Augmenting Turn-taking Prediction with Wearable Eye Activity During Conversation

Hang Li[1,2], *Siyuan Chen*[1], *Julien Epps*[1,2]

[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2]Affiliation Data61, CSIRO, Australia
hang.li@unsw.edu.au, siyuan.chen@unsw.edu.au, j.epps@unsw.edu.au

Abstract

In a variety of conversation contexts, accurately predicting the time point at which a conversational participant is about to speak can help improve computer-mediated human-human communications. Although it is not difficult for a human to perceive turn-taking intent in conversations, it has been a challenging task for computers to date. In this study, we employed eye activity acquired from low-cost wearable hardware during natural conversation and studied how pupil diameter, blink and gaze direction could assist speech in voice activity and turn-taking prediction. Experiments on a new 2-hour corpus of natural conversational speech between six pairs of speakers wearing near-field eye video glasses revealed that the F1 score for predicting the voicing activity up to 1s ahead of the current instant can be above 80%, for speech and non-speech detection with fused eye and speech features. Further, extracting features synchronously from both interlocutors provides a relative reduction in error rate of 8.5% compared with a system based on just a single speaker. The performance of four turn-taking states based on the predicted voice activity also achieved F1 scores significantly higher than chance level. These findings suggest that wearable eye activity can play a role in future speech communication systems.

Index Terms: voice activity prediction, turn-taking prediction, multi-modal, conversation, eye activity

1. Introduction

Conversation plays an important role in our daily life, where our feelings and ideas are expressed and exchanged, as well as other important information. Developments in human-like robotics and computer-aided mediations in human-human conversation (for example during COVID-19) create an imperative demand on the understanding of conversations automatically. This results in an increasing number of studies [1]–[6] focusing on building a human-like dialogue system that responds and engages correctly. To smooth conversation engagement, one of the aims is to predict conversation turn-taking automatically [1], [7], [8], i.e. identify the time point where people wish to speak or to end speech ahead of time.

Recently, many dialogue systems (e.g. [9], [10]) have provided rigid turn-taking estimation by detecting the end of the utterance, which relies on a duration threshold of the non-speech part to predict the next turn. Apart from speech, different modalities, e.g. behavioral signals [11], [12], physiological signals [3], [13], and multimodal signals [1], [14], have also been investigated. Compared with speech signals, information captured in eye contact, head movement or respiration provides different informative clues. An advantage of these non-speech signals is that they contain useful cues of not only the speaker but also the listener. In eye activity, gaze has been widely accepted as an important clue to predict the turns during a conversation. However, other eye-related information, like blink and pupillary response, have not been fully researched to aid turn-taking prediction. The reason for selecting eye activity apart from speech is that it is 'always-on', can be acquired from mobile and wearable devices, and can be analyzed continuously even when speech is inactive, making it attractive to investigate herein. Also, when backgrounds are crowded and noisy, speech may not be reliably available for analysis.

2. Related Work

Speech processing and nonverbal behavioral signals have previously been researched for turn-taking detection. In speech processing, most studies (e.g. [2], [7], [15]) focused on detecting the end of a turn using prosodic, acoustic and syntactic features extracted from the participant's speech. For example, in [2], conversations were segmented into inter-pausal units (IPU) and turn-yielding cues were extracted from acoustic, prosodic, syntactic sources. Their results showed that a linear-kernel SVM classifier achieved the highest accuracy. Considering the influence of previous speaking activity on the turn-taking prediction, Skantze [16] applied Recurrent Neural Networks with Long Short-Term Memory (LSTM) as a general continuous model of turn-taking. There is no denying the importance of the speaker's speech on turn-taking prediction, but the state of his/her partner (i.e. the listener) may also be valuable to explore. To our knowledge, how listener's state contributes to their decision to end their turn has not been investigated.

Among nonverbal behaviors, eye gaze, head movement and other physical motion have been evaluated during turn-taking. Many studies have suggested that eye gaze is an effective nonverbal behavioral feature [17] for turn-taking prediction. It can be more useful when combined with prosodic features [4]. As eye gaze is under voluntary control, its patterns can be easily influenced by individual preferences and conversational contexts, such as remote communication via teleconference or teleconsultation. In [18], the pattern of eye gaze together with head movement and perceived emotion were explored in multi-party conversation conditions, showing that turn-taking detection performances with and without gaze features were not significantly different. This indicates that eye gaze may not be a consistently effective feature for turn-taking detection due to its task-specific nature.

Unlike gaze, pupillary response and blink are two kinds of eye activity which are involuntary and less task-specific, but have seldom been investigated in turn-taking study although they have been often employed in mental workload estimation [19]–[24]. Different studies [19]–[21] have suggested that pupil diameter represents a good index of mental activity since the pupil dilates while we are performing more difficult cognitive tasks. Palinko et al. found that the pupil dilates more significantly during speaking than during listening [25]. Blink has also been researched and is believed to be a good communication index in dyadic conversation [26], and a kind of 'punctuation mark' of mental activity changes [22]. With recent advances in eye computing hardware and algorithms [27], eye activity can be easily recorded and acquired. A recent study [28] explored pupillary response and blink in low and high communication loads, showing that pupil size changes differently in listening and speaking segments during low communication load.

In controlled laboratory experiments, the conducted conversations are often not as natural as real conversations in our daily life [29], however herein we set few restrictions on physical movements and investigate eye activity change during dyadic conversations. Since eye activity can be obtained continuously from two interlocutors, it is of interest to explore whether information from two conversation sides could infer their future intention to speak, and to predict turn-taking. To our acknowledge, no paper has previously investigated the performance of pupillary response and blink on turn-taking detection, especially from the perspective of exploring the state of the current listener.

3. Proposed Prediction System

Since this paper aims to investigate whether including both interlocutors' states can indicate speech activity and turn-taking state, we proposed a system to predict speech and non-speech state of a future period of time by extracting their continuous audio-visual information and then deciding the turn-taking status based on associated rules. Figure 1 shows a block diagram of the proposed system which fuses information from two speakers.

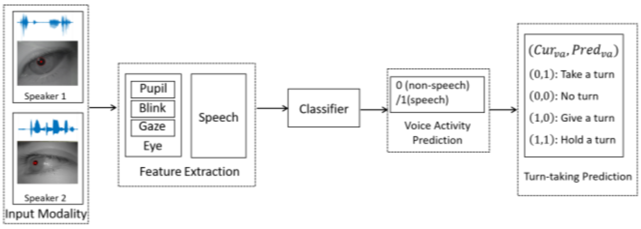


Figure 1: *Block diagram of the proposed system. Two speakers' speech and eye activity information comprise the input modalities. Based on the prediction of voice activity, four types of turn-taking status are then detected.*

The feature set of two speakers (S_i where $i = 1,2$) during a dyadic conversation are organized as $[Fea_{s1}^T,\ Fea_{s2}^T]^T$ at 20 fps which is shown in Figure 2, where Fea_{si}^T represents the features from the ith speaker. In this study, three main types of eye activity are extracted from wearable eye videos: pupil size, blink status and gaze direction. Considering the important role of speech in turn-taking, we also include current voice activity (speech (1)/no speech (0)) and prosodic features of two speakers in this system. Overlapped speech, which is inevitable in natural conversation, has often been ignored in speech processing system, however since both interlocutors' information is included in this system, there is no need to treat overlapped speech differently.

An intermediate output of the system is the prediction of voice activity (0 and 1) in the following period T, as seen in Figure 2. Based on the predictions $Pred_{VA}$ and the current voice activity Cur_{VA}, the future turn-taking onset is inferred. If Cur_{VA} is non-speech and $Pred_{VA}$ is speech state, this predicts that turn-taking will occur in the ensuing period of time T. If Cur_{VA} and $Pred_{VA}$ are both non-speech states, a no-turn status (i.e. listening) will continue. If Cur_{VA} is speech and $Pred_{VA}$ is non-speech, then turn-giving will occur next. If Cur_{VA} and $Pred_{VA}$ are both speech states, the turn will be held by the speaker.

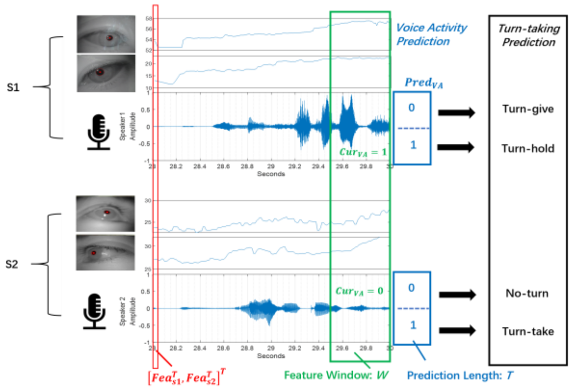


Figure 2: *Illustration of feature extraction from 2 speakers* $[Fea_{s1}^T,\ Fea_{s2}^T]^T$*, a feature window of duration W and the turn-taking prediction based on voice activity prediction. Red rectangle: extracted frame, green rectangle: a feature window W, blue rectangle: prediction for the future period T, black rectangle: turn-taking prediction based on the two states of voice activity.*

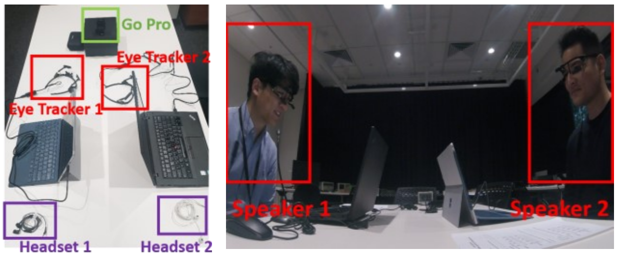
Different classification algorithms have been researched in this field, including logistic regression, support vector machine (SVM), but also recurrent neural networks. LSTMs also previously achieved excellent performance of turn taking using speakers' speech features only [16]. Herein, we employ SVM to predict the following speech state and also explore the performance using LSTM.

4. Experiment Setting

4.1. Data Collection

To investigate eye activity during natural conversation, we designed a conversational task and collected eye and speech recordings from 12 volunteers (9 females, 3 males, average age: 25.6). The task was a role play survival task, in which the participants' ship has been forced to land at a remote location 200 miles away from their destination. They were given 15 items from which they needed to select 10 items to take and needed to rank them in an order of importance. 12 Participants were randomly divided into 2-person groups. Each participant had 5 minutes to solve the survival task individually, and then they had a maximum of 10 minutes to discuss with their partner and reach a consensus about the selected items and ranking. For the purposes of this research, only the conversations were analyzed in this study.

Throughout the experiment, each participant was required to wear a close-talk headset and a glasses-like hardware (Pupil Labs [30]) with two small infrared webcams pointing towards the two eyes. A scene camera recorded the experiment and two laptops displayed the survival task prompts and recorded individual videos. The experiment was conducted in a quiet laboratory with ripple-free lighting. Figure 3(a) shows the equipment used, and Figure 3(b) shows a scene view of the data collection from one group. Differently from other eye activity datasets in previous studies, eye videos of both participants in the conversation were recorded in this dataset.



(a) (b)

Figure 3: *Illustration of the data collection. (a) The equipment in this data collection: two headsets, two Pupil Labs hardware and one Go-Pro. (b) Two participants completed the conversational task face-to-face.*

4.2. Data Pre-processing and Feature Extraction

Two video files were recorded from each participant. One was the eye video captured by a Pupil Labs eye tracker at 30fps, and the other one was the personal view captured by the laptop. To easily synchronise near-field eye videos and audios, we asked each participant to close their eyes and clap their hands three times at the start of the experiment. Based on the logged files provided by Pupil Labs software, the pupil sizes of the both eyes were extracted. To obtain accurate blink status and gaze direction, we annotated these videos using Anvil [31]. A binary feature was used to represent blink status (eye close/eye open). Three types of gaze direction (*GD*) were involved: looking at computer (1), partner (2) and other (3), which was manually annotated. Three raw features, i.e. pupil size, blink and gaze, were down sampled to 20fps for analysis. Besides the near-field eye features, prosodic features were also extracted from speech recordings as paper [2]: intensity, pitch, jitter, shimmer and noise-to-harmonics ratio (NHR) were extracted using openSMILE [32]. Voice activity was extracted from the defined Inter-Pausal Units (IPU) using SPPAS [33].

In this study, a total of 117 minutes of data were collected from 12 participants (6 pairs of conversations). Due to individual pupil size differences, the percent change in pupil size (*PC*) relative to a pupil size baseline was adopted, like prior studies [34]. $P_{current}$ and $P_{baseline}$ represent the current and baseline pupil size respectively, where $P_{baseline}$ is the average pupil size over the 10s before the start of the conversation.

$$PC = (P_{current} - P_{baseline})/P_{baseline} \qquad (1)$$

To extract more information from the pupillary response, a feature window (see in Figure 2) was used to extract average percent change in pupil size (*MeanPC*), range of percent change in pupil size (*RP*), standard deviation of percent change in pupil size (*STD*), maximum percent change in pupil size (*MaxPC*), minimum percent change in pupil size (*MinPC*) from the raw pupil size data. For gaze, since each frame was annotated to three types of direction, a 3-dimensional one-hot feature was used. In each window, blink information (*BI*), *GD*, and speech features were averaged across all frames. Considering the average value of voice activity was used in feature window, Cur_{VA}, which we considered for two states in Section 3 (i.e. 0 or 1), had another state which between 0 and 1. Therefore, we added two rules: if Cur_{VA} is between 0 and 1 and $Pred_{VA}$ is 0, then we predicted turn-giving will occur in the following *T*. Otherwise, turn-taking will occur. For the voice activity of target label, a value of 1 was adopted if at least one frame during the prediction window *T* was speech.

In total, two main groups of features were extracted (see in Table 1) from the source of eye activity and speech. 36 features were extracted from two speakers' near-field eye videos and from the speech recordings during a conversation. As two speakers' data were all captured during a conversation, 12 participants' data were from 6 pairs of conversations. We employed 5-fold leave-two-pairs-out cross validation experiment, meaning that each pair of participants were included only in the training set or the test set.

Table 1: *Two groups of feature set from one speaker*

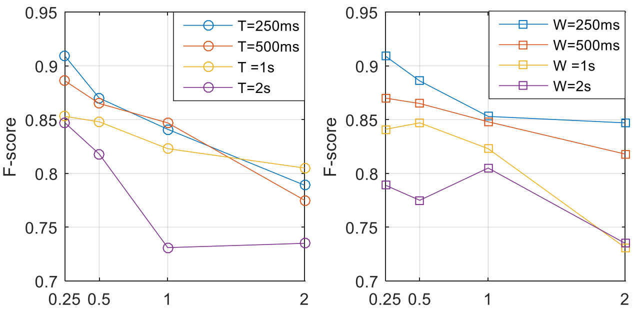
Eye Activity (12)			Speech (6)
Pupil Response (from two eyes, 8)	Blink (1)	Gaze (3)	
MeanPC *RP* *MaxPC* *MinPC*	*BI*	*GD*	*Voice activity* *Intensity* *Pitch* *Jitter* *Shimmer* *NHR*

5. Results and Discussion

5.1. Prediction of Voice Activity based on SVM

5.1.1. Feature Window Duration and Prediction Length

Two parameters exist in the proposed system: feature window duration *W* and prediction length *T*, and 4 values (i.e. 0.25s, 0.5s, 1, 2s) were chosen for both. Smaller values were not considered, for example, if *W* is smaller than 0.25s then the signal extracted from each window will be less than 4 frames, during which eye-related activity does not change much. F-scores of predictions for varying *W* and *T* are shown in Figure 4. Figure 4(a) shows that the accuracy drops with a larger *W*. All four prediction lengths have the highest accuracy when *W* is 0.25s. The highest accuracy achieved is 90.9% when *W* is 0.25s and *T* is 0.25s. Figure 4(b) shows a decreasing trend when predicting a further into the future (larger *T*). As discussed before, we found that when *W* is 2s, the accuracy is the lowest in Figure 4(b) and it may or may not reliably predict short times into the future.



(a) Feature window duration *W* (b) Prediction length *T*

Figure 4: *F-scores for (a) varying W in different T settings and (b) varying T in different W settings to explore the effect of two parameters in voice activity prediction*

5.1.2. One-side vs. Two-side

It is of interest to compare performance between using only one speaker's features and using both. We denote the feature window duration and prediction length setting as (W, T). Two settings are shown, based on favorable parameter choices from Figure 4. Voice activity prediction results are shown in Table 2, which demonstrate that using two-side features provides a relative reduction in error rate of 8.5% compared with using speech one-side features. It makes sense that the accuracy improves with more information obtained from both interlocutors.

Table 2: *Summary of accuracy comparisons for* ***one-side feature set*** *and* ***two-side feature set*** *in (0.25s, 1s) and (1s, 1s).*

	(0.25s, 1s)		**(1s, 1s)**	
	One-side	Two-side	One-side	Two-side
Precision	95.1%	95.8%	87.1%	88.5%
Recall	75.0%	76.9%	75.4%	76.9%
F-score	**83.8%**	**85.3%**	**80.8%**	**82.3%**

5.1.3. Eye Features vs. Speech Features

Comparing the use of eye activity only with speech only from two sides (Tables 3 and 4), speech features outperform eye-related features in voice activity prediction in general. With prediction length fixed at 1s, it seems that eye activity performs better for larger feature window durations (Table 3), but the speech shows the opposite trend (Table 4). It makes sense that better eye information can be extracted when the window is longer since pupil changes need time [35]. However, for speech, long feature windows may include too much variability.

Table 3: *Summary of accuracy comparisons for* ***eye activity***

	(0.25s, 1s)	**(0.5s, 1s)**	**(1s, 1s)**	**(2s, 1s)**
Precision	66.9%	60.7%	57.3%	58.0%
Recall	42.0%	46.9%	59.2%	72.3%
F-score	**51.6%**	**52.8%**	**57.7%**	**63.7%**

Table 4: *Summary of accuracy comparisons for* ***speech***

	(0.25, 1s)	**(0.5s, 1s)**	**(1s, 1s)**	**(2s, 1s)**
Precision	95.3%	92.3%	87.5%	81.0%
Recall	74.1%	74.6%	76.2%	77.3%
F-score	**83.4%**	**82.5%**	**81.4%**	**79.1%**

5.2. Prediction of Next Turn Onset based on Voice Activity Prediction

Based on the predicted voice activity, combined with current voice activity state, the turn-taking state can be inferred (Figure 2). According to the above investigation of the proposed system, we chose (1s, 1s) which had a high accuracy in voice activity prediction, but also good performance achieved by using both eye activity and speech features. For (1s, 1s), there are 174 turn-taking instances, 199 turn-giving instances, 978 no-turn instances and 745 turn-hold instances. Based on the rules, the prediction accuracies of the four types of states during a conversation are shown in Table 5. Hold and no-turn are easily predicted, while turn-taking and turn-giving are more difficult, although much higher than chance level. The accuracy is low because we checked the turn shifts on the exact time window which means we did not include the instances which may also detected the shifts but in advance.

Table 5: *Summary of accuracies of four types of decisions in turn-taking (1s, 1s)*

	Take	**Give**	**No turn**	**Hold**
Instances	174	199	978	745
Chance level	8.3%	9.5%	46.7%	35.5%
SVM (F-score)	**36.8%**	**31.5%**	**98.4%**	**82.8%**
LSTM (F-score)	**42.4%**	**59.8%**	**94.3%**	**79.3%**

5.3. LSTM Method

Regarding the high accuracy of LSTM method achieved in turn-taking prediction [16], we replaced SVM with LSTM as the backend to predict the voice activity and then used rules for the turn-taking decision. Similarly with [16], one LSTM layer and one dense layer were used. 40 hidden units were used in the LSTM layer. The tanh function was used in the LSTM layer and the sigmoid function was used in dense layer. The learning rate was set to 0.001 and the L2 regularization was set 0.001. Although the feature sets and data are not identical, the results for different T are consistent with results in [16]. The shorter the prediction length, the higher the accuracy that can be achieved; however, we used eye activity and speech features from two speakers. Turn-taking decisions were also checked by using the LSTM outputs. The results are shown in Table 5. Compared with voice activity using SVM, LSTM method can achieve a higher accuracy in turn-taking.

6. Conclusion

In this paper, we proposed a system to predict turn-taking states during dyadic conversation using low-cost wearable hardware. Voice activity, as an intermediate output of the proposed system, was predicted by extracting near-field eye activity and speech. To our knowledge, this is the first work to use pupil size and blink in voice activity and turn-taking prediction. Through experiments on a natural conversation dataset, we found that combining eye activity and speech is useful, improving voice activity prediction by a relative reduction in error rate of 5%-12.9% compared with speech features alone and 8.5% compared with one person's features. In voice activity prediction, we found that a shorter feature window and a shorter prediction can increase the accuracy when fusing eye activity and speech. For eye activity, a larger feature window duration provides more information while the speech features show an opposite trend. Based on the predicted voice activity and turn-taking rules, we can achieve turn-taking prediction accuracy that is much higher than the chance-level. This work shows the potential for using eye information to augment the performance in turn taking. One limitation of our work is that pupil size is sensitive to light and may be influenced by the gaze direction change between computer and partner, which might explain a lower performance than using speech features.

7. Acknowledgements

This work was supported in part by US Army ITC-PAC, through contract FA5209-17-P-0154. Opinions expressed are the authors' and may not reflect those of the US Army. The work was also partly supported by Data61, CSIRO, Australia.

8. References

[1] I. de Kok and D. Heylen, "Multimodal end-of-turn prediction in multi-party meetings," *Proc. 2009 Int. Conf. Multimodal interfaces - ICMI-MLMI '09*, p. 91, 2009.

[2] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Comput. Speech Lang.*, vol. 25, no. 3, pp. 601–634, 2011.

[3] H. Wesselmeier and H. M. Müller, "Turn-taking: From perception to speech preparation," *Neurosci. Lett.*, vol. 609, pp. 147–151, 2015.

[4] T. Kawahara, T. Iwatate, and K. Takanashi, "Prediction of Turn-Taking by Combining Prosodic and Eye-Gaze Information in Poster Conversations," *Interspeech*, pp. 727–730, 2012.

[5] N. G. Ward, O. Fuentes, and A. Vega, "Dialog Prediction for a General Model of Turn-Taking," *Proc. Interspeech-2010*, no. September, pp. 2662–2665, 2010.

[6] G. Skantze, A. Hjalmarsson, and C. Oertel, "Turn-taking, feedback and joint attention in situated human-robot interaction," *Speech Commun.*, vol. 65, pp. 50–66, 2014.

[7] D. Schlangen and D. Schlangen, "From Reaction To Prediction Experiments with Computational Models of Turn-Taking," *Word J. Int. Linguist. Assoc.*, pp. 1–4, 2006.

[8] S. C. Levinson, "Turn-taking in Human Communication - Origins and Implications for Language Processing," *Trends Cogn. Sci.*, vol. 20, no. 1, pp. 6–14, 2016.

[9] R. Hariharan, J. Hakkinen, and K. Laurila, "Robust end-of-utterance detection for real-time speech recognition applications," *2001 IEEE Int. Conf. Acoust. Speech, Signal Process. Proc. (Cat. No. 01CH37221)*, vol. 1, pp. 249–252, 2001.

[10] Q. Li, J. Zheng, Q. Zhou, and C. H. Lee, "A robust, real-time endpoint detector with energy normalization for ASR in adverse environments," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 1, pp. 233–236, 2001.

[11] M. Garau, M. Slater, S. Bee, and M. A. Sasse, "The impact of eye gaze on communication using humanoid avatars," *Proc. SIGCHI Conf. Hum. factors Comput. Syst. CHI 01*, pp. 309–316, 2001.

[12] R. Ishii, S. Kumano, and K. Otsuka, "Predicting next speaker based on head movement in multi-party meetings," *2015 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 2319–2323, 2015.

[13] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, "Analysis of Respiration for Prediction of 'Who Will Be Next Speaker and When?' in Multi-Party Meetings," *NTT Tech. Rev.*, vol. 13, no. 7, pp. 18–25, 2015.

[14] R. Ishii, S. Kumano, and K. Otsuka, "Multimodal Fusion using Respiration and Gaze for Predicting Next Speaker in Multi-Party Meetings," *Proc. 2015 ACM Int. Conf. Multimodal Interact. - ICMI '15*, pp. 99–106, 2015.

[15] L. Ferrer, E. Shriberg, and A. Stolcke, "Is the speaker done yet? Faster and more accurate end-of utterance detection using prosody," *Proc. ICSLP*, pp. 2061–2064, 2002.

[16] G. Skantze, "Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks," no. August, pp. 220–230, 2018.

[17] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, "Analysis and modeling of next speaking start timing based on gaze behavior in multi-party meetings," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 694–698.

[18] K. Jokinen, M. Nishida, and S. Yamamoto, "Eye-gaze experiments for conversation monitoring," *ACM Int. Conf. Proceeding Ser.*, no. May 2014, pp. 303–308, 2009.

[19] M. K. Eckstein, B. Guerra-Carrillo, A. T. Miller Singley, and S. A. Bunge, "Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?," *Dev. Cogn. Neurosci.*, vol. 25, pp. 69–91, 2017.

[20] B. Laeng, M. Ørbo, T. Holmlund, and M. Miozzo, "Pupillary stroop effects," *Cogn. Process.*, vol. 12, no. 1, pp. 13–21, 2011.

[21] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources," *Psychol. Bull.*, vol. 91, no. 2, pp. 276–292, 1982.

[22] G. J. Siegle, N. Ichikawa, and S. Steinhauer, "Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses," *Psychophysiology*, vol. 45, no. 5, pp. 679–687, 2008.

[23] S. Chen and J. Epps, "Using task-induced pupil diameter and blink rate to infer cognitive load," *Human-Computer Interact.*, vol. 29, no. 4, pp. 390–413, 2014.

[24] A. Hall, "The origin and purposes of blinking," *Br. J. Ophthalmol.*, vol. 29, no. 9, p. 445, 1945.

[25] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman, "Estimating cognitive load using remote eye tracking in a driving simulator," *Eye Track. Res. Appl. Symp.*, pp. 141–144, 2010.

[26] K. Hirokawa, A. Yagi, and Y. Miyata, "Comparison of blinking behavior during listening to and speaking in Japanese and English," *Percept. Mot. Skills*, vol. 98, pp. 463–472, 2004.

[27] S. Chen and J. Epps, "Efficient and Robust Pupil Size Estimation and Blink Detection From Near-Field Video Sequences," *IEEE Transit. Cybern.*, vol. 44, no. 12, pp. 2356–2367, 2014.

[28] H. Li, J. Epps, and S. Chen, "Think before you speak: An investigation of eye activity patterns during conversations using eyewear," *Int. J. Hum. Comput. Stud.*, vol. 143, no. May, p. 102468, 2020.

[29] R. Cañigueral, A. Hamilton, and J. A. Ward, "Don't look at me, I'm wearing an eyetracker!," *UbiComp/ISWC 2018 - Adjun. Proc. 2018 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. Proc. 2018 ACM Int. Symp. Wearable Comput.*, pp. 994–998, 2018.

[30] M. Kassner, W. Patera, and A. Bulling, "Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction," pp. 1151–1160, 2014.

[31] M. Kipp, "Anvil: The video annotation research tool," *Handb. corpus Phonol.*, pp. 420–436, 2014.

[32] F. Eyben *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[33] B. Bigi, "SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech.," *the Phonetician*, pp. 54–69, 2015.

[34] B. P. Bailey and S. T. Iqbal, "Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management," *ACM Trans. Comput. Interact.*, vol. 14, no. 4, pp. 1–28, 2008.

[35] H. K. Wong and J. Epps, "Pupillary transient responses to within-task cognitive load variation," *Comput. Methods Programs Biomed.*, vol. 137, pp. 47–63, 2016.