Multimodal End-of-Turn Prediction in Multi-Party Meetings

Iwan de Kok
Human Media Interaction
University of Twente
P.O. Box 217, 7500AE
Enschede, The Netherlands
i.a.dekok@ewi.utwente.nl

Dirk Heylen
Human Media Interaction
University of Twente
P.O. Box 217, 7500AE
Enschede, The Netherlands
d.k.j.heylen@ewi.utwente.nl

ABSTRACT

One of many skills required to engage properly in a conversation is to know the appropiate use of the rules of engagement. In order to engage properly in a conversation, a virtual human or robot should, for instance, be able to know when it is being addressed or when the speaker is about to hand over the turn. The paper presents a multimodal approach to end-of-speaker-turn prediction using sequential probabilistic models (Conditional Random Fields) to learn a model from observations of real-life multi-party meetings. Although the results are not as good as expected, we provide insight into which modalities are important when taking a multimodal approach to the problem based on literature and our own results.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Discourse*; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Intelligent agents*

General Terms

Performance, Theory

Keywords

Multimodal, End-of-Turn Prediction, Machine Learning

1. INTRODUCTION

With recent advancements in human-like robotics [26, 19] and virtual humans the age of active androids in our society draws nearer. Before this age is here, a lot of research will be needed to be able to generate human-like conversational behavior that is as fluent as real-life conversations between humans. Over the years many dialogue systems have been developed. In the eighties and nineties of the previous century most of these relied on speech or keyboard input and produced either speech or text output only. The TRAINS dialogue system can be considered a classic example [28]. Since the start of this century the study and development of embodied conversational agents and humanoid robotics has lead to consider also nonverbal means of communication in tandem with speech and natural language [6] (see also the proceedings of the Intelligent Virtual Agents conference).

Engaging properly in a conversation requires many skills. On the input side, it involves skills such as being able to process the speech, understand what is being said and inferring what is intended. In making a contribution to a conversation one not only needs to be able to formulate and articulate an utterance properly, but also needs to know the proper rules of engagement, i.e. knowledge of when it is appropriate or desired to say something. For this a robot or virtual human should know when it is being addressed or when the speaker is about to hand over the turn. This latter point is what we will address in this paper.

Most current dialogue systems are purely reactive when it comes to deciding when it is their turn to speak. They simply (or mainly) rely on the detection of a significant pause in the speech of the speaker to determine when to start a contribution. In real life conversations, people may anticipate the ending of a turn from what is being said or from paraverbal and nonverbal cues. Thus they may start as soon as the turn ends or even before the end of a turn with the next contribution. This idea of anticipation can already be found in Sacks et al. [25] account of turn-taking. In this paper the authors claim that turn-taking decisions must be based to some extent on prediction of end-of-speaker-turns that does not depend on pauses because pauses between turns are sometimes shorter than pauses within turns. Considering this, reactive systems are prone to make errors using a pause threshold as 'the' cue for turn-taking. Instead, a predictive model needs to be developed in order to improve the turn-taking behavior of dialogue systems.

In this paper we present a multimodal probabilistic approach to the prediction of end-of-speaker-turns in multi-party meetings. In Section 2 we will discuss the state of the art in the field of end-of-speaker-turn prediction and how our approach differs from previous research. Our approach is explained in more detail in Section 3. The methodology we used for our experiments is reported in Section 4 and the results of these experiments are discussed in Section 5. We conclude this paper with suggestions for future work in Section 6.

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

2. RELATED WORK

We are not the first to tread on the path of end-of-speaker-turn prediction. Schlangen [27] tried to predict end-of-speaker-turns using a machine learning approach. In his work he used prosodic and syntactic features to learn a model which classified words to be the 'last word' in a turn or not and compared the results of his system to humans performing the same task. He made comparisons between approaches which use pause thresholds and shows that the problem increases in complexity when no pause information is available, thus shifting the problem from detection to prediction. He reaches an F_1 score of 0.355 in the condition that no pause information is used.

Atterer et al. [2] expand on the work of Schlangen by introducing more syntactic features to the feature set. The study also focuses more on end of utterance prediction instead of end-of-speaker-turn prediction. Although they call their research *detection* their approach could as well be called *prediction*, since no information of pause at the end of an utterance is used, only features that are available from the turn itself. They compare results between the prediction of end of utterance words for utterances which are turn internal and utterances which are turn final. Furthermore the authors compare results between different feature sets. They reach an F_1 score of 0.564 with a feature set using word/POS n-gram features.

When detection is considered instead of prediction, higher F_1 scores are achieved. Fuentes et al. [11] report an F_1 score of 0.841. The authors utilize left-right-filters for the generation of their features. Since information from the future is used in their approach, it is not suitable for real-time dialogue systems. Fung et al. [12] report an F_1 score of 0.679 for English spoken conversations. They report that pause is their best feature, a feature which we do not intent to use, since we are interested in predicting end-of-speaker-turns.

Note that all these studies look at two person face to face conversations, whereas we will look at multiparty interaction.

To the best of our knowledge no computational approach up until now, whether it entails prediction or detection of end-of-speaker-turns, have combined prosodic and syntactical features with visual cues. This is remarkable since the literature has proposed several visual cues that play a role in turn-taking.

Barkhuysen et al. [3] performed a study in which they compare the performance of humans in the task of end of utterance detection by providing them fragments of utterance varying in length from 1 to 2 words. These fragments were played to the participants with auditory cues only, visual cues only or both auditory and visual cues. They found that humans performed best when having both modalities available to them. Although differences are small, humans performed better with only visual cues than with only auditory cues. This supports our feeling that incorporating visual cues to end-of-turn prediction is key to mastering the task.

One of the visiual cues relevant for turn-taking is gaze behavior. Duncan [9], Kendon [17], Argyle and Cook [1], amongst others have studied the relation between gaze and turn-taking. According to this research speakers tend to look away from the listener as a turn begins and towards the listener when the turn draws to a close. In the study by Cassell et al. [7] it appeared that information-structure is an important factor in this behavior. Goodwin [13] also discussed several patterns of gaze related to the start of turns. For instance, at the start of a turn, a speaker may gaze to the addressees to check whether they are paying attention and pause or restart when this is not the case.

The shift in focus of attention signaled by gaze is often accompanied by a head shift [10]. It has also been observed that speakers often make tiny nods or shakes at the end of questions, eliciting confirmation from the addressees [15, 14, 24]. Barkhuysen et al. [3] found in their fragments that turn-final fragments included more head nods than non turn-final fragments.

Given these observations we decided to consider a multimodal approach including visual cues.

3. APPROACH

The goal of our research is to predict the end of a speaker turn based on multimodal features. The model we use to predict the end of a speaker turn is introduced by Morency et al. [21, 20]. Based on multimodal features this model is able to learn human behavior from recordings of real life conversations between humans through means of machine learning. The approach uses sequential probabilistic models such as Conditional Random Field [18] and Hidden Markov Models [23] to learn the relations between the observations from the real life conversations and the desired behavior.

In this approach the observations are represented by features which are sampled at a frame rate of 30Hz. From these features a model is inferred. The learned model returns a sequence of probabilities. The probabilities returned by the model are smoothed over time. The model can be used for generation by identifying peaks in this probability curve. These peaks represent good opportunities to display the learned behavior. The height of the peak can be used as the predicted probability of this opportunity. This probability can be used to adjust the expressiveness of the model.

This model has proven to improve the state of the art in backchannel prediction and generation [21] as well as backchannel recognition [20]. Since the comparable nature of the task at hand we deem this approach suitable to predict the end of a speaker turn.

4. EXPERIMENTS

In this section the experiments we conducted are discussed. In Section 4.1 the data set we used for our experiments is described. What we define as features and which features we derived from the data set is explained in Section 4.2. The way we use these features is explained in Section 4.3 and finally the methodology used for the machine learning is presented in Section 4.4.

4.1 Data

For the experiments we use the AMI Meeting Corpus [16]. This is a multimodal data set consisting of 100 hours of meeting recordings. From this data set we use the 14 meetings[1] for which dialogue act, focus of attention and head gesture annotations are available. We use the close-talking microphone audio recordings for the audio based features.

[1]These meetings are: ES2008a, IS1000a, IS1001a, IS1001b, IS1001c, IS1003b, IS1003d, IS1006b, IS1006d, IS1008a, IS1008b, IS1008c, IS1008d and TS3005a

Each meeting has 4 participants. Since we only use the features of the individual speaker for end-of-turn prediction at this point we can regard each participant in a meeting as a separate session. This gives us 56 ($= 4 \times 14$) sequences totalling approximately 24 hours of usable data.

4.2 Feature Extraction

We define a feature as a series of events. An event is defined as the time window (defined by a start and end time) in which the criteria describing the feature are met. Using this structure we can capture the different modalities in a uniform format. This makes the addition of extra modalities to our model an easy process.

From the data set we extract a total of 40 features which are divided into 4 modalities. In Table 1 an overview of the features is presented.

The first modality of features concerns the dialogue act annotations. We extracted the features in this modality using the annotations available in the AMI Meeting Corpus. Annotators have divided the meetings into segments and labelled each segment with the dialogue act according to the coding scheme. The coding scheme identifies 15 different dialogue acts [16], namely *Inform*, *Offer*, *Suggest*, *Assess*, *Comment*, *Elicit Inform*, *Elicit Offer*, *Elicit Assess*, *Elicit Comment*, *Backchannel*, *Stall*, *Fragment*, *Be Positive*, *Be Negative* and *Other*. We use 14 of these dialogue acts as features for our model. The only dialogue act we do not use in our experiments is *Other*. This dialogue act is not useful in the prediction model of end of turn. Even if it proves to be useful, there is no way to use it in a real life system since it is basically the garbage bin of the coding scheme.

The second modality of features is focus of attention. The features in this modality are also extracted using the annotations available in the AMI Meeting Corpus. In the coding scheme seven relevant places are identified and the eye gaze of the participants are labelled according to them [16]. These seven places are *Participant A*, *Participant B*, *Participant C*, *Participant D*, *Table*, *Slidescreen*, *Whiteboard*. We use each of these seven places as a feature. We add an eighth feature which we calculate by combining the features Participant A, B, C and D. We argue that it is not relevant whether the speaker is looking at participant A or B, but the fact that he/she is looking at *a* person. By combining the four features of the individual persons we capture this information.

The third modality of features are head gestures for which we also use the annotations from the AMI Meeting Corpus. These annotations reflect the intention of the produced head gesture rather than form of the gesture. In the coding scheme a distinction is made between the communicative head events and the remaining head events [16]. We only use the six communicative head events. These events are *Concord*, *Discord*, *Negative*, *Turn*, *Deixis* and *Emphasis*. Note that the turn event is a head gesture from a listener to take the speaking turn rather than from the speaker to keep the speaking turn.

The final modality are the prosodic features we extracted from automated computations of pitch and intensity of the audio signal. For this computation we used PRAAT version 5.1.03 [4]. We extracted the signals at 10ms intervals. From these raw signals we extracted 12 features, 6 based on the pitch signal and 6 based on the intensity signal. The features we extract from these signals are partially based on the work of Ward et al. [31].

After calculating the raw pitch signal at intervals of 10ms and with a lower and upper boundary for 50Hz and 500Hz respectively, we cleaned up the signal as PRAAT calculated pitch at times at which no speech was present, but only breathing into the microphone. The pitch values of those moments were between 50 and 150 Hz. Therefore we decided to filter out all pitch values below 150 Hz.

The first feature we derived from the pitch signal is *Pitched*. An event in this feature starts when the signal becomes larger than 0 and ends when it becomes 0 again. The next features are *Rising Pitch* and *Falling Pitch*. These events start when the pitch rises or drops for at least 100ms and end when they stop rising or falling. The next two features are *Fast Rising Pitch* and *Fast Falling Pitch*. These events start when the pitch rises or drops at least 10Hz every 10ms for at least 50ms. Another pitch feature is *Low Pitch*. These events start when the pitch drops below a certain threshold and stays there for at least 100ms. The threshold is determined by the taking the mean of the lowest encountered pitch and average pitch of the speaker. The final pitch feature is *High Pitch*. The threshold of this feature is obtained by taking the mean of the average pitch and the highest encountered pitch. Events started after the pitch is above this threshold for at least 100ms.

The first feature we derived from the intensity signal is *Intensity Above Noise*. An event of this feature starts when the intensity level rises above the standard noise level. The threshold of the noise level is determined by calculating the histogram of the intensity signal and smoothing it. We look up the first local minimum and use this as our threshold. The events in the feature *Long Intensity* start when the intensity rises above noise level for at least 700 ms. Just as with the pitch feature we also have *Rising Intensity* and *Falling Intensity* which record moments of at least 50ms of rising or falling intensity. The features *Fast Rising Intensity* and *Fast Falling Intensity* records events of at least 30ms of rising or falling intensity by at least 5 dB per frame.

The class feature we use for learning is based on the dialogue act annotations. O'Connell et al. [22] define a turn as follows: "all of the speech of one participant until the other participant begins to speak". This formed a first condition for our definition of turn. But someone may interrupt the speaker while the turn was not yet finished. We therefore also required that the speaker had finished his dialogue act. Backchannels or other short utterances are also included in this definition. However as we are more interested in predicting the end of a longer utterance we only looked at turns longer than 5 seconds. Since we are interested in predicting the end of a turn we label the last second of a turn as our class feature. Our data set contains a total of 1327 class labels.

4.3 Feature Representation

We represent the events of each feature in more than one way to enrich the information provided by each feature. This is done by using the encoding dictionary introduced by Morency et al. [21, 20]. By representing the events in various ways we simulate the different kind of relationships the feature may have with the to be predicted behavior. We use three different encoding templates for our features and 9 different encodings in total. These encoding templates and the number of encodings we use of each template are:

Modality	Source	Features
Dialogue Acts	Annotation	Inform, Offer, Suggest, Asses, Comment, Elicit Inform, Elicit Offer, Elicit Assess, Elicit Comment, Backchannel, Stall, Fragment, Be Positive, Be Negative
Focus of Attention	Annotation	Person, Participant A, Participant B, Participant C, Participant D, Table, Slidescreen, Whiteboard
Head Gestures	Annotation	Concord, Discord, Negative, Turn, Deixis, Emphasis
Prosody	Automatic Generation	Pitched, Rising Pitch, Falling Pitch, Fast Rising Pitch, Fast Falling Pitch, Low Pitch, Intensity Above Noise, Long Intensity, Rising Intensity, Falling Intensity, Fast Rising Intensity, Fast Falling Intensity, High Pitch

Table 1: Overview of the 40 features we used in our research. The features are divided into four modalities. In the second column the source of the features is presented.

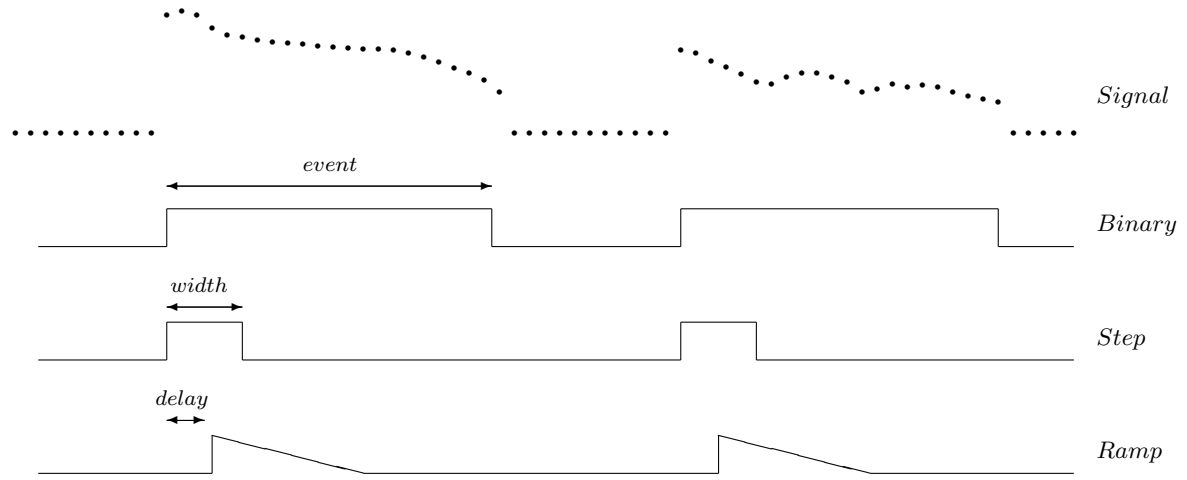

Figure 1: Explanation of the encoding dictionary. The top graph represents the raw signal of a feature, in this case the pitch signal. For instance if we create a feature which represents the areas where the pitch signal is above 0, we record all the events where this happens. With *binary* encoding the resulting feature looks like the second graph. A *step* encoded version of this feature with a delay of 0 and a width of 1 looks like the third graph. A *ramp* encoded version of the feature with a delay of 0.5 and a width of 2 looks like the bottom graph.

- Binary (1 encoding)
- Step (9 encodings)
- Ramp (9 encodings)

The first encoding template we use is *Binary*. This is the most straightforward encoding in which the signal of the encoded feature is 1 between the start and end time of the event and 0 in between events. The result of this encoding is shown as the second graph in Figure 1. This encoding is useful when the fact whether or not this event is happening in itself is a condition for behavior. For instance it may be the case that it is unlikely that someone stops talking when he is not looking at the listeners [7].

The *Step* encoding template models events in a way which captures timing since the start of an event and a limited time window. Two parameters are introduced in this encoding representing the delay after the first occurrence of the event and the width of the time window. The signal of the encoded feature is 1 after the delay since the start time has passed and will remain 1 for the width of the time window. This encoding is useful if for instance only the first second since the start of the event is relevant (delay 0, width 1). An example encoding with a delay of 0 and a width of 1 is shown as the third graph in Figure 1. In our experiments we use 9 variations of this encoding, namely (delay, width): (0,0) (0,1) (0,2) (1,0) (1,1) (1,2) (2,0) (2,1) (2,2).

The third encoding template we use is *Ramp*. We use the same two parameters as in the Step encoding template. Instead of remaining 1 for the width of the time window, the signal of the encoded feature will now decrease linearly from 1 to 0. The influence of a feature may change over time and this template captures this. An example encoding with a delay of 0.5 and a width of 2 is shown as the bottom graph in Figure 1. In our experiments we use 9 variations of this encoding, namely (delay, width): (0,0) (0,1) (0,2) (1,0) (1,1) (1,2) (2,0) (2,1) (2,2).

Features Set	F_1	Precision	Recall
Head Gestures	0.090	0.070	0.149
Dialogue Acts	0.063	0.061	0.081
Prosody	0.047	0.025	0.339
Focus of Attention	0.032	0.020	0.101
Multimodal all	0.061	0.041	0.225

Table 2: In this table the performance of our experiments are presented. In the multimodal feature set we use all the features of every modality.

Note that all these encoding templates can be generated using only the start time and are therefore suitable for real time generation.

4.4 Methodology

To train our prediction model we split the 56 sequences into 3 sets, a training set, a validation set and a test set. This is done by doing a 4-fold testing approach. 14 sessions are left out for test purposes only and the other 42 are used for training and validation. This process is repeated 4 times in order to be able to test our model on each session. Validation is done by using the hold out cross-validation strategy. A subset of 14 sessions is left out of the training set. This process is repeated 4 times and then the best setting for our model is selected based on the performance of our model.

Our data set contains an unbalanced number of end-of-spea-ker-turn frames compared to background frames. To balance our training set and to reduce training time without losing valuable information we preprocess our training set. From the complete sequences of the training set we randomly selected the same number of samples that contain an example of an end-of-speaker-turn as samples containing only background frames. The samples containing an example of an end-of-speaker-turn contained a buffer before and after the example with background frames. The size of the buffer randomly varied between 3 and 60 frames. The background samples ranged in size from 30 to 50 frames.

The machine learning technique used in all experiments is Conditional Random Fields (CRF). This technique has proven to out perform Hidden Markov Models in a comparable task [21]. The regularization term for the CRF model was validated with values $10^k, k = -1..3$.

The performance of our model is measured by using the F-measure. This is the weighted harmonic mean of precision and recall. Precision is the probability that predicted turn ends correspond to actual end-of-speaker-turns in our data. Recall is the probability that an end of a speaker turn in our test set was predicted by the model. We use the same weight for both precision and recall, so called F_1. During testing we identify all the peaks in our probabilities. A turn end is predicted correctly if a peak in our probabilities (see Section 3) occurs during an actual end-of-speaker-turn.

5. RESULTS AND DISCUSSION

We designed our experiment to investigate the influence of different modalities on the prediction of end-of-speaker-turn. Therefore we learn different models for each modality individually. Furthermore we learned a model which combines the different modalities. The performance of these models is presented in Table 2.

Unfortunately the results are not as good as expected. With only an F_1 score of 0.090 as our best result the performance is a lot lower than 0.355 Schlangen [27] reports and the 0.564 Atterer et al. [2] report. Looking for an explanation we will discuss for each modality literature suggesting how this modality can contribute to end of turn prediction in more detail and reflect why it did not work in our experiment.

The best modality in our experiments to predict end of speaker turn is the head gesture modality with an F_1 score of 0.090. Several sources mention head gestures as cues for turn-taking behavior. The averting gaze of the speaker at the beginning of a turn is usually accompanied by turning their head in the same direction [14, 15]. They turn their head towards the next speaker again upon completion of their turn. The direction of gaze and head orientation is especially an important cue in multi-party meeting to regulate floor management [30, 29].

Besides turning their heads to accompany an eye gaze aversion, Barkhuysen et al. [3] noted that during the final fragments of a turn speakers display more cases of head nodding than in non-final fragments. These head nods function as requests for feedback [13] and are responded to by listeners by a mimicked head nod (or other form of short backchannel) or more elaborately by taking over the turn.

Unfortunately the annotations available in the AMI Corpus do not include these function of head gestures in their annotation scheme. This could explain the poor performance of the model. The features do not describe the useful cues for this task.

The second best performance is the dialogue acts modality with an F_1 score of 0.061. These dialogue acts contain information about the structure of the conversation. This structure can be used to predict the end of a speaker turn. Especially the elicit category of dialogue acts can be used as a signal for the end of a speaker turn.

For instance take the following example. The speaker is explaining something through a series of *Inform* dialogue acts. He concludes his explanation with "Or don't you agree?". This is an *Elicit-Assessment* dialogue act which clearly signals the end of the explanation by the speaker and the moment the speaker is ready to hand over the turn.

Even though the model can detect that an elicit dialogue acts is happening, the model does not know when this dialogue act is finished. Other modalities should provide this information to be able to predict the end of a speaker turn.

The role of prosody in end of speaker turn prediction is very marginal. The performance of our features is low (0.047) with a high recall rate, but low precision. This suggests that prosody features do contain cues, but by itself these cues are not discriminative enough. In the approaches of Schlangen [27] and Atterer et al. [2] prosody was only marginally responsible for their performance as well. They mainly rely on syntactic features to reach their performance.

There are conflicting views on the importance of prosody for end of turn prediction. De Ruiter et al. [8] conducted an experiment to determine whether end of turn prediction by humans is based on lexicosyntactic cues or intonational contours. Subjects were presented fragments from conversations and were asked to push a button at the moment they thought the speaker would end his or her turn. The fragments were either the original fragment, the fragment with a flattened pitch or the fragment with indiscernible words, but

with the intonational contours intact. They found no difference in the performance of the task between the original fragment and the fragment with a flattened pitch. However performance dropped significantly when the subject could not hear the words of the conversation anymore.

On the other hand Barkhuysen et al. [3] found that in fragments at the end of a turn seem to either have a higher or lower boundary tone, while fragments in the middle of a turn usually have a boundary tone around the mean of the speaker.

The final modality in our experiment is focus of attention or eye gaze. It is reported by several studies that speakers tend to look away when starting to speak and look back to the listener when they conclude their turn [1, 3, 10, 17]. Since the annotations currently available in the AMI corpus described the object the person is looking at, this information is only implicitly available in the object transitions. Results would probably be better when the annotations described the moments the speaker looked away and back, which are according to the literature the cues relevant for end of turn prediction, than the object the speaker is looking at.

The multimodal model actually performs worse than the head gestures modality or the dialogue acts modality by itself. This can be attributed to the fact that this probabilistic approach works better when the feature set is limited. Only with large amounts of data the model will be able to discriminate the features (and encodings) with relevant information from the noise features. A cumulative feature selection approach as employed by Morency et al. [21] is able to make this discrimation on smaller sets of data, but we were unable to employ this technique at the moment.

Besides the modalities covered in this approach literature suggests two other cues for end of speaker turn prediction. In the comparison Barkhuysen et al. [3] made between fragments at the end of a turn and non turn-final fragments, they observed that in turn-final fragments more people blinked with their eyes.

Another potential cue for multimodal end of speaker turn prediction is body posture. Cassell et al. [5] designed an embodied conversation agent, based on an empirical study, that exhibits appropiate posture shifts during dialogues with human users. One of the rules in their design is a low probability of a posture shift at the end of an turn. If such a posture shift is generated at this point, it is an posture shift with a long duration, high energy and with the lower body part, while at the beginning of a turn the probability of a posture shift is higher, with shorter durations and not only limited to the lower body part.

6. CONCLUSION AND FUTURE WORK

Even though the performance of our probabilistic approach to end of speaker turn prediction was unsuccessful, this research has provided valuable insight into the various modalities that play their part in turn-taking behaviour. Based on these insights a new attempt at multimodal end of turn prediction will be conducted.

The annotations of the AMI corpus proved not to include the most valuable information for this task. New annotations should be collected either through manual coding, derivatives of the current annotations or most ideally through automatic generation using head, eye gaze and body posture trackers. This would have the advantage of giving insight into the performance of the approach in a real time system.

All studies on end of turn prediction and detection, including this one, only used features from the speaker itself, but interaction is a joint activity in which your behavior is co-determined by the actions of the other participants. If during a turn of a speaker a listener starts expressing turn-taking behavior, for instance by starting to lean forwards and using hand gestures, the probability the speaker will end his turn will increase. Therefore it will be interesting to use such features in future work as well.

7. ACKNOWLEDGEMENTS

We would like to thank Khiet Truong for her advice on the Praat software and the prosody features. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211486 (SEMAINE).

8. REFERENCES

[1] M. Argyle and M. Cook. *Gaze and mutual gaze*. Cambridge University Press, London, United Kingdom, 1976.

[2] M. Atterer, T. Baumann, and D. Schlangen. Towards incremental end-of-utterance detection in dialogue systems. In *Proceedings of International Conference on Computational Linguistics*, 2008.

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

[4] P. Boersma and V. van Heuven. Speak and unspeak with praat. *Glot International*, 5(9-10):341–347, November 2001.

[5] J. Cassell, Y. I. Nakano, T. W. Bickmore, C. L. Sidner, and C. Rich. Non-verbal cues for discourse structure. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 114–123, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[6] J. Cassell, J. Sullivan, S. Prevost, and E. F. Churchill. *Embodied Conversational Agents*. MIT Press, Cambridge Massachusetts, London England, 2000.

[7] J. Cassell, O. E. Torres, and S. Prevost. Turn taking vs. discourse structure: How best to model multimodal conversation. In *Machine Conversations*, pages 143–154. Kluwer, 1998.

[8] J. de Ruiter, H. Mitterer, and N. Enfield. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515 – 535, 2006.

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

[10] S. Duncan and G. Niederehe. On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10:234–247, 1974.

[11] O. Fuentes, D. Vera, and T. Solorio. A filter-based approach to detect end-of-utterances from prosody in dialog systems. In *HLT-NAACL (Short Papers)*, pages 45–48. The Association for Computational Linguistics, 2007.

[12] J. Fung, D. Hakkani-Tur, M. Magimai-Doss, E. Shriberg, S. Cuendet, and N. Mirghafori. Prosodic features and feature selection for multi-lingual sentence segmentation. In *Proceedings of Interspeech 2007*, pages 2585–2588, 2007.

[13] C. Goodwin. *Conversational Organization: interaction between speakers and hearers*. Academic Press, 1981.

[14] D. Heylen. Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, 3(3):241–267, 2006.

[15] D. Heylen. Listening heads. In I. Wachsmuth and G. Knoblich, editors, *Modeling Communication with robots and virtual humans*, volume 4930 of *Lecture Notes in Artificial Intelligence*, pages 241–259. Springer Verlag, Berlin, 2008.

[16] http://corpus.amiproject.org. The AMI Meeting Corpus, May 2009.

[17] A. Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[18] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labelling sequence data. In *ICML*, 2001.

[19] T. Minato, Y. Yoshikawa, T. Noda, S. Ikemoto, H. Ishiguro, and M. Asada. CB^2: A child robot with biomimetic body for cognitive developmental robotics. In *IROS 2008: Proceedings of the IEEE/RSJ 2008 International Conference on Intelligent RObots and Systems*, pages 193–200, 2008.

[20] L.-P. Morency, I. de Kok, and J. Gratch. Context-based recognition during human interactions: Automatic feature selection and encoding dictionary. In *ICMI '08: Proceedings of the 10th International Conference on Multimodal Interfaces*, pages 181–188, New York, NY, USA, 2008. ACM.

[21] L.-P. Morency, I. de Kok, and J. Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In *Intelligent Virtual Agents (IVA '08)*, pages 176–190, 2008.

[22] D. C. O'Connell, S. Kowal, and E. Kaltenbacher. Turn-taking: A critical analysis of the research tradition. *Journal of Psycholinguistic Research*, 19(6):345 – 373, 1990.

[23] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[24] R. J. Rienks, R. Poppe, and D. Heylen. Differences in head orientation behavior for speakers and listeners: an experiment in a virtual environment. *Transactions on Applied Perception*, 7(1):accepted for publication, 2010.

[25] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696 – 735, 1974.

[26] D. Sakamoto, T. Kanda, T. Ono, H. Ishiguro, and N. Hagita. Android as a telecommunication medium with a human-like presence. In *HRI '07: Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 193–200, New York, NY, USA, 2007. ACM.

[27] D. Schlangen. From reaction to prediction: Experiments with computational models of turn-taking. In *Proceedings of Interspeech 2006*, 2006.

[28] T. Sikorski and J. F. Allen. A task-based evaluation of the trains-95 dialogue system. In *ECAI '96: Workshop on Dialogue Processing in Spoken Language Systems*, pages 207–220, London, UK, 1997. Springer-Verlag.

[29] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proceedings of CHI'01*, pages 301 – 308. ACM, 2001.

[30] R. Vertegaal, G. van der Veer, and H. Vons. Effects of gaze on multiparty mediated communication. In *Proceedings of Graphics Interface*, pages 95 – 102, Montreal, Canada, 2000. Morgan Kaufmann Publishers.

[31] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 32(8):1177–1207, 2000.