The relationship between turn-final eye gaze and floor transfer offset

Tim Zee

Abstract

This paper investigates whether the floor transfer offset (FTO) of turn switches is related to eye gaze behaviour in conversational speech. Previous research has claimed that mutual gaze and looking towards a listener signals the end of a turn and that the absence of these cues results in slower turn switches. Decision trees and random forest analyses of dyadic Dutch conversations reveal that both listener and speaker gaze play an important role in the duration of the pause or overlap between turns. However, the findings do not show that mutual gaze and looks toward the listener are associated with particularly low FTO values, which suggests that eye gaze patterns do not simply reflect turn-taking strategies.

Keywords: turn-taking; floor transfer offset; eye gaze

Introduction

The *floor transfer offset* (FTO), or *turn switch delay*, refers to the delay or overlap between two consecutive turns at talk by different speakers. Previous research by Roberts, Torreira and Levinson (2015) has shown that this variable—as it occurs in conversational English—is related to cognitive limitations on the speakers as well as the "turn-allocation techniques" (Sacks et al., 1974) and *sequence organisation* methods that they employ. A different strand of research has investigated whether eye gaze patterns can be used to predict turn switches. Eye gaze data from multiparty conversations suggest that mutual gaze is used to signal a turn switch (Jokinen et al., 2010). Similarly, it has been claimed for dyadic conversations that speakers tend to look towards their interlocutor at the end of their turn (e.g. Kendon, 1967, p. 33) and that the absence of speaker gaze co-occurs with a "delayed response" (Kendon, 1967, pp. 36-37). However, subsequent research has suggested that this turn-final speaker gaze is not a consistent pattern (Torres et al., 1997). Instead, it has been proposed that gaze behaviour is influenced by whether a speaker expects a response from the listener (Rossano, 2013, p. 317). It can be concluded that the role of eye gaze in turn-taking is not completely clear yet, let alone its relationship to the FTO between turns. Moreover, previous research (Roberts et al., 2015) reminds us that turn-taking involves a plethora of other variables that should be taken into account when further exploring this topic.

This research explores whether a relationship between turn-final eye gaze and turn switches is reflected in the FTOs in conversational Dutch. Furthermore, this relationship is contextualised by investigating its importance compared to both sequence organisation and processing measures.

Methods

This section will go into more detail about the conversational data that was used, how the relevant variables were defined, and which statistical methods were used to analyse the data.

Materials

A large amount of spontaneous speech was needed in order to investigate turn-taking patterns in a naturalistic context. Furthermore, this speech data had to be richly annotated to allow for the extraction of gaze, sequence organisation, and processing information. The current study used the IFADV corpus (van Son et al., 2009), because it meets these requirements.

The IFADV corpus consists of audio and video recordings of 20 conversations between two friends or colleagues. The participants sat at opposite ends of a table and were told to talk about anything they liked. For each conversation the first 15 minutes were annotated for the features in Table 1.

Table 1: Annotations in the IFADV corpus.

Annotation	*Type*
Orthography	Manual
Part-of-Speech tags and Lemmas	Automatic
Word alignment	Automatic
Phoneme alignment	Automatic
Conversational function	Manual
Gaze direction	Manual
End intonation	Automatic

However, for a number of conversations some of these annotations were missing, resulting in a total of 8 usable conversations.

Most annotations were available in the TextGrid format of the Praat speech analysis software (Boersma & Weenink, 2017). For convenience, the remaining annotations were converted to the TextGrid format by means of a Python (available at www.python.org) script. Subsequently, the FTOs and relevant contextual information was automatically extracted from these files using a combination of Python and Praat scripts.

Variable definitions

This subsection will introduce all variables that were extracted and provides the formal definitions that were used to estimate their values. If possible variables were given identical definitions to those used in the study by Roberts and colleagues (2015).

Floor Transfer Offset A turn was defined as a sequence of speech by a single speaker (the turn holder) that was not interrupted by the speech of the other person. Stretches of speech that were completely overlapped by the turn holder's speech did not count as a change of speaker. In other words, each change of speaker resulted in the start of a new turn. The

FTO was determined by subtracting the start time of the next turn (T2) from the end time of the present turn (T1). As a result, the FTO could have a positive or negative value depending on whether T2 started after or before T1 ended.

These values were based on the automatic word alignments, because the boundaries of the manual transcriptions often included a small stretch of silence at start and end of turns. Furthermore, the mean FTO value based on automatic annotations, 146.40 ms, was closer to a previously reported FTO value for Dutch, 108.93 ms (Stivers et al., 2009), compared to the mean value based on manual transcriptions, 15.60 ms.

Eye Gaze Previous claims concerning the role of eye gaze in turn-taking mostly pertained either the presence of mutual gaze or a change in gaze direction at the end of turns (Rossano, 2013). In order to allow for the analysis of these patterns both speaker and listener gaze at the end of turns was classified as *constant gaze at interlocutor (g)*, *constant gaze away from interlocutor (x)*, *look toward interlocutor (x-g)*, *look away from interlocutor (g-x)*. Following Torres and colleagues (1997), the end of a turn was defined as the final word of that turn.

Pitch As previous research has shown that the end of a speech chunk is reflected in the accompanying intonation (Caspers, 2003), the final pitch of each turn was included as a variable. This variable was included as an annotation to the IFADV corpus and its formal definition is given in Van Son et al. (2009, p. 25).

Sequence organisation These measures were extracted from the existing annotations of conversational function for each utterance. This resulted in 3 dummy variables for the final utterance in T1 and the first utterance in T2 respectively. One variable indicated whether the utterance prompted a response or not (initiating). A second variable indicated the presence of a response (responding). A third variable codified whether it was a backchannel or not (backchannel).

Processing: Turn duration The duration of T1 and T2 could easily be calculated by subtracting their respective start times from their respective end times.

Processing: Frequency For each turn the mean word frequency was calculated. This involved extracting the lemma and Part-of-Speech (POS) tag corresponding to each word token in the corpus in order to determine the frequency for each lemma-POS combination. For many words the lemma annotation was missing. These lemmas were automatically extracted from the SUBTLEX-NL corpus (Keuleers, Brysbaert & New, 2010).

Processing: Speech rate Following Wightman, Shattuck-Hufnagel, Ostendorf and Price (1992) The mean speech rate of a turn was based on the proportion between expected phoneme duration and mean phoneme duration. This relation can be expressed formally as in (1), where $\hat{\alpha}$ stands for speech rate, N stands for number of segments, d stands for duration of segment i, and μ_p represents the mean duration of a given phoneme.

$$(1)\ \hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i}{\mu_{p_i}}$$

Processing: Clauses For each turn the number of clauses were determined based on the clause demarcations that were used to generate the POS annotation.

Processing: Concreteness Using the concreteness ratings for 30 000 Dutch words collected by Brysbaert, Stevens, De Deyne, Voorspoels and Storms (2014), a mean rating was calculated for each turn. If a word was not included in the corpus by Brysbaert et al., it did not count towards the mean concreteness score.

Other variables In addition to nonverbal cues, processing measures, and sequence organisation variables, the current analysis included the duration of the conversation at the time of the turn switch and the sex of both participants.

Statistical methods

Due to the large amount of variables—and the even greater amount of interactions—that potentially influence the FTO values in the IFADV corpus, a conventional regression analysis of the present dataset would be hard to interpret. More importantly, such an approach would not allow for robust comparisons of different variables given the many associations between the predictors (Roberts et al., 2015, p. 7; see Appendix A). Therefore, conditional decision trees are used to explore possible trends concerning gaze patterns and a random forest analysis is conducted to estimate the relative importance of gaze behaviour.

A conditional decision tree works by iteratively splitting the data into subsets based on the variable that explains the most variation in the FTO, but only if the null hypothesis that the FTO and a predictor are independent is not supported at a given significance level (Hothorn, Hornik & Zeileis, 2006, p. 655). Although the resulting tree allows for relatively straightforward interpretations of complex interactions, the effects it shows are still sensitive to multicollinearity in the data. Furthermore, the structure of single decision trees can be highly variable depending on small changes in the data because decisions made higher up in the hierarchy influence those that are made in lower levels. In order to prevent these problems a random forest design contains hundreds of these decision trees, each of which is based on a random sample of the data and a given number of randomly picked predictors. For each variable an estimate of importance can be made by taking all trees that include that variable and calculating the mean increase in prediction error of those trees if that variable is randomly permuted (Roberts et al., 2015. p. 9).

The single decision trees presented in the current research were implemented using the `partykit` package (Hothorn &

Zeileis, 2015) in the R data analysis software (R Core Team, 2016). All trees had a conditional *p*-value of .05. The random forest analysis was implemented using R package `party` (Strobl, Hothorn, Zeileis, 2009) with the number of decision trees set to 500 and the number of variables chosen for each tree set to 3.

Results

This section first presents descriptive statistics on the variables in the data set. Subsequently, single decision trees are presented to illustrate the trends of eye gaze and its interaction with sequence organisation and processing variables. Finally, the random forest analysis is presented to give an indication of the relative importance of gaze patterns for FTO.

Descriptive statistics

Tables 2-5 summarize the distribution of all variables in the analysis.

Table 2: Descriptive statistics of continuous variables.

Variable	*Min*	*Max*	*Mean*	*SD*
FTO (s)	-2.13	2.62	0.15	0.56
Conv. time (s)	0.35	904.00	435.95	260.55
T1 duration (s)	0.05	50.93	3.49	4.83
T2 duration (s)	0.05	50.93	3.50	4.84
T1 frequency	1	2702	820.73	790.35
T2 frequency	1	2702	822.23	791.06
T1 speech rate	0.47	26.36	1.23	1.03
T2 speech rate	0.47	26.36	1.23	1.03
T1 clauses	1	25	2.07	2.07
T2 clauses	1	25	2.07	2.08
T1 concreteness	1	4.80	1.81	0.39
T2 concreteness	1	4.80	1.81	0.39

Table 3: Counts of categorical non-verbal communication variables.

Variable	*g*	*x*	*g-x*	*x-g*
Speaker gaze	1511	300	70	80
Listener gaze	1425	311	161	66
Variable	*low*	*mid*	*high*	
End pitch	576	1017	376	

Table 4: Counts of categorical sequence organisation variables.

Variable	*Yes*	*No*
T1 initiating	267	1702
T2 initiating	213	1756
T1 responding	138	1831
T2 responding	242	1727
T1 backchannel	397	1572
T2 backchannel	450	1519

Table 5: Counts of speaker sex.

Variable	*Female*	*Male*
T1 sex	1182	787
T2 sex	1184	785

Many of the predictor variables in this study are correlated. This is reflected in Table 6, which shows the mean association effects between different types of variables.

Table 6: Mean association scores derived from the effect sizes in Appendix A.

	Processing	Sequence organisation	Non-verbal cues	Other
Processing	.15	.11	.07	.04
Sequence organisation		.17	.06	.04
Non-verbal communication			.04	.07
Other				.04

This table shows that the strongest associations are between different processing variables and sequence organisation variables. It also shows that eye gaze and turn-final pitch do not have very strong associations with either processing or sequence organisation variables.

Eye gaze

The decision tree in Figure 1 splits the data based on gaze patterns only.

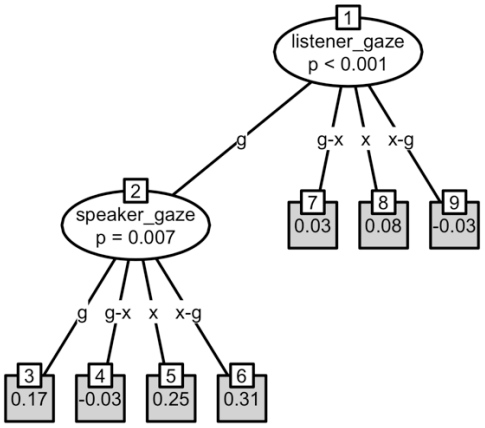

Figure 1: A decision tree of the data exclusively based on eye gaze patterns. The numbers in the grey boxes represent predicted FTO values in seconds.

The tree shows that both listener and speaker gaze make significant contributions to the model. The first splits, which are based on the gaze patterns of the listener, show that if the listener looks towards the speaker (*x-g*), the predicted FTO value is lowest, and if the speaker consistently looks away (*x*) a higher FTO value is predicted. However, if the listener directs a constant gaze towards the speaker, the predicted FTO value depends on the gaze pattern of the speaker. Interestingly, a look-away pattern (*g-x*) is associated with the lowest FTO value and a look-towards pattern is predicted to have a relatively high FTO.

Figure 2 shows a decision tree that was based on eye gaze patterns as well as all sequence organisation variables.

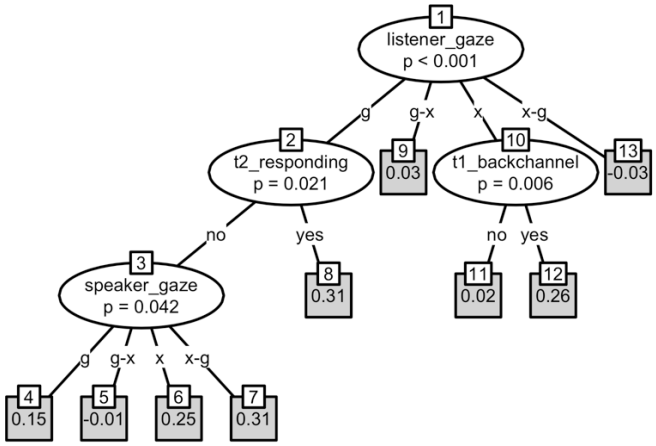

Figure 2: A decision tree of the data based on eye gaze patterns and sequence organisation. The numbers in the grey boxes represent predicted FTO values in seconds.

This decision tree has the same basic structure as the one in Figure 1, but it provides additional nuance for two nodes. Firstly, when the listener has a constant gaze towards the speaker, this tree distinguishes between responses and non-responses by the listener. In case of a response, a relatively high FTO is expected. When the second turn does not consist of a response the familiar general pattern based on speaker gaze applies. Secondly, if the listener has constant gaze away from the speaker the predicted FTO depends on whether the first turn is a backchannel (high FTO) or not (low FTO).

Figure 3 shows a decision tree that was based on processing variables in addition to eye gaze patterns.

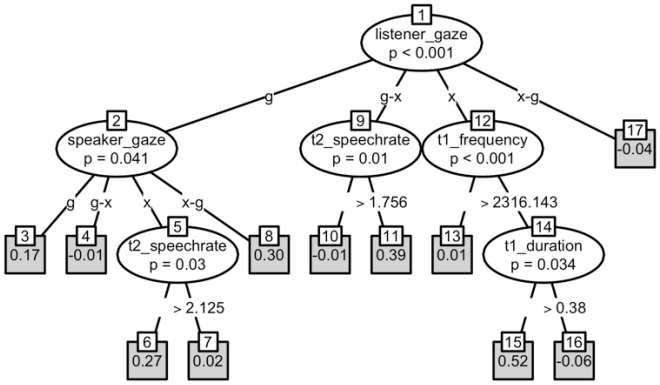

Figure 3: A decision tree of the data based on eye gaze patterns and processing. The numbers in the grey boxes represent predicted FTO values in seconds.

The figure shows that T2 speech rate, T1 word frequency and T1 duration interact with gaze patterns in certain ways. For instance, if a listener looks away from the speaker, the FTO is predicted to be relatively high if that listener's next turn has a high speech rate (above 1.756). However, the opposite pattern holds if the listener shows a constant gaze towards a speaker who does not look at the listener.

Variable Importance

By using a random forests approach, a robust model incorporating all predictor variables was created. Predictions made by this model were correlated to the actual FTOs at r = .66. Furthermore, this model was used to estimate the importance of the different variables. These values were used to make a ranking of importance, as exemplified in Figure 4.

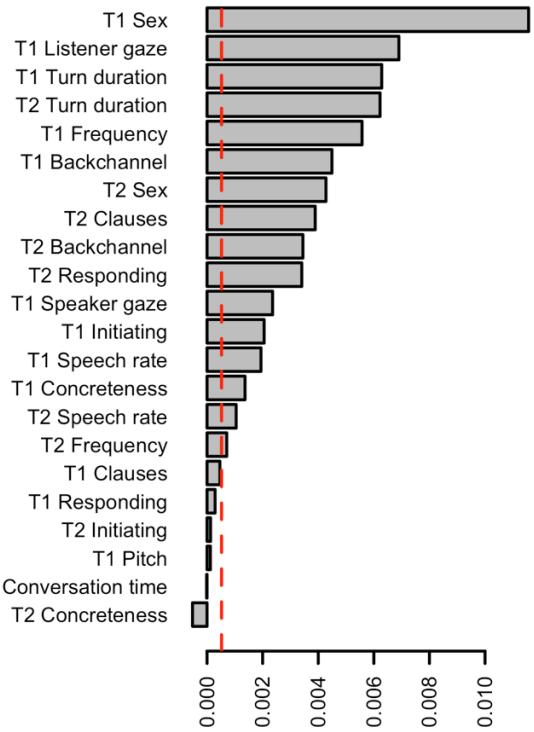

Figure 4: Variable importance measured as mean increase in prediction error when a variable is randomly permuted. Variables that do not cross the threshold indicated by the red line, do not make an important contribution to the model.

In addition to highlighting the importance of eye gaze, this ranking has some key differences to previous findings. Roberts et al. (2015) found that *T1 Responding* was the most important variable in their data, whereas the current research ranks that variable below the importance threshold. However, a medium-sized correlation (r = .36) does exist between this ranking and the ranking in Roberts et al. (2015).

Discussion

The decision tree analysis did not find that mutual gaze or a look toward the listener by the speaker result in particularly low FTOs, as might be expected if these gazing behaviours are solely associated with the end of the speaker's turn. Furthermore, the results show that both sequence organisation and processing variables interact with gaze patterns regarding FTO values. These results are in line with claims that gaze patterns are not simply related to turn-taking strategies but reflect the social interactions expressed by those turns (Rossano, 2012).

Finally, the variable importance results highlight the need to incorporate eye gaze as a factor in FTO research.

References

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer Software]. Retrieved from http://www.praat.org/.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica, 150*, 80-84.

Caspers, J. (2003). Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics, 31*, 251-276.

Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research, 16*, 3905-3909.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics, 15*(3), 651-674.

Jokinen, K., Harada, K., Nishida, M., Yamamoto, S. (2010, September). *Turn-alignment using eye-gaze and speech in conversational interaction*. Paper presented at the 11th International Conference on Spoken Language Processing, Makuhari, Japan.

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

Keuleers, E., Brysbaert, M. & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods, 42*(3), 643-650.

R Core Team. (2016). R: A language and environment for statistical computing [Computer Software]. Retrieved from https://www.R-project.org/.

Roberts, S.G., Torreira, F., & Levinson, S.C. (2015). The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology, 6*, 1-16.

Rossano, F. (2012). *Gaze behavior in face-to-face interaction*. Nijmegen: Ipskamp Drukkers.

Rossano, F. (2013). Gaze in conversation. In J. Sidnell & T. Stivers (Eds.), *The Handbook of Conversation Analysis* (pp. 308-329). Chicester: Blackwell.

Sacks, H., Schlegloff, E.A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696-735.

Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J.P., Yoon, K., & Levinson, S.C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences, 106*(26), 10587-10592.

Strobl, C., Hothorn, T., & Zeileis, A. (2009). Party on!—A new, conditional variable-importance measure for random forests available in the party package. *The R Journal, 1-2*, 14-17.

Torres, O., Cassell, J., & Prevost, S. (1997, July). *Modeling gaze behavior as a function of discourse structure*. Paper presented at the first International Workshop on Human-Computer Conversation, Bellagio, Italy.

van Son, R., Wesseling, W., Sanders, E., & van den Heuvel, H. (2009). Promoting free dialog video corpora: The IFADV corpus example. In M. Kipp, J. Martin, P. Paggio & D. Heylen (Eds.), *Multimodal Corpora* (pp. 18-37). Berlin: Springer-Verlag.

Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P.J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America, 91*(3), 1707-1717.

Appendix A

Association between variables. Measures: Pearson's *r* for 2 continuous variables (absolute value), Pearson's *r* for 1 continuous and 1 categorical variable, and Cramér's *V* for 2 categorical variables.

	Speaker gaze	Listener gaze	Conversation time	T1 Sex	T2 Sex	T1 Initiating	T2 Initiating	T1 Responding	T2 Responding	T1 Backchannel	T2 Backchannel	T1 Turn duration	T2 Turn duration	T1 Frequency	T2 Frequency	T1 Speech rate	T2 Speech rate	T1 Clauses	T2 Clauses	T1 Concreteness	T2 Concreteness
Pitch	.06	.03	.04	.04	.07	.13	.02	.06	.08	.06	.04	.04	.06	.05	.10	.04	.03	.04	.05	.06	.04
Sp. gaze		.04	.09	.08	.07	.10	.02	.02	.09	.11	.07	.06	.05	.09	.06	.05	.09	.06	.03	.07	.06
Listener gaze			.08	.09	.11	.03	.10	.06	.02	.05	.11	.05	.21	.04	.12	.15	.06	.03	.20	.07	.07
Conversation time				.02	.02	.06	.03	.04	.05	.01	.02	.06	.04	.03	.02	.01	.02	.06	.04	.02	.02
T1 Sex					.07	.03	.01	.02	.01	.10	.09	.08	.01	.15	.09	.06	.00	.03	.01	.01	.03
T2 Sex						.01	.00	.02	.01	.09	.09	.01	.08	.09	.15	.00	.06	.01	.03	.03	.01
T1 Initiating							.07	.03	.81	.20	.20	.05	.03	.12	.06	.08	.01	.06	.02	.07	.08
T2 Initiating								.08	.05	.09	.19	.07	.13	.08	.10	.01	.05	.07	.12	.08	.08
T1 Responding									.04	.14	.00	.11	.06	.04	.04	.02	.00	.11	.05	.04	.02
T2 Responding										.17	.20	.05	.01	.13	.03	.07	.01	.04	.01	.07	.05
T1 Backchannel											.23	.29	.24	.60	.19	.22	.05	.19	.21	.11	.02
T2 Backchannel												.24	.25	.21	.57	.08	.21	.19	.15	.10	.13
T1 Turn duration													.16	.27	.17	.13	.07	.91	.14	.10	.01
T2 Turn duration														.16	.27	.12	.13	.14	.91	.07	.10
T1 Frequency															.14	.09	.02	.19	.12	.39	.00
T2 Frequency																.06	.09	.11	.19	.09	.39
T1 Speech rate																	.01	.12	.10	.04	.00
T2 Speech rate																		.08	.12	.01	.05
T1 Clauses																			.10	.07	.02
T2 Clauses																				.05	.07
T1 Concreteness																					.04
T2 Concreteness																					

Sign. at p < .05 | Sign. at p < .01 | Sign. at p < .001