Audio Features for the Classification of Engagement

Christy Elias, João P. Cabral, and Nick Campbell

School of Computer Science and Statistics, Trinity College Dublin,
College Green, Dublin 2, Ireland
{eliasc,cabralj,nick}@tcd.ie

Abstract. This paper analyses the features encapsulated in the speech signal to estimate the engagement of the interlocutors in a conversation. The features used for classifying engagement in a multiparty dialogue corpus (TableTalk corpus) are the prosodic parameter F0, glottal parameters correlated with voice quality (open quotient, return quotient, speed quotient), and Mel-frequency cepstral coefficients (MFCCs). Different combinations of these features were used in a random forest classifier and results show that the use of voice quality features improve the classification results.

Keywords: conversational engagement, voice quality analysis

1 Introduction

Engagement detection can be applied to improve the quality of interactions in dialogue systems or to help to make human intervention decisions in automated dialogue systems such as in call centres. Researchers have used different methods to detect engagement in the past. Acoustic, temporal and emotional information from telephone calls were used in the work by Yu et al. [1] in which emotional levels estimated from acoustic information from utterences and this information was used to predict engagement. The parameters used consisted of pitch, spectral energy and duration parameters. Gustafson and Neiberg [2] modelled engagement based on listener responses where change in syllabicity, pitch slope and loudness in non lexical response tokens in Swedish were used to detect engagement. These studies were based on dyadic telephone conversations whereas Bohus and Horvitz modelled engagement in dynamic environments where the participants enter, leave and interact in a very natural manner [3]. Gaticia-Perez [4] considered the degree of engagement displayed by a person to be an expression of internal state of "interest" of that person resulting from the attraction towards the interlocutor, interest in the theme of the conversation or the social rapport. The definition of engagement varied according to the contexts of these studies and in the context of this work an interlocutor is considered to be engaged if he/she is in overall involved in the conversation and is interacting with others.

Voice quality has been found to be associated with speaker's emotion, mood and attitude [7]. Charfuelan et al. [8] used voice quality to predict the social status of participants in scenario meetings and reported the use of "louder-than-average" voice quality for the most dominant speaker and "softer-than-average" in the case of the least dominant speaker. Prosodic parameters and MFCCs have been used in previous works on engagement. For example, Hsiao et al. [9] used acoustic patterns and turn-taking patterns to detect continuous social engagement in dyadic conversations. Gupta et al. [10] analysed engagement behaviour in children with vocal cues in non-verbal vocalizations and their results suggests that vocal cues can be used effectively to detect engagement. In this work, the classification of engagement is extended to other voice quality parameters (glottal parameters), in addition to the previously used F0 and MFCCs.

The remainder of this paper is organised as follows. Section 2 describes the TableTalk corpus and engagement annotations used in this study. The feature extraction and classification of engagement are explained in Section 3. The results of classification is analysed in section 4 and conclusions are presented in section 5.

2 TableTalk Corpus

2.1 Data

The data used in the study is part of the TableTalk[1] corpus , which was collected at ATR Research Labs in Japan for studies in social conversations. The corpus collected over three days contains free conversations among four people, with exception of day 2 which has five participants. The conversations were in English although only one of the participants was a native speaker (the others were Japanese, French and Finnish). The recording was performed using a 360 degree camera to captures the faces of the participants who sat around a table, and a centre mounted microphone was used to record the audio.

2.2 Engagement Annotation

Bonin et al. [5] conducted an experiment for annotation of the day one recording (35 minutes long) for individual and group engagement. The segments were marked discretely with "+" for engaged and "-" for not engaged. Five psychology students were recruited as annotators and were given no restriction on the length of each annotated segments. Interlocutors were considered engaged if they were interacting with others or actively listening using backchannels, gestures or mimicry. The group was considered to be engaged if three out of the four interlocutors were considered to be engaged.

The first five tiers shown in Figure 1 are the annotations of group engagement obtained from the annotations in [5]. The timespans were not predefined for the annotations so as to let the annotator rate without any restriction of time.

[1] http://sspnet.eu/2010/02/freetalk/

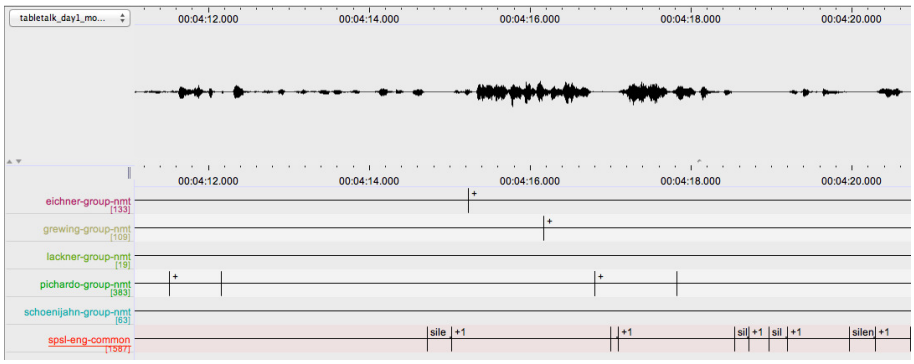

Fig. 1. Annotation of engagement in the TableTalk Corpus. The last tier shows the derived labels of engagement.

For this study new labels were derived from the original annotations by combining the labels of different annotators. The data was segmented into silence and speech. The new labels are shown in the last tier in Figure 1. In this case, each segment was given a numeric label of +1 for "engaged" or -1 for "not-engaged" based on the maximum number of annotations that were marked as engaged/not-engaged (respectively) in overlapping segments of the annotations.

3 Experiment

3.1 Feature Extraction

The speech features were estimated from the audio recording down-sampled to 16 kHz. The estimated features consisted of prosodic (F0), Mel-frequency cepstral coefficients (MFCCs) and glottal parameters. Features were estimated on frames 25 ms long and using a frame shift of 5 ms.

F0 and 12 MFCC coefficients (including log energy) were extracted using the SPTK toolkit (`http://sp-tk.sourceforge.net`). The glottal parameters open quotient (OQ), return quotient (RQ) and speed quotient (SQ) were estimated using the method described in [6]. These three parameters are strongly correlated with voice quality. Open quotient (OQ) is the ratio of the duration of the open phase of the glottal cycle (when the glottal folds are open) by the pitch period. Speed quotient (SQ) represents the asymmetry of the glottal pulse. The return quotient (RQ) is related with the abruptness of the transition between open phase and the closed phase which is proportional to the spectral tilt.

3.2 Engagement Classification

A random forest learning algorithm implemented in Weka [11] was used for the classification of engagement. The classification experiment was performed using

different combinations of the speech features together with the engagement annotations. Seven feature sets were used: F0, MFCC, VQ, F0+VQ, MFCC+F0, MFCC+VQ, MFCC+F0+VQ. A 10-fold cross validation approach was performed to assess the classifier.

4 Results

The results are shown in Table 1. The distribution of engaged and not-engaged classes were not balanced in the TableTalk corpus. For this reason, the average per-class accuracy (unweigthed accuracy) was calculated to assess the results of the classification.

The combination of MFCC, F0 and voice quality parameters resulted in the highest unweighted accuracy (88.24%) among the combinations used. The results were statistically significant with p-value$<$0.05. The F-measure for the engaged and not-engaged segments was higher when voice quality features were used. The increase in the F-measure shows that the classifier performed better when voice quality features were combined with F0 and MFCCs.

Table 1. Results of random forest learning for engagement classification

Features	Accuracy	F-Measure	
		Engaged	Not-Engaged
F0	65.39%	0.777	0.228
VQ	77.72%	0.861	0.439
MFCC	86.22%	0.916	0.621
F0+VQ	77.53%	0.862	0.398
MFCC+F0	87.55%	0.923	0.669
MFCC+VQ	87.77%	0.925	0.677
MFCC+F0+VQ	88.24%	0.927	0.693

5 Conclusions

In this study the use of audio features (F0, MFCC, Voice quality) for classification of engagement was analysed. It can be seen that the combination of voice quality features with F0 and MFCC features improved the classification of engaged and not-engaged segments, compared with the commonly used set of acoustic parameters (F0 and MFCC). The classification of engagement using more voice quality features on a larger data set will be the immediate future work.

Acknowledgments. This research is supported by the Science Foundation Ireland through the CNGL Programme (Grant 12/E/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin.

References

1. Yu, C., Aoki, P.M., Woodruff, A.: Detecting user engagement in everyday conversations. CoRR. vol. cs.SD/0410027 (2004).
2. Gustafson, J., Neiberg, D.: Prosodic cues to engagement in non-lexical response tokens in Swedish. In: DiSS-LPSS, pp. 63–66 (2010).
3. Bohus, D., Horvitz, E.: Models for Multiparty Engagement in Open-world Dialog. In: Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 225–234. Association for Computational Linguistics, Stroudsburg, PA, USA (2009).
4. Gatica-Perez, D.: Modeling interest in face-to-face conversations from multimodal nonverbal behavior. Multi-Modal Signal Processing: Methods and Techniques to Build Multimodal Interactive Systems, pp. 309–323 (2009).
5. Bonin, F., Bock, R., Campbell, N.: How Do We React to Context? Annotation of Individual and Group Engagement in a Video Corpus. In: 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on Social Computing (SocialCom), pp. 899–903 (2012).
6. Cabral, J.P., Renals, S., Richmond, K., Yamagishi, J.: Towards an improved modeling of the glottal source in statistical parametric speech synthesis. In: Proceedings of the 6th ISCA Workshop on Speech Synthesis, pp. 113–118 (2007).
7. Gobl, C., N Chasaide, A.: The role of voice quality in communicating emotion, mood and attitude. Speech communication. vol. 40, 189–212 (2003).
8. Charfuelan, M., Schrder, M., Steiner, I.: Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings. In: INTERSPEECH. pp. 2558–2561 (2010).
9. Hsiao, J.C., Jih, W., Hsu, J.Y.: Recognizing continuous social engagement level in dyadic conversation by using turntaking and speech emotion patterns. In: Activity Context Representation Workshop at AAAI (2012).
10. Gupta, R., Lee, C.-C., Bone, D., Rozga, A., Lee, S., Narayanan, S.: Acoustical analysis of engagement behavior in children. In: WOCCI. pp. 25–31 (2012).
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10–18 (2009).