Turn-alignment using eye-gaze and speech in conversational interaction

Kristiina Jokinen¹, Kazuki Harada², Masafumi Nishida², Seiichi Yamamoto²

¹Department of Speech Sciences, University of Helsinki, Finland
²Department of Information Systems Design, Doshisha University, Japan

Abstract

Spoken interactions are known for accurate timing and alignment between interlocutors: turn-taking and topic flow are managed in a manner that provides conversational fluency and smooth progress of the task. This paper studies the relation between the interlocutors’ eye-gaze and spoken utterances, and describes our experiments on turn alignment. We conducted classification experiments by Support Vector Machine on turn-taking using the features for dialogue act, eye-gaze, and speech prosody in conversation data. As a result, we demonstrated that eye-gaze features are important signals in turn management, and seem even more important than speech features when the intention of utterances is clear.

Index Terms: eye-gaze, dialogue, interaction, speech analysis, turn-taking

1. Introduction

The role of eye-gaze in fluent communication has long since been acknowledged ([2]; [7]). Previous research has established close relations between eye-gaze and conversational feedback ([3]), building trust and rapport, as well as focus of shared attention ([15]). Eye-gaze is also important in turn-taking signalling: usually the interlocutors signal their wish to give the turn by gazing up to the interlocutor, leaning back, and dropping in pitch and loudness, and the partner can, accordingly, start preparing to take the turn. There is evidence that lack of eye contact decreases turn-taking efficiency in video-conferencing ([16]), and that the coupling of speech and gaze streams in a word acquisition task can improve performance significantly ([11]).

Several computational models of eye-gaze behaviour for artificial agents have also been designed. For instance, [9] describe an eye-gaze model for believable virtual humans, [13] demonstrate gaze modelling for conversational engagement, and [10] built an eye-gaze model to ground information in interactions with embodied conversational agents.

Our research focuses on turn-taking and eye-gaze alignment in natural dialogues and especially on the role of eye-gaze as a means to coordinate and control turn-taking. In our previous work [5,6] we noticed that in multi-party dialogues the participants head movement was important in signalling turn-taking, maybe because of its greater visibility than eye-gaze. (This is in agreement with [12], who noticed that in virtual environments, head tracking seems sufficient when people turn their heads to look but if the person is not turning their head to look at an object, then eye-tracking is important to discern the gaze of a person.) The main objective in the current research is to explore the relation between eye-gaze and speech, in particular, how the annotated turn and dialogue features and automatically recognized speech properties affect in turn management. Methodologically our research relies on experimentation and observation: signal-level measurements and analysis of gaze and speech are combined with human-level observation of dialogue events (dialogue acts and turn-taking).

We use our three-party dialogue data that is analysed with respect to the interlocutors’ speech, and annotated with dialogue acts, eye-gaze, and turn-taking features [6]. The experiments deal with the classification of turn-taking events using the analysed features and the results show that eye-gaze speech information significantly improves the accuracy compared with the classification with dialogue act information only. However, what is also interesting that the difference between gaze and speech features is not significant, i.e. eye-gaze and speech are important signals in turn management, but their effect is parallel rather than complementary. Moreover, eye-gaze seems to more important than speech when the intention of the utterance is clear.

The paper is structured as follows. We first describe the research on turn-taking and the alignment of speech and gaze in Section 2. We then present our data and speech analysis in Section 3, and experimental results as well as discussion concerning their importance in Section 4. Section 5 draws conclusions and points to future research.

2. Alignment and turn-taking

We base our analysis on the hypothesis that human–human interaction is cooperative activity which emerges from the speakers’ capability to act in a relevant and rational manner ([1]; [4]). The basic enablements of communication, Contact, Perception, and Understanding (CPU) must hold for the interaction to proceed smoothly, and consequently, the agents’ cooperation manifests itself to the extent in which the agents can observe and provide relevant feedback on the CPU enablements. Such aspects as looking at the conversational partner or looking away provide indirect cues of the partner’s willingness to continue interaction, whereas gazing at particular elements in the vision field tells us what the partner’s focus of attention is, and thus they give guidance for appropriate presentation of information as well as suitable analysis and response to the partner’s contribution. The agents thus constantly monitor each other, the partner’s activities, and the communicative situation and, if some of the enablements are not fulfilled, react to the problems.

We assume that turn management is the interlocutors’ coordinated action of speaking and listening so that only one of the interlocutors holds the floor at any one time. Natural conversations also contain overlaps and silences which can give feedback about the CPU and the participants’ emotional stance: they signal excitement, cooperation, ignorance, etc. Usually they are very short as the speakers take their partners’ enablements into consideration: it is impossible to get one’s message across if the speakers speak at the same time.
As discussed above, eye-gaze and speech prosody are recognized as important signals for turn-taking, besides grammatical structure (i.e. completed or incomplete sentence structure). Timing of the signals is also important: successful turn-taking usually requires mutual gaze contact, and simultaneously an appropriate boundary tone at syntactically appropriate pause. We study the relation of these signals in turn taking, and how they indicate if the speaker is holding a turn (speaking), giving a turn, taking a turn, or having no turn (listening). In this study, we paid attention especially to turn-give and turn-hold.

3. Conversational data

3.1. Data collection and annotation

We collected data from speakers participating in natural, free-flowing dialogues with no particular task to elicit particular behaviours. The collection setup is shown in Figure 1. Three participants sit in a triangle formation, and in front of one of them there is the eye-tracker to record the person’s eye movements (ES, or the eye-tracked speaker, is the leftmost partner in Figure 1). The two other participants, the left-hand speaker (LS) and the right-hand speaker (RS), are videotaped with a digital camera and they provide the reference point to what ES sees and where his gaze is focused on. The eye-gaze is tracked by the NAC EMR-AT VOXER eye-tracker. To avoid accommodation, ES was always a different person in each conversation, and the others rotated so that no group had exactly the same participants. Data collection and annotation are discussed in more detail in [6].

For the experiment, we chose 4-6 minutes long excerpts from four different dialogues. Three annotators annotated the behaviour of the three dialogue participants considering dialogue acts, gaze, face, head, turn-taking, feedback, and emotion/attitude. The dialogue act annotation followed the guidelines of the AMI corpus (www.amiproject.org), while non-verbal communication was annotated according to a modified MUMIN scheme [1]. Annotation was done with the Anvil annotation tool [8], and the inter-annotator agreement was measured by Cohen’s kappa-coefficient. We reached the kappa value 0.46, which corresponds to moderate agreement.

The annotation features are listed in Table 1. The eye-gaze of LS and RS is manually annotated on the basis of the video data, while the eye-tracked person’s gaze was based on the eye-tracker analysis. The eye-tracker data is not always clear as the gaze trace sometimes breaks. If the break (= no gaze) is shorter than 0.2 seconds, the gaze elements were regarded as part of the same gaze event (unless there was a shift), otherwise they were considered separate events.

3.2. Speech signal segmentation

We analysed the speech of the six speakers in the annotated 4 conversations (4 different ES + 2 LS/RS who all rotated in the triads). Speech segmentation was done with WaveCutter, and Figure 2 shows how the speech signal was segmented.

The conversation data were automatically divided into utterance units, based on signal energy. An utterance unit is defined as a speech spurt such that its energy level is higher than a given threshold, and it is separated by silence or by a signal with the energy less than the threshold, lasting for 200 ms or more. As a result, we obtained 110 utterances which were used for the analysis of turn management.

We then used WaveSurfer to extract F0 and power (intensity) from the 110 utterances. An example analysis is shown in Figure 3. As the speech features for each utterance, we used the maximum, minimum, and mean values of the extracted F0 and power, as well as the range which is the difference between the respective minimum and the maximum values of F0 and power (altogether 8 features). Since the loudness and intonation are different depending on the speaker (loud vs. quiet voice, high vs. low pitch) we distinguished the speakers in order to normalize their speech features. However, for the classification experiments, the data was combined since we were interested in the relation of speech and turn taking in general, rather than in the characteristics of each individual.

Speech features were normalized with respect to each speaker using Equation (1).

\[ Z_{ij} = \frac{X_{ij} - \mu_{ij}}{\sigma_{ij}} \]
In this equation, \( x \) is the maximum, minimum, average, or range value of F0 or power for every utterance by speaker \( i \), \( \mu \) is the mean value of the maximum, minimum, average, or range values of F0 or power for speaker \( i \), \( \sigma \) is the standard deviation of the maximum, minimum, average or range values of F0 or power for speaker \( i \), and \( z \) is a normalized value of the maximum, minimum, average, or range values of F0 or power. Each feature is normalised to a normal distribution with mean = 0 and variance = 1 by the processing.

For each utterance, also the features concerning the eye-gaze which synchronizes with the utterance as well as the dialogue act which corresponds to utterance were extracted besides the speech features. These manually annotated features were vectorized by assigning integers to the nominal values of the features, and generating a vector consisting of these numbers as elements. Analogously to the speech data of the individual speakers, also the gaze data of the individuals ES, LS, and RS was combined, i.e. individual gazing behaviour was not distinguished.

Finally, the speech features were integrated with the corresponding dialogue act and eye-gaze features into a single vector representation. The speech features are used as such as elements in a vector, and thus the result is a 12-dimensional integrated vector, which consists of one dialogue act feature, three eye-gaze features, and eight speech features.

4. Results and discussion

Our experiments deal with the classification of the utterance vectors with respect to turn-taking events on the basis of the speech signal and its manually annotated dialogue act and eye-gaze features. This gives us an indication of the use of the features in turn management: how well the extracted features signal a change of the speaker or a continuation of the same speaker.

We used the Support Vector Machine (SVM) technique in the Weka software [17] to classify turn-taking events. We used the polynomial kernel function, and the hyper-parameters of SVM were set experimentally. Classification was conducted through 10-fold cross-validation.

The results obtained by the different feature sets are given in Table 2. “DialogueAct only” is the result when only the dialogue act feature was used, i.e. classification is based on the interpretation of the content and intention of the utterance. We regard this as the baseline. “DialogueAct + Speech” is the result when speech features were added to the dialogue act feature, while “DialogueAct + Eye-gaze” is the result when eye-gaze features were added to the dialogue act feature. Finally, “DialogueAct + Speech + Eye-gaze” is the result when all the features were used.

The eye-gaze information seems to improve classification accuracy but this is not the case with the speech features: they do not improve baseline accuracy although this was expected in particular in regard to dialogue acts (cf. [14]). The reason may be that the intention of the utterance becomes clear when the dialogue act has been recognized, and the selected speech features do not affect turn-taking classification: dialogue acts already encode much of the speaker’s intentions concerning turn management (questions and suggestions presuppose answers from the partner, while backchannels and evaluative statements usually let the current speaker to continue). The speech features obviously become effective when the intention of utterance is vague or undetermined, cf. [14].

<table>
<thead>
<tr>
<th>Feature</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>DialogueAct only</td>
<td>0.669</td>
<td>0.655</td>
<td>0.581</td>
</tr>
<tr>
<td>DialogueAct + Speech</td>
<td>0.669</td>
<td>0.655</td>
<td>0.581</td>
</tr>
<tr>
<td>DialogueAct + Eye-gaze</td>
<td>0.752</td>
<td>0.755</td>
<td>0.753</td>
</tr>
<tr>
<td>DialogueAct + Speech + Eye-gaze</td>
<td>0.752</td>
<td>0.755</td>
<td>0.753</td>
</tr>
</tbody>
</table>

Table 2: Classification accuracy for each feature condition.

If the dialogue act information is combined with the eye-gaze information (the last two lines), classification accuracy is significantly improved. The last line of Table 2 shows that if the speech features are used as well, accuracy is equivalent to the case of using dialogue act and eye-gaze features only.

Considering the fact that turn management is joint activity between the speaker and the listeners, the result is understandable: while the speech features are able to express information of one single speaker only, the eye-gaze features use the information of both the speaker and listeners, i.e. they encode mutual information of turn-management. Therefore, the results allow us to conclude that eye-gaze is important in turn-management, and effective in distinguishing the turn-give and turn-hold.

The confusion matrices in the case of DialogueAct only, DialogueAct + Speech, DialogueAct + Eye-gaze, and DialogueAct + Speech + Eye-gaze are given in Tables 3, 4, 5, respectively. As can be seen, there are many confusions concerning turn-hold when only the dialogue act feature is used (Tables 3), and the same is true when the speech features are added in the vector (Table 4). In other words, when the interlocutor is speaking (holding the turn), it is difficult to distinguish whether there is a turn-taking possibility and the speaker would like to yield the turn, or if this is just a pause and the speaker intends to continue speaking.

Classification accuracy of turn-hold is greatly improved by combining the dialogue act feature with the eye-gaze features (Table 5), i.e. eye-gaze behaviour seems to provide information about the speaker’s focus of attention, and thus the speaker’s willingness to yield the turn. This supports again the hypothesis that eye-gaze is an important signal in turn-management, as it provides an extra source of information for the partners to interpret the speaker’s intentions to give the turn. Table 6 shows that speech features do not improve on the classification. What is interesting to notice in these two cases is that gaze information also seems to increase confusion on turn-giving. While there was only a few cases when the partner missed turn-giving signals as turn-holding, there now seems to be a greater possibility for “false alarms”, i.e. the partner interprets the signals as turn-hold although they actually are turn-give.

<table>
<thead>
<tr>
<th></th>
<th>Feature</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>DialogueAct only</td>
<td>x y</td>
<td>classified as</td>
<td>x=turn-give</td>
<td>y=turn-hold</td>
</tr>
<tr>
<td>Total</td>
<td>68</td>
<td>3</td>
<td>x=turn-give</td>
<td></td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>7</td>
<td>y=turn-hold</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Confusion matrix with DialogueAct only.

<table>
<thead>
<tr>
<th></th>
<th>Feature</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x y</td>
<td>classified as</td>
<td>x=turn-give</td>
<td>y=turn-hold</td>
</tr>
<tr>
<td>Total</td>
<td>68</td>
<td>3</td>
<td>x=turn-give</td>
<td></td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>6</td>
<td>y=turn-hold</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Confusion matrix with DialogueAct + Speech.
5. Conclusions

Gaze has many functions in face-to-face interaction. In this paper we experimented with turn-taking signals such as the speech features and information about the gaze, and investigated their effect on turn management. In the experiments, we focussed on the classification of turn-give and turn-hold because it is important to predict whether the interlocutor has the right to speak, i.e. if the signalling shows that the turn is passed to the partner in the turn-management.

The result confirms the earlier findings [7] that mutual gaze is needed for successful turn change. Eye-gaze gives the interlocutors guidance as to when to take the turn, or if the partner wants to continue talking or is preparing to talk, and thus helps in smooth and effective interaction without explicit spoken expressions.

There are further challenges that concern the correlations between eye-gaze and turn-management and we will investigate these further. On the basis of the gaze elements and the features that were annotated in our data, we noticed that turn-taking could be predicted fairly accurately. However, aspects such as how long and how often the partner is looked at, are also relevant from the communicative point of view, as well as the whole eye region (eyebrows, eyelids, wrinkles), and the upper body and gesture movements. Previously we noticed that in multiparty dialogues also head turns play an important role in signalling turn-taking [5]. It would thus be useful to combine all feature sets and look for correlations among the various non-verbal signals.

6. Acknowledgements

We would like to thank all the students who helped us with the data collection and annotation. The first author would also like to thank NiCT for the grant that allowed her to conduct research at Doshisha University in Japan.

7. References