Research Article

Pupil Responses During Interactive Conversation

Benjamin Masters[1] [ID], Susan Aliakbaryhosseinabadi[2,3], Dorothea Wendt[2,3] [ID]
and Ewen N MacDonald[1] [ID]

Abstract

Pupillometry has been used to assess effort in a variety of listening experiments. However, measuring listening effort during conversational interaction remains difficult as it requires a complex overlap of attention and effort directed to both listening and speech planning. This work introduces a method for measuring how the pupil responds consistently to turn-taking over the course of an entire conversation. Pupillary temporal response functions to the so-called conversational state changes are derived and analyzed for consistent differences that exist across people and acoustic environmental conditions. Additional considerations are made to account for changes in the pupil response that could be attributed to eye-gaze behavior. Our findings, based on data collected from 12 normal-hearing pairs of talkers, reveal that the pupil does respond in a time-synchronous manner to turn-taking. Preliminary interpretation suggests that these variations correspond to our expectations around effort direction in conversation.

Introduction

In recent decades, pupillometry has become a well-known tool for assessing cognitive effort. The so-called task-evoked pupillary response (TEPR) refers to the dilation of the pupil in response to an increase in mental effort and load induced while performing a variety of cognitively demanding tasks, including memory recall, language processing, and quantitative reasoning (Beatty, 1982; Kahneman & Beatty, 1966).

In addition to explaining general cognitive effort, the TEPR also has specific applications in auditory sciences, where pupil response has been interpreted as an indicator of listening effort. For example, pupil dilation has been observed to reflect increased effort based on the signal-to-noise ratio in speech intelligibility tasks (Wendt et al., 2018; Zekveld et al., 2010) and age and hearing status in speech reception threshold tests (Zekveld et al., 2011). The pupil response also indicates increased effort when switching attention between acoustic sources (McCloy et al., 2017) and when dividing attention between multiple streams of speech simultaneously rather than focusing on one (Koelewijn et al., 2014).

For these studies, it has been standard to report summary metrics of pupil response, such as mean and peak dilation, by evaluating changes in pupil size relative to a baseline window immediately before the presentation of a stimulus. Although listening effort can be directly inferred from these metrics in passive listening experiments, interpretation is more difficult in situations where there are cognitive demands unrelated to auditory processing or speech understanding. For example, when a participant response is required (e.g., when asked to press a button) pupil size has been shown to be significantly larger and more sustained than when a response is not required (Privitera et al., 2010). For this reason, pupil dilation during participant response is typically disregarded. It is worth noting, though, that pupil response across different experimental conditions can still be compared if the same participant response is required regardless of condition, but that

[1]Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada
[2]Eriksholm Research Centre, Snekkersten, Denmark
[3]Hearing Systems, Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark

Corresponding author:
Benjamin Masters, Systems Design Engineering, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.
Email: bpmasters@uwaterloo.ca

this may mask changes in pupil size related to the presentation of stimuli. Interpreting pupil response as a measure of effort is further complicated by the fact that a variety of other factors can induce dilations or constrictions of the pupil, such as emotional arousal (Bradley et al., 2008), depth of focus (or accommodation; Kasthurirangan & Glasser, 2005), and light exposure. There can also be measurement artifacts due to distortion of the shape of the pupil during blinking and eye movement that can have confounding effects (Gagl et al., 2011; Yoo et al., 2021).

Evaluating effort in dynamic, interactive environments also presents challenges, for example, sequential stimuli may have overlapping cognitive effects, making it difficult to clearly define these baseline and response windows. One such environment is conversation, within which participants must fluidly switch between listening to and understanding their partner's speech and planning and producing their own speech. Pupil response has recently been evaluated in conversation by considering first-order statistics of pupil size during separate phases of the conversation or in different conversational conditions. Li et al. (2020) found that pupil response between speaking and listening times varied significantly during tasks with a lower communication load but not during tasks with a higher load. Aliakbaryhosseinabadi et al. (2023) observed a larger pupil size when conversing in noise than in quiet. These studies take a conversation-level approach to analyzing differences in effort between speaking and listening time or based on background noise condition.

However, assessing listening versus speaking effort in conversation may not be as simple as analyzing the overall cognitive effort exerted while listening or speaking. Levinson and Torreira (2015) suggested that there must be simultaneous predictive components of both comprehension and production of speech for fluid turn-taking to take place, given that there is considerable latency involved in speech production. Therefore, there is value in determining if we can use pupil response to measure differences in effort at a finer temporal scale, such as around turn-taking in conversation where there is likely to be overlap of listening and speaking effort and a reallocation of attentional resources.

Turn-taking is a coordinated process that requires participants to not only listen to their partner(s) and plan their own speech but also to monitor a variety of other acoustic, behavioral, and contextual cues to interpret when they should take their turn (Brusco et al., 2020; Gravano & Hirschberg, 2011; Hjalmarsson, 2011). Measures of the temporal dynamics of turn-taking, such as interpausal units and floor transfer offsets (FTO), have been shown to significantly vary based on background noise, native versus second language, hearing status, and hearing aid amplification (Petersen et al., 2022; Sørensen et al., 2021, 2024), and it is been suggested that the overall difficulty level of a conversation can be inferred by characterizing the turn-taking dynamics over the course of that conversation. Since the changes in overall conversational difficulty level are reflected in the dynamics of turn-

taking, it seems reasonable to expect that effort may be reflected locally around turn-taking, in the pupil response, as well.

Our current understanding of turn-taking suggests that there is a complex interaction and overlap of both predictive and reactive effort exerted during conversation. However, traditional pupillary analysis methods tend to investigate changes in pupil size in clearly defined windows relative to discrete events, typically on a trial-by-trial basis. Adapting these methods to conversation presents two challenges. The first is that there are no clearly separable trials within conversations around which we can measure the pupil response. To address this, we can define discrete events in conversation around which to measure the pupil response. Given the objective of measuring effort related to turn-taking, we propose defining these events as the starts and ends of conversational turns. The second challenge is that we expect overlap of effort directed to different cognitive processes as turn-taking occurs, which could affect measurements taken with reference to some baseline. For example, if the reference point is defined as when a person starts listening to their partner's speech, the interval immediately prior to this point could still include lingering effects related to that person's speech production, resulting in biased baseline measures. To overcome this challenge, an approach must be taken which can disentangle components of the pupil response, which are related to different cognitive processes and may overlap in time. To enable the disentanglement of pupil responses to different cognitive processes, it is evident that the time course of pupil responses within conversation should be analyzed.

Some studies have assessed the temporal dynamics of pupil responses, for example, by analyzing the response of the pupil, over time, to various stimuli by fitting an Erlang gamma function (Hoeks & Levelt, 1993; McCloy et al., 2016) or fitting polynomials of varying orders using growth curve analysis (e.g., fourth order polynomial models in Wagner et al., 2019). However, these studies operate on an experimental design that is expecting a causal response to some event and also assume a general shape of the pupillary response function. In the context of conversation, we also want to measure predictive components related to turn-taking and expect that assuming a general shape of the response may not be appropriate, as there is not a clear onset of a stimulus to measure the response to. Another approach that has been adopted recently in analyzing the time course of pupil response is the generalized additive mixed model (Van Rij et al., 2019), which does not assume a general shape of the pupil response or causality. Although the application of this modeling approach to analyzing trial-based data is straightforward, it remains unclear how it could be adapted to be analyze interactive conversations, where there are closely spaced events with overlapping effects. One potential option would be to segment the longer recordings into what could effectively be thought of as trials. However, in our case, if a conversation was explicitly segmented, then information about overlapping effects would

be lost. It may be possible to do this type of analysis using interaction effects and carefully designed time-lagged predictors, but it has not yet been done to our knowledge. The response of the pupil to closely spaced sequential stimuli has been successfully analyzed using a deconvolution approach, though. Wierda et al. (2012) used such an approach to identify differences in pupil responses when people are presented with a single visual stimulus versus multiple sequential stimuli, the results of which showed clear peaks in dilation corresponding to the number of stimuli presented. This deconvolution approach appears promising for investigating conversation, where the pupil responses to multiple different sequential turn-related events are measured.

One deconvolution-based method to determine how a response signal is temporally related to a stimulus signal is to estimate a temporal response function (TRF), which is a linear filter (or kernel) that is derived to optimally map from one signal to another (Theunissen et al., 2000). It is often used in neuroscience to model how the brain responds (e.g., via electroencephalography (EEG) or magnetoencephalography) to a continuous stimuli, such as the acoustic envelope of speech (Ding & Simon, 2012). However, the underlying mathematical approach is applicable to any set of time series. This modeling approach accepts input and output signals of arbitrary lengths and derives a filter of a specified size. In the case of a discrete input signal, such as the turn-taking events, the resulting TRFs represent how the response signal varies in a time-synchronous manner around the discrete events (i.e., it derives the optimal impulse response). Additionally, the model can be generalized to include multivariate input signals, and a TRF will simultaneously be derived for each, effectively adding covariates to the model. This multivariate deconvolution is then capable of disentangling components of the response which are attributable in a time-synchronous manner to different stimulus signals (Crosse et al., 2016), defined in this study as the starts and ends of conversational turns. This directly addresses the issue of overlapping temporal effects in the pupil response.

In the present study, we investigate an approach that estimates pupillary TRFs to turn starts and ends during conversation. By doing so, we aim to develop a method with the potential to disentangle and investigate cognitive processing related to speaking and listening using pupil response measurements. We also compare the TRFs derived for conversations taking place in quiet versus noise, as background noise is well known to significantly impact communication difficulty.

Methods and Materials

Participants and Experimental Design

The data analyzed here were previously collected as part of Aliakbaryhosseinabadi et al. (2023). In summary, 12 pairs of older (average age of 63.2 ± 6.4 years) normal-hearing (age-adjusted as per the International Organization for Standardization-7029) Danish talkers were seated face-to-face across a table, approximately 1.5 m apart, and held dyadic conversations in a variety of acoustic conditions, described below. Conversations were elicited using a randomized subset of the DiapixUK spot-the-difference tasks (Baker & Hazan, 2011), modified to include Danish signage. In this collaborative task, each participant received a physical copy of an image, positioned about 40 cm in front of them on a stand, which contained 12 subtle differences from the image their partner received. The participants were instructed to work together to identify the differences between the two pictures through verbal interaction, without seeing each other's images. Participants completed two practice rounds to familiarize with the task before the experiment began. In the context of the experiment, we will use the term "conversation" to refer to the attempt to complete this task for one pair of images. The conversations ended after all the differences were found or 4 min had passed, whichever was first, at which point they were provided a new image and the next conversation began. During the conversations, there was no moderation by the experimenter. The order of experimental conditions was randomized but counterbalanced based on image scene type (street, beach, and farm) and background noise condition (quiet, 60 dBA noise, 70 dBA noise, and simulated conductive hearing loss [SHL]). In the SHL condition, the participants wore earplugs that provided, on average, 25 dB of attenuation. In the noise conditions, a calibrated loudspeaker array played a 10-talker babble noise at the appropriate level. Two replicates of each condition were performed, resulting in eight total conversations for each pair. Speech signals were recorded using headset microphones. Pupil response and eye gaze behavior were recorded simultaneously for both talkers using Tobii Pro 3 eye tracking glasses.

Participants signed an informed consent form that was approved by the Science Ethics Committee for the Capital Region of Denmark (No. H-16036391) and were compensated for their time. Secondary analysis of the data performed at the University of Waterloo was approved by the university's Research Ethics Committee (REB No. 45442).

Speech Preprocessing

Voice activity detection (VAD) was performed on the speech signals using individually defined root mean square thresholds in 5 ms windows with 1 ms of overlap. Windows containing a root mean square power greater than the threshold were classified as containing speech and windows with a power below the threshold as not containing speech. As recommended in Heldner and Edlund (2010), segments containing less than 90 ms of continuous voice activity were relabeled as nonspeech acoustic bursts, and nonspeech segments less than 180 ms were interpreted as silences related to the production of plosives and relabeled as containing speech. VAD signals were then resampled using nearest-neighbor interpolation to 20 Hz.

Pupil Response Preprocessing

The pupil diameter data were extracted from the Tobii recordings and time aligned to the speech signals. Any sample greater than 3 standard deviations from the mean was identified as an artifact due to blinking. A fixed window of 50 ms before and 150 ms after each of these artifacts was removed to mitigate the effects of blinking on pupil response, as suggested in Winn et al. (2018). The eye with the least missing data was selected for further analysis for each person in each conversation. If more than 40% of a participant's pupil diameter samples in any conversation were missing, their pupil response for that conversation was excluded from further analysis. Based on these criteria, 35 pupil response signals were excluded (~18.2%). Thirty-one of the excluded responses belonged to only four participants, who may have been particularly susceptible to eye tracking errors. Fifteen additional responses were excluded due to equipment problems during data collection (~7.8%). The distribution of the responses that were removed was relatively balanced across conditions (quiet: 12, SHL: 17, N60: 11, N70: 10). In the pupil response signals that remained, missing values between valid samples were interpolated using a cubic spline method. Leading and trailing missing data were set to the mean pupil diameter of that participant for that conversation.

A minimalist filtering approach for the pupil data was chosen given the potential for temporal artifacts (De Cheveigné & Nelken, 2019). To mitigate any initial effects from arousal at the task onset, pupil data were first detrended by fitting a fifth order polynomial to each conversation using a robust detrending approach (De Cheveigné & Arzounian, 2018). The pupil responses were then low-pass filtered at 10 Hz using a 50th order Hanning window-based linear phase finite impulse response filter, which served as an antialiasing filter, implemented using a zero-phase approach. The filtered signals were then down sampled to 20 Hz to reduce computational complexity, under the assumption that components of the pupil response related to turn-taking would fall within this 10 Hz bound. Previous observations of the turn-taking rate in Diapix conversations are much slower than 10 Hz, typically falling between 0.4 and 0.6 floor transfers per second (Sørensen et al., 2021), and we do not expect the evoked pupil responses to contain any meaningful content above 10 Hz. The pupil data were then standardized within people, such that the distribution of all pupil diameter measurements across all conditions and replicates for each participant had zero mean and unit variance. This normalization procedure was selected to maintain any changes in the size and variability of the pupil across the different conditions.

Conversational State Changes

Corresponding pairs of VAD signals were input into an automated conversational state labeling algorithm, which identified the start and end times of turns by associating differences in the VAD signals between interlocutors. A conversational turn was defined as when one talker had been speaking for at least 90 ms, and the other talker had simultaneously been silent for at least 180 ms based on the suggestions in (Heldner & Edlund, 2010). For the purposes of this method, state changes are defined as the points in time at which speakers start and stop their turns. Given that this experimental setup is dyadic and different responses are expected to occur whether a person is speaking or listening, state changes are further classified into two categories: belonging to oneself or belonging to one's partner. As seen in Figure 1, this classification scheme yields the following four state change events: self-start, self-stop, partner-start, and partner-stop. The state changes are identified as discrete points in time during each conversation and arranged as a set of impulse trains. Note that it is also possible to have an exchange where one talker begins their turn before their partner has ceded theirs, which corresponds to a negative FTO. In such a case, the beginning and end of the interpausal units were still identified as the start and stop events, even though they appear to occur out of order. An example of this scenario is denoted in Figure 1 at the turn-taking, which contains the first partner-stop and second self-start events.

Gaze Behavior

A measure of gaze distance was used to account for components of the pupil dilation response that could be attributed to the near/far pupil response or differences in luminance between the Diapix image and a person's conversational partner. One of the eye-related time series provided by the Tobii Pro 3 glasses is the gaze points, which are estimated using the intersection of gaze vectors projected from the center of each eye. The depth component of these points provides an indication of the depth of focus of the wearer. In the experimental setup, participants were seated about 1.5 m apart from each other. Each participant had their Diapix images placed directly in front of them, approximately 40 cm away. Thus, the estimated gaze depth at each point in time can be used to infer whether the participant was looking at their picture or their partner.

The gaze depth estimates were preprocessed by first removing outliers. Rather than using statistical methods to do this, it was instead assumed, based on the experimental setup, that any measurement indicating a gaze point greater than 3 m away must be an artifact due to blinking or eye movement. As such, estimates larger than this threshold were removed and replaced by interpolating through the remaining estimates using a cubic spline. The gaze depth estimates were then low pass filtered with a cutoff frequency of 10 Hz. To match the sample rate of the VAD and pupil dilation data, the gaze estimates were down sampled to 20 Hz.

To integrate gaze behavior as a covariate into the model, it must be of the same form as the state change signals (one or multiple impulse trains). For this, a threshold gaze depth of 95 cm, halfway between the expected distances of the image and
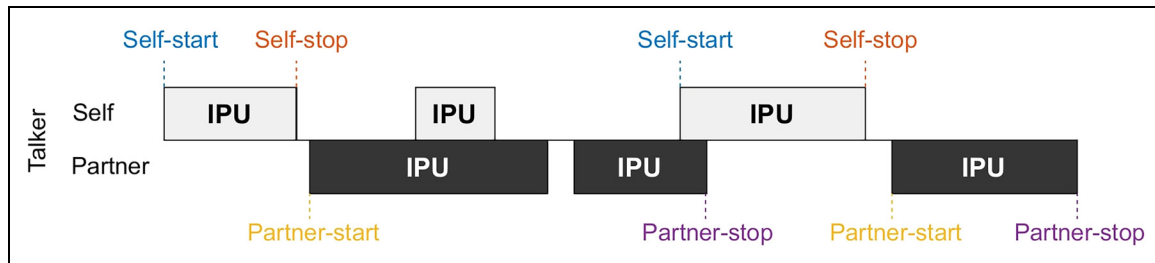
Figure 1. The four conversational state change events denoted in a sample turn-taking exchange between two talkers. IPU = interpausal unit.

partner, was applied to the Tobii gaze depth estimates. If the depth of a person's gaze was below this threshold, their fixation target was assumed to be the picture, whereas if it was above the threshold, it was assumed they were looking at their partner. From this, we computed the amount of time in each conversation spent with a participant looking at their partner as the proportion of total samples with a gaze depth estimate greater than the threshold. We also computed the duration of each glance at the partner, as the duration of continuous segments of the gaze depth signals that were greater than the depth threshold. To extract the differences in pupil response that occur at the change between regions, the points in time when a person's gaze crossed the previously defined threshold were identified. This results in two impulse trains, the first contains the points at which a person looks up at their partner (i.e., transition from near to far) and the second contains the points at which they look down at their image (i.e., transition from far to near).

One key consideration with this approach is that the signals we are using to determine whether a person is looking at the image or their partner are only based on the depth of a participant's gaze and therefore cannot be definitively stated to belong to these regions. Therefore, careful interpretation of these results must be made, especially given that eye gaze behavior has been shown to be a significant factor in regulating turn-taking (Degutyte & Astell, 2021). However, the goal of including gaze behavior as a covariate is to capture the pupil response to distance-related changes in gaze to avoid misclassifying a luminance based or near/far pupil response as a response to a conversational state change. Therefore, the inclusion of this signal should still capture the pupil response appropriately whether a talker is, in fact, looking at their partner as expected or instead averting their gaze to somewhere else in the room.

Estimating Pupil Responses to State Changes

The initial objective of this analysis is to estimate a general pupillary TRF corresponding to each type of conversational state change. To do this, a time-lagged multivariate ridge regression, based on the approach commonly utilized in EEG data analysis (Theunissen et al., 2000), was performed on

each conversation. We define the stimulus signals as the set of impulse trains indicating the locations of state changes in a conversation. For the model to account for changes in gaze target (i.e., from the partner to the image, or vice versa), the stimulus also includes the extracted change of gaze target signals. The response signal is the preprocessed pupil data for the same conversation. Thus, the resulting model computes six pupillary TRFs: one for each of the four state changes, one for when a talker looks from the image to their partner, and one for when a talker looks from their partner to the image.

With this modeling approach, there are four hyperparameters to consider: the minimum and maximum time lags, the time lag step, and the regularization parameter. The minimum and maximum time lags dictate the duration of the estimated TRF, which can be thought of as a window size. In this analysis, the minimum and maximum lags were selected such that the window spanned 2.5 s before and after the state changes. These values were selected based on the expectation that other state changes would likely occur within the window. Figure 2 shows that there is an approximately 50% likelihood of each other state change having occurred by 2.5 s in either direction, enabling the model to disentangle the effects from neighboring state changes. It is important that this window size considers possible overlap of responses from neighboring events, as if the responses to neighboring events overlap but the analysis windows don't, then it is not possible for the TRF approach to disentangle responses from one another. The time lag step was selected as one sample, at the down sampled 20 Hz rate, to maximize temporal resolution. The regularization parameter was optimized via a 10-fold cross-validation performed on each conversation and selected to minimize the average residual sum of squares across all conversations ($\lambda = 5.2269$). This procedure ensures that the same regularization value is used in the derivation of all models, which prevents artifacts that could be introduced due to differing amounts of smoothing between conversations introduced by the regularization.

The TRFs to each of the conversational state changes and gaze target changes were estimated for each talker in each conversation with the MATLAB multivariate temporal response function toolbox using the parameters specified above
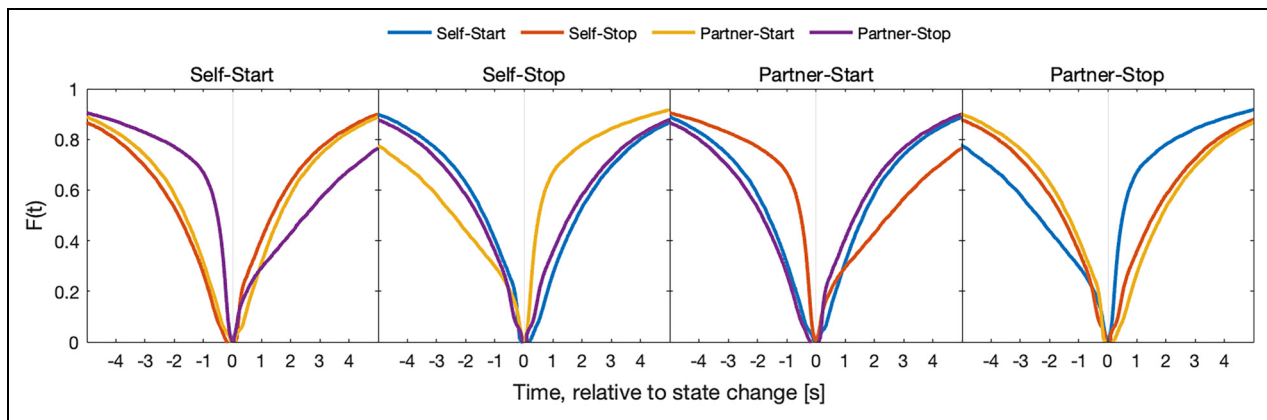
Figure 2. Cumulative distribution functions of the state changes relative to other state changes, over time. Each panel corresponds to a different type of state change. The colored curves indicate the probability that each of the other state changes has occurred, as a function of time, with respect to the state change each panel belongs to. In computing these curves, only positive FTOs are included, such that the probability at time 0 will be 0 for all resulting CDFs.
FTO = floor transfer offset; CDF = cumulative distribution function.

(Crosse et al., 2016). To perform a group-level analysis of the results, combinations of these TRFs can be averaged together or statistically analyzed as a sample. The set of TRFs to be averaged is determined based on the desired analysis. To compare between conditions, for example, between quiet and 70 dBA noise, one would average all TRFs within those conditions, resulting in a set of pupillary TRFs for each condition, with each TRF corresponding to one of the stimulus signals.

Given the objectives of this work are to analyze the pupillary response to turn-taking in conversation, the results presented will generally only include the response curves corresponding to the so-called conversational state changes. However, in all results presented, the gaze target changes were included as covariates in the model; thus, the response curves are derived while accounting for gaze behavior.

Statistical Methods

To assess when the state change responses were significantly different from 0, a temporal cluster-based permutation testing approach was employed (Maris & Oostenveld, 2007). This approach simultaneously accounts for autocorrelation in the pupil dilation data and the multiple comparisons problem because it compares the collection of pointwise statistics from TRFs that are derived using the actual, unmanipulated, pupil response signals but with randomly permuted stimulus signals or condition labels. The routine was implemented as follows. The arrays containing the points of conversational state change and gaze change were randomly permuted across time, and a set of null pupillary TRFs was calculated for all conversations. Using a P value of .05, pointwise one-sample t tests were performed on the set of null TRFs. The t statistics were summed within continuously significant segments and the largest sum of t statistics was stored. This approach was chosen over simply using the largest continuous length of

significant regions as it also considers the degree of the significance within each segment. This process was then repeated over 1000 iterations to form a null distribution of the largest summed significant t statistics. From this null distribution, the 95th percentile (chosen based on a P value of .05) was found and set as the threshold defining the minimum significant sum of t statistics. Therefore, when assessing the significance of the actual TRFs, a pointwise t test was performed and the t statistics in continuously significant segments were summed. If the sum exceeded the threshold defined from the 95th percentile of the null distribution, then the region was assumed to be significant.

A similar process was employed to compare TRFs across experimental conditions. Instead of permuting the arrays of state changes, the condition labels are randomly shuffled for each participant, and all possible pairs of conditions are compared using pointwise two-sample t tests. The largest sum of continuously significant t statistics was once again used to form a null distribution from which a threshold was defined based on a P value of .05.

Results

Pupil Response Across All Conversations

Figure 3 shows the state change responses found by averaging across all conversations, also denoted are the segments within which the curves have a value significantly different from zero. Significant responses were found to all state changes except for self-start. As can be observed, there is an elevated pupil response around both self-stop and partner-stop, and there is a sharper pupil response that peaks ~1 s after partner-start. The segments of significance in the results reveal that a consistent time-synchronous change in pupil
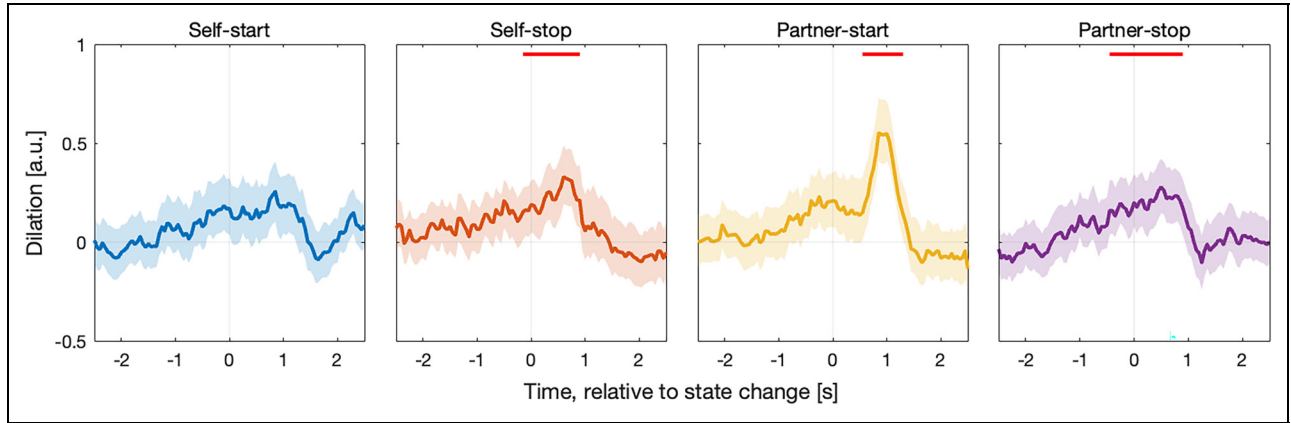
Figure 3. State change response curves obtained by averaging the results for all participants in all conversations. The highlighted regions indicate the pointwise confidence intervals. The red bars indicate consecutive pointwise significant intervals over which the summed One-sample *t* statistic exceeds a significance threshold set based on a *P* value of .05, denoting that the curves are significantly different from 0.

response is observed for three of the four conversational state changes.

By Condition

To assess how the pupil responses may vary based on expected difficulty of the conversation, the state change responses can be found by pooling TRFs only across conversations that took place in the same conditions. For this analysis, we compared the results between the quiet and 70 dBA noise condition, with the expectation that this combination will have the highest disparity in perceived difficulty and therefore emphasize any processing differences that exist.

Figure 4 shows the conditional state change responses along with the segments within which the two curves are significantly different from each other. We see that the TRFs corresponding to each condition have intervals, which are significantly different around three of the four state changes. It is observed that the pupil is significantly larger after starting one's turn in quiet than in noise. Conversely, the dilations of the pupil that occur in noise after ending one's turn and after the start of a partner's turn are larger in noise than in quiet.

Comparison to Grand Averaging

As mentioned previously, a common analysis method for TEPR is to average the pupil's dilation trace over many individual trials. This approach can be applied to our case by averaging across fixed windows around every state change, which we will call a grand average, with cases where the window would extend beyond the bounds of conversation excluded. This is distinct from the analysis method used to generate the results previously as no deconvolution is performed rather the conversations are directly segmented into windows around state changes and the responses in these windows

are averaged. This analysis is performed to assess the potential benefits of the deconvolution-based TRF method.

Figure 5 plots the previously derived response curves, from Figure 3, against the curves obtained from a grand average of pupil responses around every state change, across all participants and conversations. Due to the differences in scaling that happens because of regularization, both sets of curves are normalized such that the set of all TRFs has zero mean and unit variance, and the set of all grand average curves has zero mean and unit variance. In general, the two sets of curves agree and resemble one another, although there are some differences of note that could be related to smearing. For example, the amplitude of the dilation after partner-start is greater in the TRF derived curve than in the grand averaging curve, suggesting that the TRF approach attributes this effect more strongly to the partner-start event than the averaging approach does.

Analysis of Gaze Correction

In the following analysis of gaze correction, we assume that a gaze depth below the previously defined threshold indicates a talker looking at their image, and a depth beyond the threshold indicates a talker looking at their partner. Given that the Diapix task requires participants to predominantly look at the image to complete the task, we think this is a reasonable assumption.

The estimates of the distributions of gaze depth and duration of fixations to the partner for each condition (i.e., pooled across replicates and participants) are plotted in Figure 6, revealing that most of the time participants looked at the image and that glances at their partner are typically short (the mode of the distribution was 70 ms). Also displayed are the total percentages of each conversation that were spent looking at the partner. It is observed that in all four conditions, the
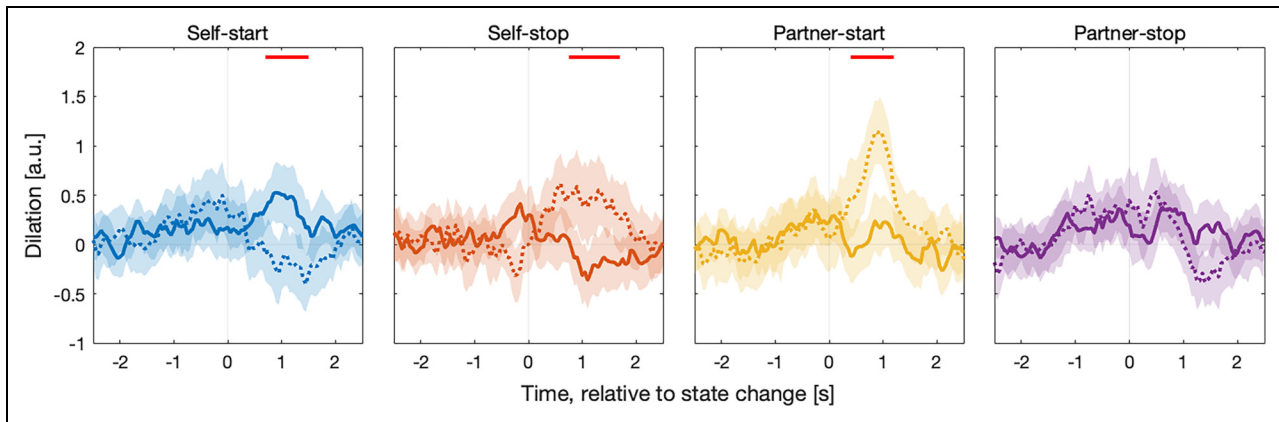
Figure 4. State change responses obtained only by averaging conversations that took place in quiet (solid line) and 70 dBA noise (dotted line). The shaded regions indicate the pointwise one-sample confidence intervals of each curve. The red bars indicate intervals over which the summed two-sample t statistic exceeds the significance threshold set based on a P value of .05, denoting the two curves are significantly different from each other.
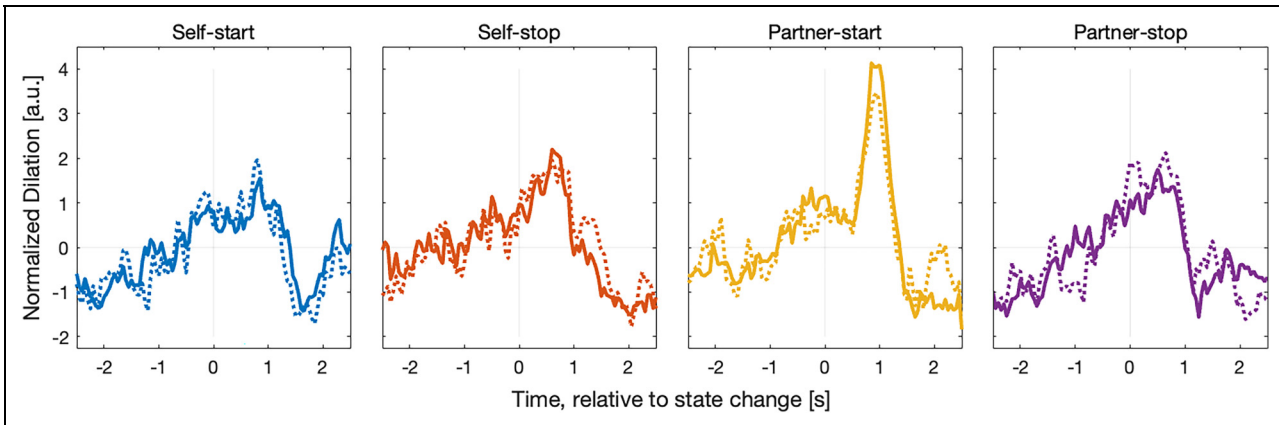

Figure 5. Derived state change responses (solid lines) plotted against the grand averages of fixed windows (dotted lines). The solid lines corresponding to the derived pupillary state change responses are identical to those presented in Figure 3.
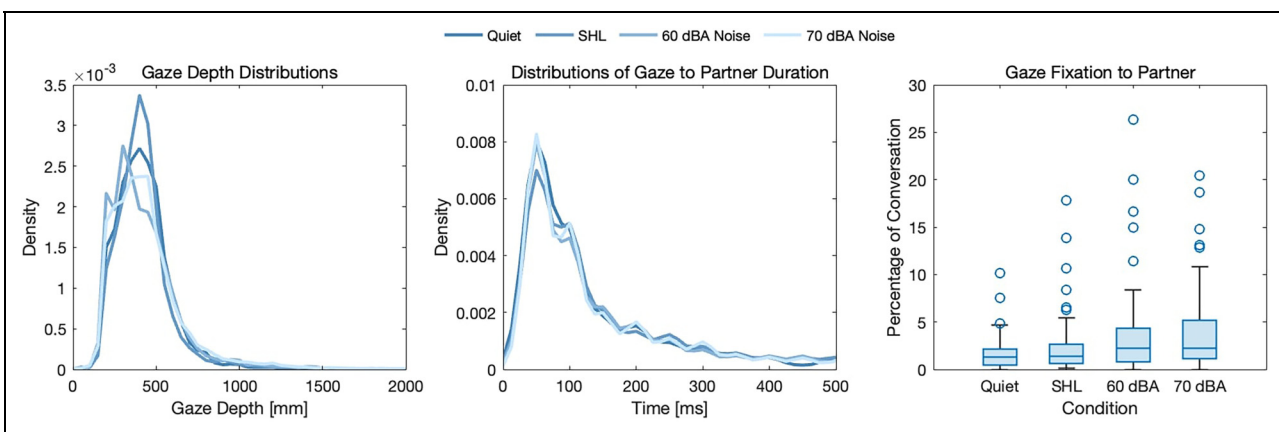

Figure 6. Distributions of gaze depth and the duration of fixations to the partner across all conversations, by condition. Color indicates the condition the conversation took place in. Also included are the overall percentages of each conversation that were spent looking at the partner, by condition, determined as the percentage of time points where the gaze depth was greater than the 95 cm threshold defined previously.
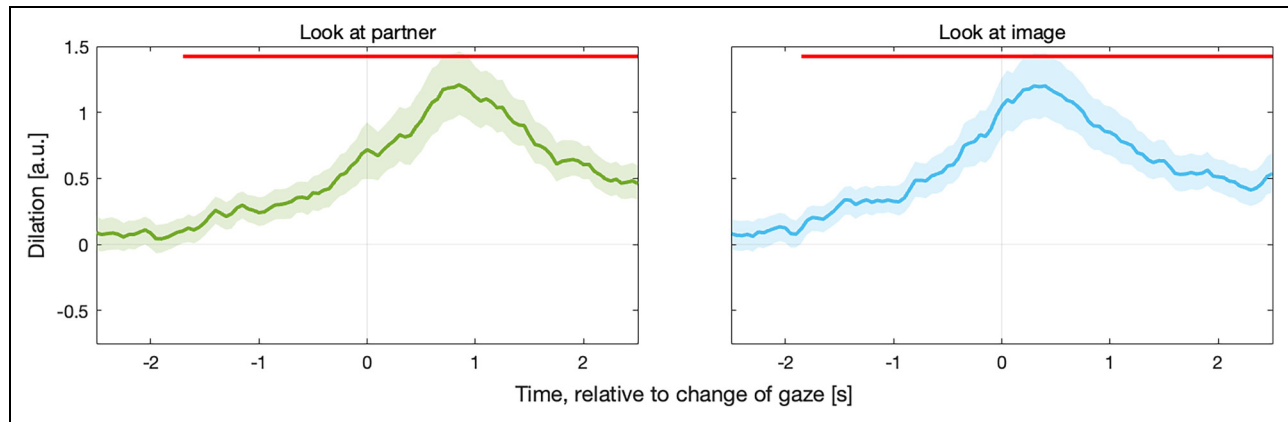
Figure 7. Pupil responses to change of gaze target. The shaded regions indicate the pointwise confidence intervals. The red bars indicate intervals over which the set of TRFs averaged to produce these curves has a mean significantly different from 0. TRF = temporal response function.

median percentage is less than 5%, confirming that, on average, people spent most of each conversation looking at the image.

The analysis of the pupil responses to a change of gaze target reveals significant effects on pupil dilation, as seen in Figure 7, but no significant difference in the gaze-related response by condition, as seen in Figure 8. It can be noted that the responses obtained for looking at the partner and looking at the image are similar, but that the response corresponding to looking at the image is left-shifted, relative to the curve corresponding to looking at their partner.

Discussion

Interpretation of State Change Response Curves

In this work, we present a method for estimating how pupil response varies around turn-taking in conversation. This method not only estimates the pupil response to a specific conversational state change (e.g., a partner starting their turn) but also simultaneously considers the other conversational state changes and gaze behavior while doing so. Importantly, we found significant pupil responses to turn-taking in conversation that also varied based on background noise condition.

The intervals which are significantly different from zero in the response curves found by averaging TRFs across all conversations seem to correspond with our expectations around effort in conversation. Most notably, there is a large pupil dilation that peaks ∼1 s after a partner begins their turn, which may be indicative of directing effort towards listening. It is possible that this could also be indicative of arousal as a reaction to hearing a new auditory stimulus. If this is the case, then there should be no difference between conditions with different background noise levels. However, when deriving the state changes only for conversations in alike conditions, this

dilation was found to be larger in noise than in quiet, suggesting that it is not an effect of arousal but rather of task demands via increased noise level, which suggests increased effort investment.

Another observation from the curves derived across all conversations is the significant dilation after a person stops talking. This effect is also likely related to the reallocation of effort toward listening. Intuitively, this effect should be attributed to both the end of a talker's turn and the start of their partner's turn, as it is unlikely that devoting effort toward listening is purely reactionary to a partner beginning to speak. Instead, it is expected that talkers would begin reallocating effort toward listening as they finish their speech production and planning. However, the interval to the start of the subsequent partner's turn (the FTO) can exhibit large variability, especially in more difficult communication environments (Sørensen et al., 2021, 2024). Therefore, the timing of the preparatory effects may be more highly correlated with the end of the preceding turn rather than the start of the following turn. A similar dilation is observed shortly after the end of a partner's turn, which could be related to speech planning. Previous studies of neural correlates of turn-taking have shown results that suggest speech planning starts well before a conversational partner's turn ends and begins as soon as possible based on the content of the partner's speech (Bögels, 2020; Bögels et al., 2018). This observation could support the hypothesis that an increase in effort directed toward speech planning is temporally correlated with the end of a partner's turn. Given that preparatory effort has been shown for speaking, it seems reasonable to suggest that preparatory effort could also be directed toward listening.

Further differences were found between the response curves derived based on the condition of the conversation. The pupil generally seems to constrict as one starts speaking in noise but dilates as one starts speaking in quiet. We suggest that this effect is related to an increase in effort required to
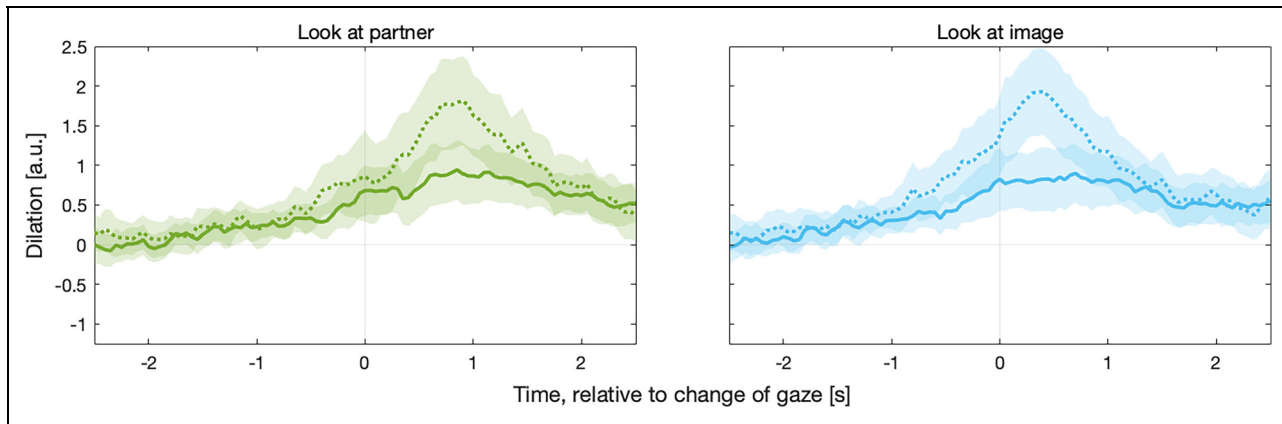
Figure 8. Pupil responses to change of gaze by condition, where the solid line corresponds to the quiet condition and the dotted line corresponds to the 70 dBA background noise condition. The shaded regions indicate the pointwise confidence intervals of each curve. No significantly different intervals were identified between conditions for these curves.

listen and comprehend in noise resulting in a relative decrease in pupil size when going from listening to speaking. This suggestion is supported by the findings from Li et al. (2020), which showed that pupil size while listening during conversations with a higher communication load was larger than in conversations with a lower communication load. Although communication load in this context was not increased via background noise, both have the effect of introducing difficulty into the conversation. To further support this suggestion, a post hoc statistical analysis was performed on the participants' pupil diameters during different phases of the conversation. The mean pupil diameter during speaking and listening time was computed for each talker in each conversation. A linear mixed effects model of the form Mean Diameter $\sim$ Action $\times$ Condition $+ (1 \mid$ Talker$) + (1 \mid$ Replicate$)$ was fit to the mean diameter measurements from the quiet and 70 dBA noise conditions, where "Action" was a categorical variable that specified whether the mean diameter value corresponded to speaking or listening time, "Condition" represented the level of background noise, "Talker" was a random effect corresponding to individual participants, and "Replicate" was a random effect that denoted the repetition of a given condition. An analysis of variance revealed a significant positive effect of the interaction, implying that pupil size was significantly larger while listening in noise, $F(1, 144) = 6.82$, $P < .01$, but no other significant effects. These results support the previous suggestion that the changes in pupil size observed in the TRFs are related to differences in effort between speaking and listening in the more difficult condition. A similar but opposite effect is observed when one stops talking. The pupil dilates in noise while preparing/starting to listen and does not in quiet, likely for the same reasons. Further support of this claim can be found by assessing the levels of speech in quiet and noise. In the quiet condition, the average speech level was 59.5 dBA in a room with an ambient A-weighted equivalent continuous sound level of less than 35 dBA, whereas in the 70 dBA noise condition, the signal-to-noise ratio was found to be only 1.2 dB. Given this relatively low SNR in noise, it is expected that listening would be considerably more effortful than in the quiet condition, a notion which is supported by previous research on speech understanding in noise (Ohlenforst et al., 2018).

Advantages of the Proposed Method

One of the proposed advantages to using the regression-based method over simply averaging a window around every turn-taking event was the ability to separate (i.e., demix) the effects of each state change on the pupil response. The results in Figure 5 support this argument. For example, the increased amplitude of the partner-start peak in the TRF relative to the grand averaging curve suggests the TRF approach finds this feature more strongly time synchronous, and therefore attributable, to the partner-start event. However, the two sets of curves are similar. One explanation for the degree of similarity is that, since there is a large quantity of turns ($n = 5038$ in the conditions analyzed), the grand-averaging approach is smoothing out most components of the pupil responses which are not highly time synchronous to the state changes. However, if there were fewer turns or less variability in the interval lengths between neighboring state changes, then time-synchronous averaging is less capable of isolating the responses from different events (e.g., portions of the response related to self-stop may appear as an artifact in the grand-averaged result for partner-start). The participant pool in this study consisted of older talkers. Previous research, which has analyzed differences in conversational dynamics between young normal-hearing talkers and older hearing-impaired talkers, has found that the older hearing-impaired talkers exhibit longer turns and FTOs and more variable FTOs when compared to their younger normal-hearing counterparts (Sørensen et al., 2024). In addition to conversations between

young normal-hearing interlocutors, there are other possible scenarios where the demixing capabilities of the TRF approach may have an important role in analysis, such as examining effort in multitalker settings (e.g., target vs. masked speech), where the pupil response to the target and masking signals could be assessed separately by deriving TRFs for each.

Another proposed advantage of the TRF method is the ability to account for covariates in the model, which helps prevent misattribution of pupil response to, for example, looking at your partner to one of the state changes. This is a valid concern, as gaze plays a significant role in turn-taking behavior (Degutyte & Astell, 2021). Our findings revealed significant responses of the pupil to a change of gaze target. The response curves corresponding to the change of region of interest indicate a dilation after a talker looks toward their partner, aligning with the expected effect of the near/far pupil response (Kasthurirangan & Glasser, 2005). However, the response to looking back at the image is less intuitive, as one may expect to observe a clear constriction of the pupil, which is not the case. It can be noted that the response curve corresponding to looking at the image resembles a left-shifted version of the response to looking at the partner. Therefore, if considering the response latency of the pupil, one can think of the response curve representing the pupil size around a change of gaze to the image as being a return to the pregaze change pupil size. Given the short durations of the glances at the partner and the low variability thereof, as demonstrated in Figure 6, it is likely that these events may not be distinct enough from each other, or frequent enough, for the responses to be completely pulled apart. Comparing the median number of glances at the partner by talker in each conversation, which is 17, to the median number of turns by each talker in each conversation, which is 51, supports this as a possibility. These metrics show that turn-taking is occurring three times as often as changes of gaze target while completing this task. In general, these results suggest that the model is accounting for differences in gaze behavior when deriving the effects of turn-taking. Although there is no significant difference between the gaze-change response curves by condition they do appear to be, qualitatively, of a different shape. One possible explanation for this difference is the purpose of glances at the partner. In noise, the change of gaze target may be effort related, whereas in quiet they may more often be used as a social cue. If people are looking at their partner in noise when they are experiencing difficulty in the conversation, it would make sense that the dilation would be stronger, as it would be temporally correlated with both a distance-based pupil accommodation and an increase in effort. Whereas in quiet, there may be no effort related effect that covaries with the near/far pupil response. This idea can be partially supported by previous studies of gaze behavior in conversation, which have shown that talkers tend to spend more time looking at their conversational partner's mouth when conversing in noise than in quiet (Hadley et al.,

2019). This implies that when conversational difficulty increases, talkers are relying upon the long-studied benefits of visual information for increased speech intelligibility (Erber, 1975; Sumby & Pollack, 1954). However, the previously mentioned role of eye gaze in regulating turn-taking behavior (Degutyte & Astell, 2021) would suggest that talkers also direct their gaze toward their partner in quiet, where it is less likely that the talkers would be doing so for the benefits of visual information on listening.

Although there are interesting results and potential interpretations related to change of gaze target, most of the time people looked at the image rather than their partner in all conversations. Across all conditions and conversations, the median percentage of time each participant spent looking at their partner was less than 5%. This is likely an effect of the task used to elicit the conversations, as the Diapix task necessitates near constant reference of the image to be able to determine if differences exist. For this reason, it is likely that accounting for the gaze behavior in this experiment had little effect, which was verified by comparing to response curves derived without the gaze points included. However, in more naturalistic communication settings, such as a free conversation, the role of gaze, and the correction thereof, would likely be more important given the previously mentioned role of eye-gaze as a turn-taking cue.

Conclusions

This study presented a method for deriving pupillary responses in interactive conversations based on turn-taking. This approach revealed consistent pupil responses to turn-taking in conversation, which can potentially be used to infer how cognitive effort varies during the transition from listening to speaking in conversation and how these effects change based on background noise condition. In addition to being applicable to interactive conversation, advantages of this method over other trial-based approaches include the ability to account for external covariates in the model (such as gaze) rather than having to process pupil data to account for these effects, and the capability of demixing overlapping responses from neighboring events. Further work is needed to understand how the interpretation of the results here can provide insight into changes in processing demands during conversation. A potential follow-up would be to have a listener–observer follow along, while talkers participate in a subset of the Diapix tasks and measure their pupil response. This would separate out the cognitive effects from speaking and those associated with task demands and listening. The results from such a study would provide further insight into how attention is divided and effort is reallocated during conversation.

ORCID iDs

Benjamin Masters   https://orcid.org/0009-0006-3479-5384
Dorothea Wendt   https://orcid.org/0000-0002-3340-1180
Ewen N MacDonald   https://orcid.org/0000-0002-6973-2926

Ethical Considerations

This experiment was approved by the Science Ethics Committee for the Capital Region of Denmark (H-16036391). Secondary analysis of the data performed at the University of Waterloo was approved by the university's Research Ethics Committee (REB 45442).

Consent to Participate

Participants signed an informed consent form which was approved by the Science Ethics Committee for the Capital Region of Denmark (H-16036391).

Funding

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability

The data that support the findings of this study are openly available in Borealis (Masters et al., 2025).

References

Aliakbaryhosseinabadi, S., Keidser, G., May, T., Dau, T., Wendt, D., & Rotger-Griful, S. (2023). The effects of noise and simulated conductive hearing loss on physiological response measures during interactive conversations. *Journal of Speech, Language, and Hearing Research*, *66*(10), 4009–4024. https://doi.org/10.1044/2023_JSLHR-23-00063

Baker, R., & Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, *43*(3), 761–770. https://doi.org/10.3758/s13428-011-0075-y

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Bögels, S. (2020). Neural correlates of turn-taking in the wild: Response planning starts early in free interviews. *Cognition*, *203*, 104347. https://doi.org/10.1016/j.cognition.2020.104347

Bögels, S., Casillas, M., & Levinson, S. C. (2018). Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question. *Neuropsychologia*, *109*, 295–310. https://doi.org/10.1016/j.neuropsychologia.2017.12.028

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Brusco, P., Vidal, J., Beňuš, Š, & Gravano, A. (2020). A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool. *Speech Communication*, *125*, 24–40. https://doi.org/10.1016/j.specom.2020.09.004

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604. https://doi.org/10.3389/fnhum.2016.00604

De Cheveigné, A., & Arzounian, D. (2018). Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data. *NeuroImage*, *172*, 903–912. https://doi.org/10.1016/j.neuroimage.2018.01.035

De Cheveigné, A., & Nelken, I. (2019). Filters: When, why, and how (not) to use them. *Neuron*, *102*(2), 280–293. https://doi.org/10.1016/j.neuron.2019.02.039

Degutyte, Z., & Astell, A. (2021). The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. *Frontiers in Psychology*, *12*, 616471. https://doi.org/10.3389/fpsyg.2021.616471

Ding, N., & Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*(1), 78–89. https://doi.org/10.1152/jn.00297.2011

Erber, N. P. (1975). Auditory-Visual perception of speech. *Journal of Speech and Hearing Disorders*, *40*(4), 481–492. https://doi.org/10.1044/jshd.4004.481

Gagl, B., Hawelka, S., & Hutzler, F. (2011). Systematic influence of gaze position on pupil size measurement: Analysis and correction. *Behavior Research Methods*, *43*(4), 1171–1181. https://doi.org/10.3758/s13428-011-0109-5

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, *25*(3), 601–634. https://doi.org/10.1016/j.csl.2010.10.003

Hadley, L. V., Brimijoin, W. O., & Whitmer, W. M. (2019). Speech, movement, and gaze behaviours during dyadic conversation in noise. *Scientific Reports*, *9*(1), 10451. https://doi.org/10.1038/s41598-019-46416-0

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568. https://doi.org/10.1016/j.wocn.2010.08.002

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, *53*(1), 23–35. https://doi.org/10.1016/j.specom.2010.08.003

Hoeks, B., & Levelt, W. J. M. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers*, *25*(1), 16–26. https://doi.org/10.3758/BF03204445

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kasthurirangan, S., & Glasser, A. (2005). Characteristics of pupil responses during far-to-near and near-to-far accommodation. *Ophthalmic and Physiological Optics*, *25*(4), 328–339. https://doi.org/10.1111/j.1475-1313.2005.00293.x

Koelewijn, T., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2014). The pupil response is sensitive to divided attention during speech processing. *Hearing Research*, *312*, 114–120. https://doi.org/10.1016/j.heares.2014.03.010

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, 731. https://doi.org/10.3389/fpsyg.2015.00731

Li, H., Epps, J., & Chen, S. (2020). Think before you speak: An investigation of eye activity patterns during conversations using eyewear. *International Journal of Human-Computer Studies*, 143, 102468. https://doi.org/10.1016/j.ijhcs.2020.102468

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Masters, B., Aliakbaryhosseinabadi, S., Wendt, D., & MacDonald, E. N. (2025). Replication data for pupil responses during interactive conversation. *Borealis*. https://doi.org/10.5683/SP3/LAYQ3J

McCloy, D. R., Larson, E. D., Lau, B., & Lee, A. K. C. (2016). Temporal alignment of pupillary response with stimulus events via deconvolution. *The Journal of the Acoustical Society of America*, 139(3), EL57–EL62. https://doi.org/10.1121/1.4943787

McCloy, D. R., Lau, B. K., Larson, E., Pratt, K. A. I., & Lee, A. K. C. (2017). Pupillometry shows the effort of auditory attention switching. *The Journal of the Acoustical Society of America*, 141(4), 2440–2451. https://doi.org/10.1121/1.4979340

Ohlenforst, B., Wendt, D., Kramer, S. E., Naylor, G., Zekveld, A. A., & Lunner, T. (2018). Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. *Hearing Research*, 365, 90–99. https://doi.org/10.1016/j.heares.2018.05.003

Petersen, E. B., MacDonald, E. N., & Josefine Munch Sørensen, A. (2022). The effects of hearing-aid amplification and noise on conversational dynamics between normal-hearing and hearing-impaired talkers. *Trends in Hearing*, 26, 23312165221103340. https://doi.org/10.1177/23312165221103340

Privitera, C. M., Renninger, L. W., Carney, T., Klein, S., & Aguilar, M. (2010). Pupil dilation during visual target detection. *Journal of Vision*, 10(10), 3. https://doi.org/10.1167/10.10.3

Sørensen, A. J. M., Fereczkowski, M., & MacDonald, E. N. (2021). Effects of noise and second language on conversational dynamics in task dialogue. *Trends in Hearing*, 25, 23312165211024482. https://doi.org/10.1177/23312165211024482

Sørensen, A. J. M., Lunner, T., & MacDonald, E. N. (2024). Conversational dynamics in task dialogue between interlocutors with and without hearing impairment. *Trends in Hearing*, 28, 23312165241296073. https://doi.org/10.1177/23312165241296073

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2), 212–215. https://doi.org/10.1121/1.1907309

Theunissen, F. E., Sen, K., & Doupe, A. J. (2000). Spectral-Temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *The Journal of Neuroscience*, 20(6), 2315–2331. https://doi.org/10.1523/JNEUROSCI.20-06-02315.2000

Van Rij, J., Hendriks, P., Van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trends in Hearing*, 23, 233121651983248. https://doi.org/10.1177/2331216519832483

Wagner, A. E., Nagels, L., Toffanin, P., Opie, J. M., & Başkent, D. (2019). Individual variations in effort: Assessing pupillometry for the hearing impaired. *Trends in Hearing*, 23, 233121651984559. https://doi.org/10.1177/2331216519845596

Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., & Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hearing Research*, 369, 67–78. https://doi.org/10.1016/j.heares.2018.05.006

Wierda, S. M., Van Rijn, H., Taatgen, N. A., & Martens, S. (2012). Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 109(22), 8456–8460. https://doi.org/10.1073/pnas.1201858109

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An Introduction for those who want to get started. *Trends in Hearing*, 22, 233121651880086. https://doi.org/10.1177/2331216518800869

Yoo, K., Ahn, J., & Lee, S.-H. (2021). The confounding effects of eye blinking on pupillometry, and their remedy. *PLOS ONE*, 16(12), e0261463. https://doi.org/10.1371/journal.pone.0261463

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear & Hearing*, 31(4), 480–490. https://doi.org/10.1097/AUD.0b013e3181d4f251

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear & Hearing*, 32(4), 498–510. https://doi.org/10.1097/AUD.0b013e31820512bb