Automatic speech/non-speech classification using gestures in dialogue

Simon Alexanderson, Jonas Beskow, David House
Department of Speech, Music and Hearing, KTH
Lindstedtsvägen 24, 10044 Stockholm, Sweden
simonal@kth.se, beskow@speech.kth.se, davidh@speech.kth.se

Abstract

This paper presents an experiment carried out to determine what aspects of motion are associated with speech and what aspects are associated with non-speech in spontaneous dyadic communication. Six dialogs were analysed, and results show that the successful prediction of speech activity from motion differs considerably depending on the characteristics of the dialogue. The classification accuracy ranged from 61% to 82% using a Naive Bayes classifier. The most salient features were found to be the angular speed of the head and speed of the dominant hand.

1. Introduction

Traditionally, the study of speech and gesture has been pursued using laboriously hand-annotated video corpora. Motion capture techniques make it possible to study motion during dialogue from a purely statistical perspective. While the studying of semantic functions of individual gestures will still require manual annotation, there are other aspects of gesturing and motion that may be examined at a high level by looking at the relation between different channels of information for example in order to see to what degree one may be predicted from the other.

The main purpose of this study is to investigate possible differences in motion between gestures produced during speech and gestures produced by dialogue partners when not speaking. By using motion capture recorded dialogues a fully automatic approach to classifying conversational dynamics enables a comparison between co-speech and non-speech gestures with the aim of isolating and defining critical properties of the co-speech gestures.

2. Method

For this investigation the Spontal corpus of Swedish dialogue provided a rich database for the statistical analysis of the co-speech motion. The database, containing more than 60 hours of unrestricted conversation in over 120 dialogues between pairs of speakers, is comprised of synchronized high-quality audio and video recordings (high definition) and motion capture for body and head movements for all recordings (Edlund et al., 2010).

The motion data consist of the 3D positions of the motion capture markers attached to each subject. The used marker set contains 12 markers placed on the upper body according to Figure 1.

Figure 1: Marker set with 12 marker per subject

Since the data does not contain enough information for calculating hand orientations, we only extracted positional features for the hands. The chest and the head were equipped with three markers each, and by assuming that these markers form a rigid body, we extracted features for all 6 degrees of freedom (DOF) for these body parts. The features were calculated for each frame either as instantaneous values or over a moving window centered on the frame. We experimented with a series of window sizes, and 3 seconds was found to generate the best results. Table 1 shows all collected features divided in the sub-groups 'hands', 'head' and 'other'. All features were calculated in global coordinates except for the hands, where we used the data transformed to a coordinate frame located at the center of the chest and oriented with the x-axis along the shoulders and the y-axis up. This was done to make the features more robust to pose shifts.

To determine the speech activity at each frame, a voice activity detection (VAD) algorithm (Laskowski, 2011) was applied to the audio recordings from the near-field microphones attached to the subjects. The process resulted in two sets of speech/non-speech segments for each dialogue, one for each subject.

3. Results

The feature extraction algorithms were applied on six dialogues from the corpus generating a total of 120 000 instances per dialogue. We then performed subject dependent classification experiments using the WEKA software package (Witten and Frank, 2005). For each subject, the binary class speech/non-speech was predicted using 10-folds cross validation. The classification was performed using a) the 28 hand features, b) the 11 head features, c) the 45 combined 'head', 'hand' and 'other' features for the subject concerned (Comb), and d) the combined 90 features for both subjects in the dialogue (Tot). Table 2 shows the response accuracy from classification with a Naive Bayes (NB) classifier. The classification accuracies ranged from 61% to 82% for the 12 subjects. In a majority of the cases classification using the head features gave a better result than the hand features. The table also shows the results of a non-informative classifier (ZeroR), used to determine a baseline for accuracy performance. The ZeroR classifier ignores the features and predicts the majority class for all instances.

For a fine grained analysis of the predictive power of the features we performed feature selection on the combined 'Comb' groups for each subject. For this task,

we used WEKA's InformationGain evaluator together with the built in Ranker. The merits of each feature, obtained from 10-folds cross validation, were averaged over the 12 subjects and sorted in descending order. The results show that features related to velocity are most salient followed by the standard deviation of the hand locations and the distance from the hands to the head.

Group	Description	Dim
Hands	Hand positions	6
	Hand velocities	6
	SD of positions*	6
	Mean hand speed*	2
	Max hand speed*	2
	Correlation of left and right hand trajectories*	3
	Correlation of left and right hand velocities*	3
Head	Orientation	3
	Angular velocity	3
	SD of orientation*	3
	Mean angular speed*	1
	Max angular speed*	1
Other	Mean distance between hand and head*	2
	Angular velocity of chest	3
	Mean angular speed of chest*	1

Table 1: Description of extracted features. Features marked with * are calculated over a window of 3 s.

Subject Id	ZR (%)	NB (%)			
		Hands	Head	Comb	Tot
S11	54.2	72.8	68.8	73.3	76.4
S12	56.5	68.0	74.3	73.4	81.9
S21	73.3	76.1	77.7	78.4	78.3
S22	50.3	56.8	58.1	64.6	63.0
S31	58.0	63.5	64.5	65.5	67.3
S32	61.7	68.2	70.0	68.8	69.7
S41	52.3	61.1	58.6	63.4	70.1
S42	64.5	66.2	66.8	66.8	62.7
S51	70.4	69.2	68.2	68.9	68.0
S52	55.4	55.7	57.7	61.3	60.5
S61	61.1	55.1	65.1	65.6	66.2
S62	58.4	52.3	61.1	54.7	68.2

Table 2: Accuracy of 10-fold cross validation for different feature groups (Sij denotes dialog i, subject j)

4. Discussion

One of the characteristics of multimodal conversational behavior in unrestricted spontaneous dialogue is the fact that the amount of motion and gestures of the speakers is subject to great individual variability. It is clear that some subjects exhibit an abundance of manual gestures, and for these subjects hand velocity is the single strongest predictor of speech activity. Head motion, on the other hand seems to be somewhat more stable as a predictor of speech/non-speech across different speakers. One reason for this could be the fact that the head seldom is perfectly still during speech production. In fact, inspecting the motion trajectories for head rotation in the corpus, different types of motion can be identified; subtle motion likely arising as a consequence of the process of articulation, and more distinct semiotic motion. The latter, which could be more clearly identified as distinct gestures, can occur during speaking as well as listening. For many of the speakers, head nods frequently accompany speech, and during listening head nods tend to occur in the form of back-channeling. Another head related feature is head turning in relation to the other interlocutor.

5. Conclusions

The bulk of previous research on the relation between speech and gesture has been carried out on corpora recorded under controlled conditions with specific tasks to be carried out by the subjects, with the goal of eliciting a large number of gestures of a certain kind, e.g. spatial description tasks or re-telling of movie scenes. In this study we analyze completely unrestricted spontaneous dialogue, which differs greatly in the amount and type of gestures exhibited. There are great challenges involved in the analysis of this type of material and we have chosen not to rely on hand labeled events but instead employed a fully automated approach based on statistics of speech and motion characteristics.

We believe that analyzing the predictive power of motion features related to speech/non-speech classification is a useful way of gaining insights into the dynamics of spontaneous face-to-face dialogue. Identifying the motion features that are most closely associated with speech is an important first step in order to further analyze the dynamic structure of face-to-face conversation. The next step in this line of research will be to automatically extract semiotic events from the motion data based on the selected features, with the goal of automatically being able to discriminate meaningful gestures from other types of motion that occur in the data.

Another potential application could be speaker diarization, i.e. the task of identifying who is speaking when, in multiparty dialog. Typically speaker diarization relies on (multi-channel) audio, while relatively little research has been done incorporating global body behaviour (Anguera Miro et al., 2012). Our study shows that the motion channel provides significant information that is likely to be useful in such systems.

Acknowledgements

The work reported here has been funded by the Bank of Sweden Tercentenary Foundation (P12-0634:1) and the Swedish Research Council (VR 2010-4646).

References

X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, & O. Vinyals. 2012. Speaker diarization: A review of recent research. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2), 356-370.

J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House. 2010. Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. *Proc. of LREC'10*. Valetta, Malta: 2992–2995.

K. Laskowski. 2011. *Predicting, detecting and explaining the occurrence of vocal activity in multi-party conversation.* Doctoral dissertation, Carnegie Mellon University.

I. H. Witten and E. Frank. 2005. Data Mining: *Practical machine learning tools and techniques.* Morgan Kaufmann Publishers, San Francisco.