Real-Time Feedback System for Monitoring and Facilitating Discussions

Sanat Sarda[1] , Martin Constable[2], Justin Dauwels[1*] , Shoko Dauwels (Okutsu)[3], Mohamed Elgendi[4], Zhou Mengyu[2]
Umer Rasheed[1], Yasir Tahir[4] , Daniel Thalmann[4], Nadia Magnenat-Thalmann[4]
[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]School of Art, Design, and Media, Nanyang Technological University, Singapore
[3] Centre of Innovation Research in Cultural Intelligence and Leadership (CIRCQL), Nanyang Business School, Singapore
[4]Institute for Media Innovation, Nanyang Technological University, Singapore
[*]Corresponding author
sanat.sarda@gmail.com, {mconstable, jdauwels, sdauwels, elgendi, danielthalmann, nadiathalmann}@ntu.edu.sg,
{zhou0138, umer1, yasir001}@e.ntu.edu.sg

Abstract–**In this paper we present a system that provides real-time feedback about an ongoing discussion. Various speech statistics, such as speaking length, speaker turns, and speaking turn duration, are computed and displayed in real-time. In social monitoring, such statistics have been used to interpret and deduce talking mannerisms of people, and gain insights on human social characteristics and behaviour. However, such analysis is usually conducted in an offline fashion, after the discussion has ended. In contrast, our system analyses the speakers and provides feedback to the speakers in real-time during the discussion, which is a novel approach with plenty of potential applications. The proposed system consists of portable, easy to use equipment for recording the conversations. A user friendly graphical user interface displays statistics about the on-going discussion. Customized individual feedback to participants during conversation can be provided. Such close-loop design may help individuals to contribute effectively in the group discussion, potentially leading to more productive and perhaps shorter meetings. Here we present preliminary results on two-people face to face discussion. In the longer term, our system may prove to be useful, e.g., for coaching purposes, and for facilitating business meetings.**

Keywords: Behaviour, social monitoring, graphical user interface, portable, real-time feedback.

I. Introduction

People have varying individual characteristics, personality, status, intelligence, maturity, language among others. All these aspects in different combinations result in individual speaking mannerisms, such as how much a person speaks during a conversation, or how much he or she interrupts another person while speaking [1]. Talking mannerisms of individuals play an important factor for meetings to be productive and achieve certain objectives. If talking mannerisms become mutually compatible or aligned, it the meetings are likely to be more productive and efficient [2]. Our long-term objective is to develop systems that provide real-time feedback about social behaviour in conversations, helping speakers to adjust their talking mannerisms to each other. Such systems may help to boost the effectiveness of job interviews, group discussions, coaching sessions, or public speaking.

In the fields of psychology and cognitive science, human behaviour is often studied from the perspective of social interactions [3-4]. Traditionally, expert observers take notes during conversations. Alternatively, audio and video recordings of conversations are analysed manually by experts. Both approaches are time-consuming and unavoidably subjective. Recent advances in recording equipment [5] and signal processing may ultimately enable automated and real-time analysis of talking mannerisms and social interactions at large, yielding more objective results. Several studies in that direction have been conducted in recent years, to deduce individual characteristics like dominance status [6-7], emerging leadership [8] and other personality related traits [9-10]. In such studies, various statistics of the conversation are extracted, e.g., natural turns, turn duration, speaking percentage, interruptions and failed interruptions. Also, the combination of speech and visual features has been shown to provide increased accuracy for detecting characteristics such as dominance and leadership [11].

However, in none of those studies [6-11], social interactions are analysed in *real-time*, instead various corpora of audio and video recordings are analysed offline [12]. Many corpora of audio and video signals are available related to small-group interactions (see [13] for a survey). These corpora are continuously updated with manual annotations, which can be used as gold standard to assess automated analysis methods. The systems presented in [13-14] have the capability to provide real-time feedback, however, those systems do not use speech or video signals, but rather crude signals. Consequently, subtle talking mannerisms may not be detectable through such approach.

In the present work, we propose a system that in real-time provides feedback about talking mannerisms, generated from speech and video signals. The system first extracts numerous speaking statistics from those signals, most of which are similar to features considered in recent studies [6-14]. Machine learning algorithms further process those statistics, to extract higher-level characterisation of the speaking mannerisms. That information is eventually exploited to generate real-time feedback for every participant in the meeting. It can inform the speakers about their speaking mannerisms, and if needed, provide guidelines.

In this paper, we limit ourselves to automatic analysis of conversations of two persons. Such scenario is relevant for coaching, interviews, and business meetings. In the future, we plan to scale our system towards small-group interactions.

This paper is structured as follows. In Section II, we elaborate on the speech statistics extracted by our system. In Section III, we explain our implementation, including the recording setup, voice activity detection (from audio and video signals), and our design of graphical user interface. In Section IV, we outline the proposed framework for offline and real-time feedback. In Section V, we present our conclusions and make suggestions for future work.

II. Non-Verbal Speech Cues

The core of our system consists of simple (and hence fast) signal processing algorithms that detect who is speaking and when, and use that information to compute various statistics about non-verbal speaking mannerisms. In this section, we elaborate on the latter statistics. In the next Section, we will explain how we determine who is speaking and when, from audio and video signals.

A. Non-Verbal Speaking Statistics

We compute various simple non-verbal speaking statistics (see Fig. 1). Each of those statistics can be computed in real-time. Specifically, the following non-verbal speaking statistics are considered.

Speaking Percentage: The percentage of time a person speaks in the conversation.

Voicing Rate: The number of syllables spoken per minute.

Pitch: Pitch of speech is calculated using the Voice-box Toolbox [15].

Natural Turn-Taking: The number of times person *'A'* speaks in the conversation without interrupting person *'B'*.

Silence: The percentage of time when both participants are silent.

Interruption: Person *'A'* interrupts person *'B'* while speaking, and takes over. Person *'B'* stops speaking before person *'A'* does (see Fig. 1).

Failed Interruption: Person *'A'* interrupts person *'B'* while speaking, but stops speaking before person *'B'* does (see Fig. 1).

Interjection: Short utterances such as 'no', 'ok', 'yeah', 'exactly' (see Fig. 1).

Speaker Turn Duration: Average duration of each speaker turn.

Overlap Percentage: Percentage of time when both persons speak at the same time during the conversation.

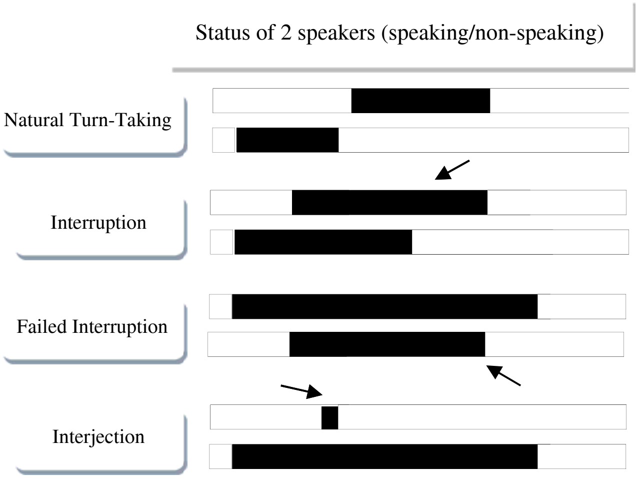


Figure 1.Illustration of turn-taking, interruption, failed interruption and interjection derived from binary speaking status (speaking and non-speaking). Periods of speaking and non-speaking are indicated in black and white respectively.

Each of the above statistics, individually and in combination, portrays different characteristics of people's behaviour during the discussion. For instance, high speaking percentage may indicate dominant behaviour of a person, whereas a high interruption count may indicate aggressive behaviour. In other words, we may be able to interpret various personality traits and dynamics of conversations from these basic statistical measures.

B. Interpretation of Non-Verbal Speaking Statistics

It may not always be straightforward to directly interpret the non-verbal speaking statistics. For instance, the number of interruptions during a debate or friendly conversation may be identical. However, the interpretation of those statistics may be quite different. In other words, the context plays an important role in interpreting the statistical measures. To address this issue, we suggest to record many conversations, in a large variety of settings. Next, from each of those recordings, one may compute the statistical measures. For each setting, histograms of the statistical measures can be generated, and based on those, one can set thresholds. The latter are then used to assess discussions. For instance, if the number of interruptions is above a threshold, the person is considered as highly interrupting; if that number is below another threshold, the speaking behaviour is considered as non-interrupting. The thresholds can be chosen as statistical quantiles. Alternatively, unsupervised learning (e.g., k-nearest neighbours clustering) may be applied to identify clusters in the statistical features across a large number of recordings. Such clusters may correspond to different speaking behaviours. By identifying to which cluster an ongoing discussion belongs, the system may be able to identify speaking behaviours in real-time.

III. Implementation

In this section, we describe our system for extracting non-verbal speaking statistics from ongoing discussions. First we explain the recording hardware, next we elaborate on speaker segmentation and our graphical user interface (GUI).

A. *Sensing and Recording*

In our setup, we use two table-top microphones, one per person (see Fig. 2), and a ZOOM H4n portable voice recorder which allows us to record from multiple speakers simultaneously. That voice recorder has flexible sampling rates and bit resolutions. We selected a sampling rate of 8kHz, as such low rate suffices for our purposes. The microphones are connected to the recorder via balanced XLR connectors. The recorder acts as an interface and is connected to the laptop via USB (see Fig.2).

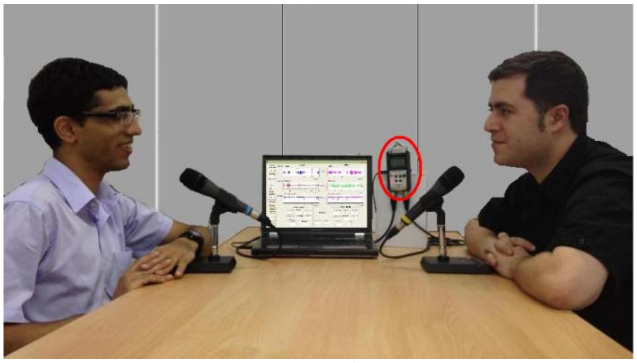


Figure 2.Recording system consisting of the Zoom H4n voice recorder (circled in red) and two Sennheiser e845s microphones, one for each participant in the conversation. The laptop can run the GUI (cf. Section III.C), which can be used to provide real-time non-verbal speech statistics to an external observer.

For online recording and real-time feedback, recordings are saved directly on the laptop. With the H4n recorder acting as interface, recordings from two separate table-top microphones are recorded synchronously without any delay. The setup with H4n recorder provides easy, quick, cost-effective, portable and most importantly, undistorted signal recording. We refer to [16] for an excellent review of computer-based audio recordings.

It is important to use suitable microphones and connectors to acquire undistorted original speech signals. Microphones are required to have a flat frequency response, in order to preserve the original speech energy and spectrum, optimally sensitive to allow talking from comfortable distance. Moreover, the microphones are recommended to be directional limit interfering signals and to reduce background noise. The connectors should be balanced to reject line noise interference. Overall, the recording system should not be imposing, in order not to disturb the speakers. For our present recording, we use Sennheiser e845s microphones with XLR connectors, as those components fulfil all the requirements. That solution is limited to two-people recordings. For small-group discussions with more than two people, we will use professional audio interfaces in the future that allow simultaneous multi-channel recording.

B. *Voice Activity Detection and Speaking Segmentation*

After the audio signals have been acquired and stored on the laptop, pre-processing is conducted, i.e., voice activity detection and segmentation of the speakers. Generally, the objective of voice activity detection is to differentiate between speech and non-speech (including silence and all kinds of noise and signals unrelated to speech). The purpose of speaker segmentation is to extract speaker turns in speech segments. Note that we conduct voice activity detection and segmentation of the speakers in real-time, while the discussion is in progress: The recording system continuously writes audio signals on the hard drive, and the preprocessing methods are continuously applied to the audio signals available at any given time.

Voice activity detection algorithms typically use audio features like frequency, energy, and spectral entropy to extract speech activity from audio recordings. There are many voice activity detection systems available; we use the algorithm proposed in [17]. For speech segmentation, we use the approach of [18-19]. For each of the two participants, we extract two binary indicators that show voice status and speaking status at each time instance (see Fig. 3). Voice status roughly corresponds to syllables and speaking status corresponds to the speaking time of a person.

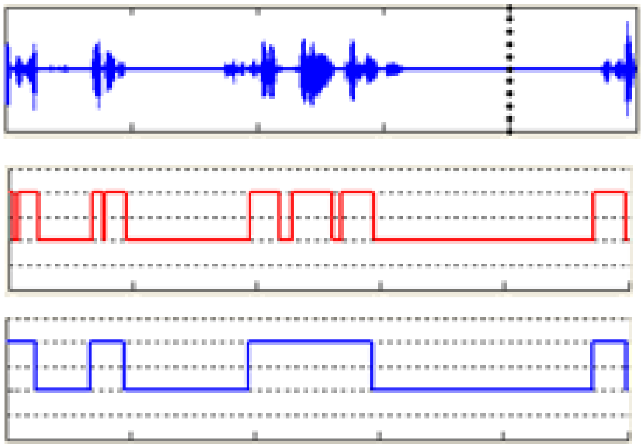
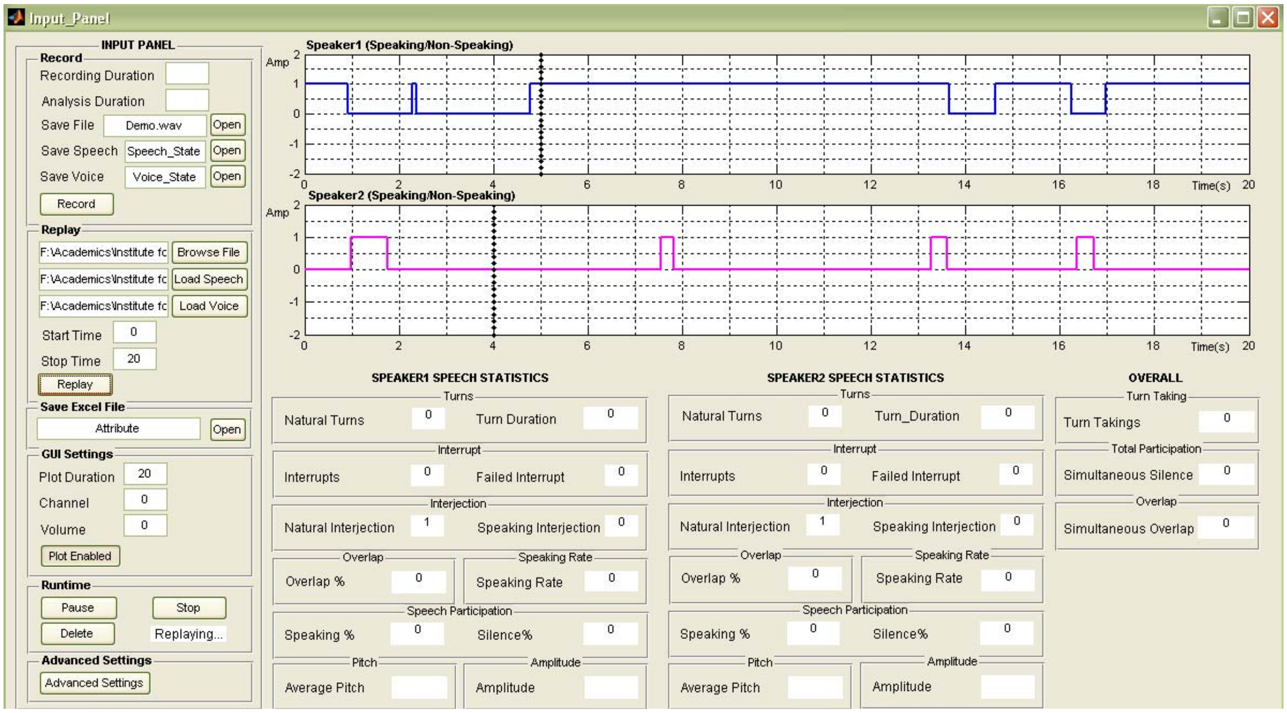


Figure 3.Illustration of voice activity detection and speaker segmentation. (top) Audio signal; (middle) Voice activity detection; (bottom) Speaker segmentation.



INPUT PANEL

Record

Recording Duration

Analysis Duration

Save File Demo.wav Open

Save Speech Speech_State Open

Save Voice Voice_State Open

Record

Replay

F:\Academics\Institute fc Browse File

F:\Academics\Institute fc Load Speech

F:\Academics\Institute fc Load Voice

Start Time 0

Stop Time 20

Replay

Save Excel File

Attribute Open

GUI Settings

Plot Duration 20

Channel 0

Volume 0

Plot Enabled

Runtime

Pause Stop

Delete Replaying...

Advanced Settings

Advanced Settings

Speaker1 (Speaking/Non-Speaking)

Speaker2 (Speaking/Non-Speaking)

SPEAKER1 SPEECH STATISTICS

Turns: Natural Turns 0, Turn Duration 0

Interrupt: Interrupts 0, Failed Interrupt 0

Interjection: Natural Interjection 1, Speaking Interjection 0

Overlap: Overlap % 0

Speaking Rate: Speaking Rate 0

Speech Participation: Speaking % 0, Silence% 0

Pitch: Average Pitch

Amplitude: Amplitude

SPEAKER2 SPEECH STATISTICS

Turns: Natural Turns 0, Turn_Duration 0

Interrupt: Interrupts 0, Failed Interrupt 0

Interjection: Natural Interjection 1, Speaking Interjection 0

Overlap: Overlap % 0

Speaking Rate: Speaking Rate 0

Speech Participation: Speaking % 0, Silence% 0

Pitch: Average Pitch

Amplitude: Amplitude

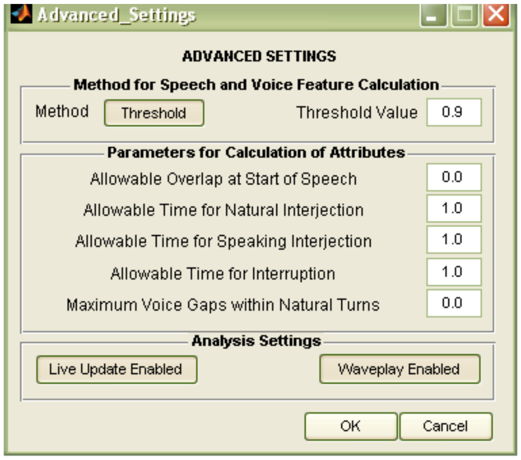
OVERALL

Turn Taking: Turn Takings 0

Total Participation: Simultaneous Silence 0

Overlap: Simultaneous Overlap 0

Figure 4.The Graphical User Interface (GUI) displays the speaker segmentation, and below those plots various non-verbal speech measures are displayed. On the left hand side, user Input Panel is at the top and controls panel is beneath it. The Advances Settings Panel (see Fig. 5) allows the user to specify parameter values for voice activity detection and speaker segmentation.



ADVANCED SETTINGS

Method for Speech and Voice Feature Calculation

Method Threshold Threshold Value 0.9

Parameters for Calculation of Attributes

Allowable Overlap at Start of Speech 0.0

Allowable Time for Natural Interjection 1.0

Allowable Time for Speaking Interjection 1.0

Allowable Time for Interruption 1.0

Maximum Voice Gaps within Natural Turns 0.0

Analysis Settings

Live Update Enabled Waveplay Enabled

OK Cancel

Figure.5. Advanced GUI Settings.

C. Graphical User Interface

After voice activity detection and speaker segmentation has been carried out, the non-verbal speaking statistics discussed in Section 2 are computed in real-time. We have designed a Graphical User Interface (see Fig. 4 and 5) that displays and continuously updates those measures (while those statistics are regularly being stored in data files).

The GUI is designed to accept a set of user inputs e.g. filenames, and parameters of voice activity detection and speaker segmentation procedures.

D. Visual Voice Activity Detection and Speaking Segmentation

To improve the robustness of voice activity detection to background noise and other non-speech sounds, we have implemented visual voice activity detection. The visual information about speaking or not speaking will be integrated with audio information in the future.

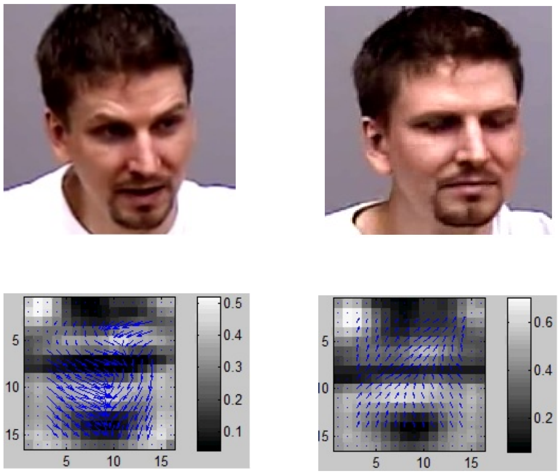


Figure 6. Face detection and optical flow for (left) Speaking sequence and (right) Silent sequence.

Our proposed algorithm first detects the faces using the method proposed in [20], and next the lip regions are extracted. We calculate optical flow of two sequential frames to infer vertical and horizontal lip motion [21], as illustrated in Fig.6. Speaking is mostly associated with vertical lip motion, which ultimately enables us to detect speaking from video.

E. Accuracy

To assess the accuracy of speaker segmentation based on audio and video signals, we have recorded 4 two-person conversations. From each of those recordings, we extracted a segment of 4 minutes. We manually labelled the speakers at each time instant, which will serve as ground truth to assess the speaker segmentation algorithms. The results for audio and video-based speaker segmentation are summarized in Table I. Both approaches seem to be reliable, and not surprisingly, audio-based segmentation is more reliable than video-based. In the future, we will combine both approaches to further improve speaker segmentation.

TABLE I. ACCURACY OF AUDIO AND VIDEO- BASED SPEAKER SEGMENTATION

Session	Audio Based	Video Based
1	96%	83%
2	93%	77%
3	94%	85%
4	94%	82%

IV. Feedback

A crucial component in the proposed system is feedback. The non-verbal speaking statistics, calculated in real-time, can be exploited to provide feedback to the speakers, either in real-time or offline fashion (after the discussion). The GUI discussed in Section III.C is one potential approach to feedback. It can be used by an external observer to analyse a conversation in real-time or after the discussion. However, as the GUI displays numerous statistics, it may not be straightforward to grasp the essential interactions, especially in real-time and for non-experts. Also for real-time feedback to the speakers, the GUI is not suitable, as the speakers cannot extract relevant information from the GUI without interrupting the conversation.

So far, we have explored two alternative means of feedback: (i) retrospective (offline) feedback, after the discussion, in the form of animations; (ii) real-time feedback to the speakers through emoticons displayed on smartphones. We will now briefly discuss those two forms of feedback.

A. Retrospective Feedback

In this approach, each participant is depicted as a character in an animation. Using a commercial game engine (Unity) data from the conversation is processed automatically to produce the animation. This animation aims to highlight selected significant non-verbal interactions in the discussion, e.g., smooth turn taking, inappropriate silence, excessive interruptions, or unbalanced voicing rates between the speakers (see Fig.7).

This approach can visualize the many complex threads of information in a manner that is relatively intuitive for participants to review and comprehend.

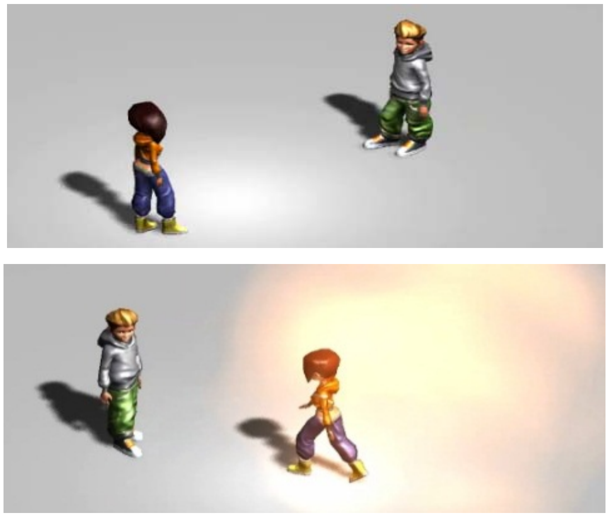


Figure 7. Retrospective feedback in the form of animations: Continuous turn taking (top) and excessive interruptions (bottom).

B. Real-Time Feedback on Smartphones

Real-time feedback to the speakers may help them to adapt their individual behaviour within the group and increase the effectiveness of a conversation. However, designing real-time feedback is a challenge, as it can easily interrupt the flow of a discussion. Auditory feedback could easily be disturbing, and therefore, we decided to design graphical feedback instead. Images can easily portray the behaviour of individuals as they can be self-explanatory.

In our approach, every participant is given feedback through emoticons displayed on a smartphone, as illustrated in Figs. 8 and 9. Whenever a significant event occurs (e.g., excessive interruptions), an emoticon is displayed. As we only trigger such feedback for the most significant interactions, the amount of feedback is limited and does not disturb the flow of a conversation. In future work, we will experiment more with such real-time feedback to assess its effectiveness and effect on conversations.

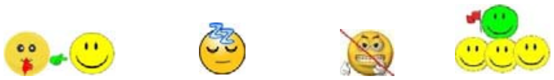


Figure 8. Examples of emoticons. From left to right: Interrupting; Monotonic speech; Aggressive behaviour; Emerging leader.

V. Conclusion and Future Work

In this paper, we have presented preliminary results on our systems for automated real-time and offline analysis of conversations from audio and video recordings. We have developed a user-friendly GUI for analysing a discussion, both in real-time (for external observer) and in retrospective fashion.

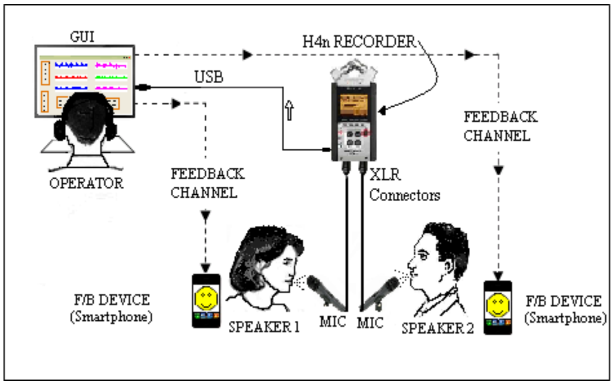


Figure 9. Real-time feedback system. The system provides feedback through the GUI, which may be used by an external observer (“operator”) to monitor the discussion; it also provides feedback to the speakers in the form of emoticons displayed on smart phones. Note that the system does not rely on any external observer or operator, as it is fully automatic.

We are in the process of designing animations that summarize salient social interactions during a discussion (e.g., interruptions), for retrospective analysis. We have also introduced a system that provides real-time feedback to individual participants during on-going conversations, in the form of emoticons on smart phones. In the longer term, such systems may help to boost the effectiveness of a diverse range of social interactions, e.g., job interviews, business meetings, group discussions, coaching sessions, or public speaking.

ACKNOWLEDGMENTS

This research project is supported in part by the Institute for Media Innovation (Seed Grant M4080824) and the Nanyang Business School, both at Nanyang Technological University (NTU), Singapore. We would like to thank Mr.VincentTeo and his colleagues at the Wee Kim Wee School of Communication of NTU, for the technical support. We are grateful to the lab managers and colleagues at Control Engineering Lab at NTU for their valuable support, and thank the participants for the test recordings.

REFERENCES

[1] A. Pentland, “Honest Signals: How They Shape Our World”, MIT Press, Cambridge, MA, 2008.

[2]A. Pentland, “Socially aware computation and communication”, IEEE Computer, vol. 38, no 3, pp. 33-40, 2005.

[3] M.S. Poole, A B. Holligshead, J. E. McGrath, R L. Moreland, and J.Rohrbaugh, “Interdisciplinary perspectives on small groups”, Small Group Research, vol. 35, no. 1, pp. 3-16, 2004, Sage Publications.

[4] E. Salas, D. E. Sims, and C. S. Burke, “Is there a big five in teamwork”, Small Group Research, vol. 36, no. 5, pp. 555-599, 2005, Sage Publications

[5] D. Gatica-Perez, “Automatic nonverbal analysis of social interaction in small groups: a review”, Image and Vision Computing, vol. 27, no 12, pp. 1775-1787, Dec 2009, Elsevier.

[6] O. Aran, H Hung, and D. Gatica-Perez, “A Multimodal Corpus for Studying Dominance in Small Group Conversations”, Proc. LREC workshop on Multimodal Corpora and 7th International Conference for Language Resource and Evaluation, Malta, 2010.

[7] R. J Rienks and D. Heylen, “Automatic dominance detection in meetings using easily detectable features”, Proc. of the Workshop on Machine Learning for Multimodal Interaction (MLMI), Edinburgh, 2005.

[8] D. Sanchez-Cortes, O. Aran, M. Schmid-Mast, and D. Gatica-Perez, “Identifying Emergent Leadership in Small Groups using Nonverbal Communicative Cues”, 12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI), pp. 39, Beijing, China, 2010, ACM.

[9] F.Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, “Multimodal recognition of personality traits in social interactions”, Proc. 10th International Conference on Multimodal Interfaces, pp. 53-60, Chania, Oct 2008, ACM.

[10] A. Pentland, “Social Dynamics: Signals and Behaviour”, International Conference on Developmental Learning (ICDL), vol. 5, IEEE press, 2004.

[11] O. Aran and D. Gatica-Perez, “Fusing Audio-Visual Nonverbal Cues to Detect Dominant People in Conversations”, 20thInternational Conference On Pattern Recognition (ICPR), pp. 3687-3690, Istanbul, Turkey, Aug 23-26, 2010, IEEE.

[12] J.Carletta et al, "The AMI meeting corpus: A pre-announcement", Proc. Machine learning for Multimodal Interaction (MLMI), pp. 28-39, Edinburgh, Jul 2005.

[13] D. Sanchez-Cortes, O. Aran and D. Gatica-Perez, “An Audio Visual Corpus for Emergent Leader Analysis”, (ICMI-MLMI), Multimodal Corpora for Machine Learning, Nov 14-18, Alicante, Spain, 2011, ACM.

[14] T. Kim, A. Chang, L. Holland and A. Pentland. “Meeting mediator: Enhancing group collaboration with sociometric feedback”, Proc. of ACM Conf. on Computer Supportive Cooperative Work (CSCW), pp. 457-466, San Diego, 2008.

[15] Mike Brooks, VOICEBOX: Speech Processing Toolbox for MATLAB, 1999.

[16] M. R. Chial, “Suggestions for Computer Based Audio Recordings of Speech Samples for Perceptual and Acoustic Analyses”, Phonology Project Technical Report No. 13, Dept. of Communicative Disorders, Phonology Project, University of Wisconsin-Madison, Oct 2003.

[17] Sumit Basu, “Conversation Scene Analysis”, PhD Thesis, MIT, Dept. of Electrical Engineering and Computer Science, 2002.

[18] W. T Stoltzman, “Towards a Social Signaling Framework: Activity and Emphasis in Speech”, Master Thesis, MIT, Sep 2006.

[19] N. Ambady and R.Rosenthal, “Thin slices of expressive behaviour as predictors of interpersonal consequences: A meta analysis”, Psychological Bulletin, vol. 111, no. 2, pp. 256-274, 1992, American Psychological Association.

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in 2001 Ieee Computer Society Conference on Computer Vision and Pattern Recognition, Vol 1, Proceedings, A. Jacobs and T. Baldwin, Eds., ed, 2001, pp. 511-518.

[21] P. Dollar, "Piotr's Image and Video Matlab Toolbox (PMT)", available from http://vision.ucsd.edu/~pdollar/toolbox/doc/