaps visual information provides an early, clearer signal that a turn exchange is about to occur, allowing for a relatively more efficient processing of the auditory information closer to the point of turn exchange. It is known that the most cognitively demanding aspects of planning a response occur closer to the point of turn exchange (Sjerps & Meyer, 2015). An early visual signal of an upcoming turn, therefore, may release part of the cognitive demand of the auditory content at this critical point of the interaction.

It is important to note that the benefit provided by the visual information appears to be conditional on the presence of auditory information. Early over the course of a turn, the presence of both auditory and visual information provides a slight advantage. However, using only visual information, participants perform significantly poorly. This suggests that there are specific visual cues that complement the auditory information that do not provide a benefit in the visual-only condition. Perhaps early in the perception of the turn exchange, auditory information is relatively more ambiguous thus allowing for visual information to play a stronger role in order to compensate. Note that such flexible use of visual information has already been shown in speech perception (Jesse & Massaro, 2010). Further, conditional use of specific cues only in the case of ambiguous information parallels findings related to other turn-taking cues. For example, it has been shown that prosodic information is most informative when other lexico-syntactic or visual cues have been compromised (Grosjean & Hirt, 1996; Keitel & Daum, 2015). It should be noted, however, that, in this particular study, we do not explicitly identify whether the source of the visual benefit is linguistic or non-linguistic in nature since we were interested in investigating the full range of audiovisual cues. Further exploration is needed to specify whether the observed visual gain is provided specifically by linguistic visual cues or whether non-linguistic visual cues also play a unique role in enhancing the auditory channel. Further, such work will also have important implications for understanding how sign language users communicate solely within the visual medium yet maintain turn exchange timing similar to audiovisual communicators (de Vos et al., 2015).

Many challenges remain before we understand the visual information processing in natural turn coordination. We have yet to determine which visual cues, in particular, are contributing to our observed early audiovisual advantage. Determining the specific cues may prove difficult considering the flexible role of visual information. It has been shown that visual cues are distributed across the length of a turn and that certain visual cues may indicate unique information depending on *when* it is available (e.g., upper body postural changes early indicate the start of a turn, while whole body postural changes indicate the end of a turn; Cassell et al., 2001). However, these findings have not consistently been demonstrated across studies (see Cook and Lalljee, 1972; Jaffe & Feldstein, 1970) leaving a comprehensive understanding of the distribution of visual information yet to be determined.

Finally, it should be noted that because our tasks involved third-person perception of turn exchanges, it is possible that they do not reflect the same constraints as those that are in place when required to respond in real time while directly engaged in a conversation. For example, it is possible that information gathering may unfold differently when one must directly plan and execute an appropriate response. Further, directly engaging in natural conversations involves additional demands that are supplementary to the content of the conversation itself such as ensuring appropriate gaze and interpreting facial expressions and vocal emotion (Goodwin, 1981; McNeill, 1992; Sacks, 1992). It is noteworthy, however, that previous studies have shown that the perception of turn exchange behavior while engaged in a conversation is, in fact, comparable to third-person perception (Holler & Kendrick, 2015).

Overall, the experiments presented here demonstrate the complementary roles of auditory and visual information when making predictions of turn exchanges during conversation. These findings provide the necessary foundation for furthering our understanding of the yet under-investigated role of visual information during turn-taking behavior. Continued work in this area will help specify the mechanisms involved in maintaining the temporal structure of everyday conversations.

Funding Funding for this research was provided by the National Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institude of Health Research (CIHR), the Ontario Ministry of Colleges, Universities and Training and Queen's University

References

Aschersleben, G & Prinz W. (1995). Synchronizing actions with events: The role of sensory information. *Perception & Psychophysics, 57(3),* 305–317.

Baayen RH, Davidson DJ, & Bates DM. (2007). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59(4),* 390.

Barkhuysen P, Krahmer E & Swerts M. (2008). The interplay between the auditory and visual modality for end-of-utterance detection. *The Journal of the Acoustical Society of America, 123(1)*, 354–365.

Bavelas J, Chovil N, Coates L & Roe L. (1995). Gestures specialized for dialogue. *Personality and Social Psychology Bulletin, 21(4)*, 394–405.

Bögels S & Torreira F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics, 52,* 46–57.

Brysbaert M. (2007). *"The language-as-fixed-effect fallacy": Some simple SPSS solutions to a complex problem (Version 2.0).* Royal Holloway, University of London. Technical Report.

Casillas M, Bobb SB & Clark EV. (2016). Turn taking, timing, and planning in early language acquisition. *Journal of Child Language, 43(6),* 1310–1337.

Cassell J, Bickmore T, Billinghurst M, Campbell L, Chang K, Vilhjálmsson H & Yan H. (1999). Embodiment in conversational interfaces: Rea. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*

Cassell J, Torres O & Prevost S. (1999). Turn taking vs. discourse structure: How best to model multimodal conversation. In Y Wilks (Ed.), *Machine Conversations*, The Hague: Kluwer.

Cassell J, McNeil D & McCullough KE. (1998). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition, 7(1)*, 1–34.

Cassell J, Nakano Y, Bichmore T, Sidner CL & Rich C. (2001). Non-verbal cues for discourse structure. *Proceedings of the Association for Computational Linguistics. ACL 2001*, 106–115.

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A & Ghazanfar A (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology, 5(7)*, e1000436.

Chen Y, Repp B & Patel AD. (2002). Spectral decomposition of variability in synchronization and continuation tapping: Comparisons between auditory and visual pacing and feedback condition. *Human Movement Science, 21(4)*, 515–532.

Clark, HH. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12(4)*, 335–359.

Cook M & Lalljee MG. (1972). Verbal substitutes for visual signals in interaction. *Semitoica, 6(3)*, 212–221.

Corsair. (2016). STRAFE mechanical gaming keyboard - Cherry MX Red. Retrieved from http://www.corsair.com/en-eu/strafe-mechanical-gaming-keyboard-cherry-mx-red.

de Ruiter, JP, Mitterer H, & Enfield, NJ. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language,* 515–535.

de Vos C, Torreira F & Levinson SC. (2015). Turn-timing in signed conversations: coordinating stroke-to-stroke turn boundaries. *Frontiers in Psychology, 6*, 268. https://doi.org/10.3389/fpsyg.2015.00268.

Duncan S (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23(2),* 283–292.

Ford CE & Thompson SA (1996). Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, EA Schegloff & SA Thompson (Eds). *Interaction and Grammar,* Cambridge: Cambridge University Press.

Fox Tree JE (2000). In L Wheeldon (Ed). Coordinating spontaneous talk. In *Aspects of Language Production.* Philadelphia: Psychology.

Goodwin C. (1981). *Conversational organization: Interaction between speakers and hearers,* Cambridge: Academic.

Grant KW, & Seitz PF. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108(3)*, 1197–1208.

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

Grosjean F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics, 28(4).* 267–283.

Grosjean F & Hirt C. (1996). Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subjects. *Language & Cognitive Processes, 11*, 107–134.

Hadar U, Steiner TJ, Grant EC & Rose FC. (1984). The timing of shifts of head postures during conversation. *Human Movement Science, 3(3)*, 237–245.

Ho S, Foulsham T & Kingstone, A. (2015). Speaking and listening with the eyes: Gaze signaling during dyadic interaction. *PLoS One.* https://doi.org/10.1371/journal.pone.0136905

Holler J & Kendrick KH. (2015). Unaddressed participants' gaze in multi-person interaction: Optimizing recipiency. *Frontiers in Psychology, 6*, 98. https://doi.org/10.3389/fpsycg.2015.00098

Jaffe, J & Feldstein, S (1970). *Rhythms of Dialogue.* New York: Academic.

Jesse A & Massaro DW (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception & Psychophysics, 72(1)*, 209–225.

Keitel A & Daum MM. (2015). The use of intonation for turn anticipation in observed conversations without visual signals as source of information. *Frontiers in Psychology, 6,* 108, https://doi.org/10.3389/fpsyg.2015.00108

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

Kendon A. (1972). Some relationships between body motion and speech. In AW Siegman & B Pope (Eds.), *Studies in Dyadic Communication,* New York: Pergamon.

Kendrick KH & Torreira F. (2014). The timing and construction of preference: A quantitative study. *Discourse Processes, 52(4)*, 255–289.

Kennington C, Kousidis S & Schlangen D. (2013). Interpreting situated dialogue utterances: An updatemodel that uses speech, gaze and gesture information. *Proceedings of the 14th Annual Meeting of the Special Interest on Discourse and Dialogue* (pp. 173–182). Metz, France, 22–24 August 2013.

Kraut RE, Fussell SR & Siegel J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human Computer Interaction, 18(1), 13–49.*

Latif N, Alsius A & Munhall KG (2017). Seeing the way: The role of vision in conversation turn exchange perception. *Multisensory Research*, https://doi.org/10.1163/22134808-00002582

Levinson SC. (2016). Turn-taking in human communication—origins and implications for language processing. *Trends in Cognitive Sciences, 20(1),* 6–14.

Magyari L & de Ruiter JP. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology, 3,* 376, https://doi.org/10.3389/fpsyg.2012.00376

Massaro DW. (1998). *Perceiving talking faces: From speech perception to a behavioral principle.* Boston: MIT Press.

McClave EZ. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics, 32*, 855–878.

McLeod RW & Ross HE (1983). Optic flow and cognitive factors in time-to-collision estimates. *Perception, 12(4)*, 417–423.

McNeill D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: The University of Chicago Press.

Miyake Y, Onishi Y & Pöppel E. (2004). Two types of anticipation in synchronization tapping. *Acta Neurobiologiae Experimentalis, 64(3)*, 415–426.

Munhall KG & Tohkura Y (1998). Audiovisual gating and the time course of speech perception. *The Journal of the Acoustical Society of America, 104(1)*, 530–539.

Niemi P & Näätänen R (1981). Foreperiod and simple reaction time. *Psychological Bulletin, 89(1)*, 133–162.

Pashler H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin, 116(2)*, 220–244.

Pfordresher PQ (2006). Coordination of perception and action in music performance. *Advances in Cognitive Psychology, 2(2–3)*, 183–198.

Plant RR, Hammond N & Whitehouse T (2003). How choice of mouse may affect response timing in psychological studies. *Behavior Research Methods, Instruments, & Computers, 35(2)*, 276–284.

Ranganathan R & Carlton LG. (2007). Perception-action coupling and anticipatory performance in baseball batting. *Journal of Motor Behavior, 9(1),* 189–200.

Repp BH & Su YH. (2013). Sensorimotor synchronization: A review of recent research (2006–2012). *Psychonomic Bulletin & Review, 20(3),* 403–452

Rickel J & Johnson WL. (2000). Task-oriented collaboration with embodied agents in virtual worlds. In J. Cassell, J. Sullivan and S. Prevost (eds.), *Embodied Conversational Agents*. Boston: MIT Press.

Riest C, Jorschick AB & de Ruiter JP. (2015). Anticipation in turn-taking: Mechanisms and information sources. *Frontiers in Psychology, 6,* 89, https://doi.org/10.3389/fpsycg.2015.00089

Rosenblum LD, Gordon MS & Wuestefeld AP. (2000). Effects of performance feedback and feedback withdrawal on auditory looming perception. *Ecological Psychology, 12(4)*, 273–291.

Sacks H. (1992). Lectures on conversation (vol. 1). Oxford: Blackwell.

Schubotz RI (2007). Prediction of external events with our motor system: Towards a new framework. *Trends in Cognitive Sciences, 11(5)*, 211–218.

Sinha P, Kjelgaard MM, Gandhi TK, Tsourides K, Cardinaux AL, Pantazis D, Diamond SP and Held, RM (2014). Autism as a disorder of prediction. *Proceedings of the National Academy of Sciences of the United States of America, 111(42)*, 15220–15225.

Sjerps MJ & Meyer, AS. (2015). Variation in dual-task performance reveals late initiation of speech planning in turn-taking. *Cognition, 136,* 304–324.

Stivers T, Enfield NJ, Brown P, Englert C, Hayashi M, Heinemann T, Hoymann G, Rossano F, de Ruiter JP, Yook KE & Levinson SC. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America, 106(26)*, 10587–10592.

Sumby WH & Pollack I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26(2)*, 212–215

Thomas AP & Bull P. (1981). The role of pre-speech posture change in dyadic interaction. *British Journal of Social Psychology, 20(2)*, 105–111.

Thórisson KR. (2002). Natural turn-taking needs no manual: Computational theory and model, from perception to action. *Multimodality in Language and Speech Systems*. Dordrecht: Springer.

Tice M & Henetz T. (2011). Turn-boundary projection: Looking ahead. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 838–843.

Torreira F & Valtersson V. (2015). Phonetic and visual cues to questionhood in French conversation. *Phonetica, 72*, 20–42.

Tresilian JR (1995). Perceptual and cognitive processes in time-to-contact estimation: Analysis of prediction-motion and relative judgment tasks. *Perception & Psychophysics*, *57(2)*, 231–245.

Warren WH Jr. (1990). The perception–action coupling. In *Sensory-motor organizations and development in infancy and early childhood* (pp. 23–37). Dordrecht: Springer.