This document is downloaded from DR-NTU (https://dr.ntu.edu.sg)
Nanyang Technological University, Singapore.

Real-time feedback system for monitoring and facilitating discussions

Constable, Martin; Dauwels, Justin; Dauwels, Shoko; Elgendi, Mohamed; Mengyu, Zhou; Rasheed, Umer; Tahir, Yasir; Thalmann, Daniel; Magnenat-Thalmann, Nadia; Sarda, Sanat

2014

Sarda, S., Constable, M., Dauwels, J., Dauwels, S., Elgendi, M., Mengyu, Z., et al. (2014). Real-Time Feedback System for Monitoring and Facilitating Discussions. In J. Mariani, S. Rosset, M. Garnier-Rizet & L. Devillers (Eds.), Natural Interaction with Robots, Knowbots and Smartphones (pp. 375-387): Springer New York.

https://hdl.handle.net/10356/93219

https://doi.org/10.1007/978-1-4614-8280-2_34

© 2014 Springer, Part of Springer Science+Business Media.

Downloaded on 30 May 2023 16:22:02 SGT

Real-Time Feedback System for Monitoring and Facilitating Discussions

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

Abstract–**In this paper we aim to provide a system that can facilitate a real-time analysis of the behaviour of speakers in an on-going conversation. The uniqueness of our system lies in its ability to compute necessary features, perform analysis and inform speakers during the discussion in real-time; whereas usually such analysis is conducted in an offline fashion. In social monitoring, various features are used to interpret and deduce talking mannerisms of people, and gain insights on human social characteristics and behaviour. In our system, several discussion statistics, such as speaking length, speaker turns, and speaking turn duration, were computed and displayed in real-time. Visual data of the on-going discussions are also used to corroborate the analysis. 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 in various applications, e.g., for coaching purposes, and for facilitating business meetings.**

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

I. Introduction

Conversations, discussions, meetings and other social interactions are integral in daily lives. 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, how much he or she interrupts another person while speaking [1]. Talking mannerisms of individuals play an important factor for meetings to be productive i.e. to achieve certain objectives. Other factors as for instance availability of data, know-how, and difference of opinions are also equally important. Many a times, to attain the objective, the meeting tend to be longer or it takes more than few meetings. Among the factors one may control, if talking mannerisms of the people become mutually acceptable or aligned, it increases chances of meetings to be more productive and efficient [2]. We intend to build such system to make the talking mannerisms of people mutually compatible.

The above premise is based on research and studies, carried over long period of time, in the fields of psychology and cognitive science, where human behaviour is studied from social interactions. [3-4]. Results obtained in above fields are recently being re-assessed using automatic detection of talking mannerisms in social computing, due to advances in recording equipment [5] and signal processing. Recent work mostly analyses the predictions to deduce individual characteristics like dominance status [6-7], emerging leadership [8] and also other personality related traits [9-10].

Different features from speech are extracted and coded to obtain statistics of speaking manners. Also, the combination of speech and visual (multimodal) features has been shown to provide increased accuracy for detecting characteristics like dominance [11] and leadership [8]. But in general, irrespective of the characteristic to be interpreted, often the same measures are used.

Corpora (databases) with numerous types of recordings are available online, including meetings in different scenarios [12]. In recent years, many corpora have been developed related to small-group interactions. A good survey of such corpora is presented in [13]. These corpora are continuously updated with different types of annotations that label individual characteristics in those meetings. These annotations are used as gold standard for speaking characteristics obtained from automated analysis methods [6].

In the present work, we develop a system that, from speech signals, in real-time calculates and display statistics of speaking mannerisms. Rather than extracting a single speaking characteristic, this approach provides us the flexibility to quantify numerous speaking characteristics. Most of those measures are similar to those considered in recent works. Real time feedback is also provided in few systems, however, not based on speech signals [13-14]. We also propose a system of

real time feedback to every participant in the meeting, to inform them about their speaking mannerisms, and if needed, to suggest alternatives.

In this paper, we limit ourselves to automatic analysis of conversations of two people. We are specifically interested in face-face discussions for applications in coaching, interviewing techniques, self-assessment, etc. In the future, we plan to scale our system towards small group interactions.

In Section 2, we explain the different audio statistics that our system uses, the voice activity detection system, and the different talking measures. Section 3 describes the GUI, the need for benchmarking, and provides a brief overview of different recording solutions. Section 4 explains the proposed framework for real-time individual feedback. Finally we present our conclusions and suggestions for future work in Section 5.

II. Nonverbal Speech Features

Since the goal of this work is to build computational models to analyse and predict, in real-time, the behaviour of the individuals in the conversation and the emerging nature of the discussion. In order to address these goals the extraction of non-verbal cues in necessary. This section illustrates the feature extraction techniques that are used.

A. Measures

The succeeding paragraphs briefly describe the nonverbal cues used. The analysis of visual features is also relevant but has been deferred as part of future work. A voice activity detection system uses audio features like frequency, energy, spectral entropy and their variants to extract the speech activity from the audio recordings of the individual participants. There are many voice activity detection systems available [15]. For each of the two participants, we extract two binary indicators that show voice status and speaking status at the each time with a rate of 8000 [16-17].

Voice status roughly corresponds to syllables and speaking status corresponds to the speaking time of a person. Based on this information, we perform coding of following measures for two participants. Note that these nonverbal cues had to be computationally efficient in order to be compatible with a real-time system.

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

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.

Voicing Rate: The number of syllables spoken by person during a conversation.

Interruption: The number of times person *'A'*, interrupts person *'B'* while speaking and takes over. Person *'B'* stops speaking before person *'A'* does.

Failed Interruption: The number of instances when person *'A'* interrupts person *'B'* while speaking but stops speaking before person *'B'* does.

Interjection: This indicates short utterances like 'no', 'ok', 'yeah', 'exactly' etc.

Speaker Turn Duration: This computes the average speaker turn duration.

Overlap Percentage: Percentage of instances when either person has overlapped each other during the conversation.

Simultaneous: It computes the instances when both people start talking simultaneously, with neither interrupting each other.

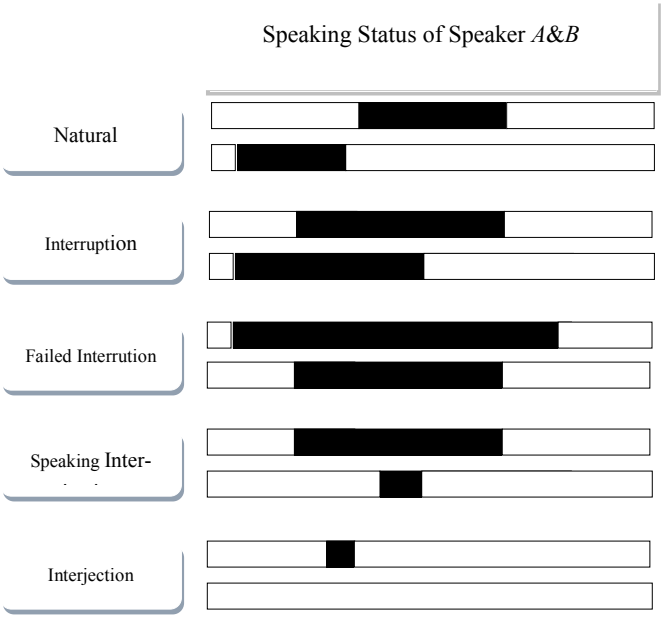

Figure 1: Describing the statistical measures of interruption, failed, interruption, turn-taking and interjection derived from binary speaking status. Speaking status is obtained from voice activity detection system.

Each of the above features, individually or in combination portrays different characteristics of their behaviour during the discussion. For instance, high speaking percentage indicates dominant behaviour of a person; a high interruption count indicates aggressive behaviour etc. In the following manner, we can interpret various personality traits and dynamics of conversations from these statistical measures.

B. Benchmarking

If the number of instances of interrupts during a debate or during friendly conversation is same, do we interpret the person as highly interrupting and therefore dominant? We commonly hypothesize that it also depends on the context. It may be possible that the number of interrupts in the debate is miniscule. This may not mean that the person has highly inter-

rupting behaviour and therefore is dominating and/or aggressive.

Thus if there is a benchmark depending on context, interpreting behaviour would be more accurate e.g. Number of instances of interrupts above a benchmark the person is considered as highly interrupting and below which as non-interrupting. Benchmarking is to be devised for any of the measures or their combination. We believe due to benchmarking, trait of a person will be truly reflected. We also believe that benchmarking the contexts, will play important role towards generalizations of systems. In future work, we will showcase results from benchmark performance. .

III. Implementation

The implementation of the above mentioned non-verbal cues is detailed in the succeeding paragraphs.

A. Voice Activity Detection and Speaking Segmentation

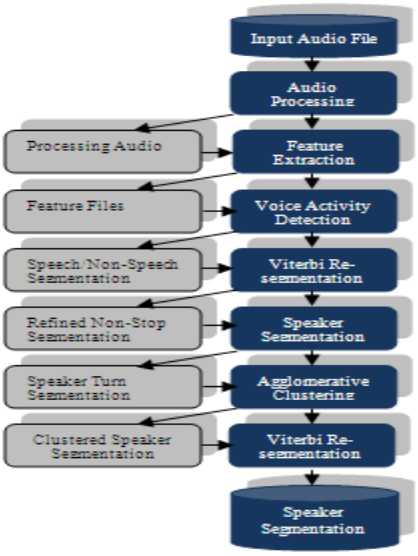

Figure: 2, Process of Voice Activity Detection and Speaker Segmentation

The voice activity detection and speaking segmentation of the speakers is the preliminary part in the system. Generally, the purpose of voice activity detection is the differentiation of the audio file between speech and non-speech (including silence and all kinds of noise). Whereas speaker segmentation tries to find speaker turns in speech segments which are long enough.

Voice Activity Detection and Speaker Segmentation are evaluated using the bottom up approach as shown in the Fig.2. and the results are illustrated in Fig.3.

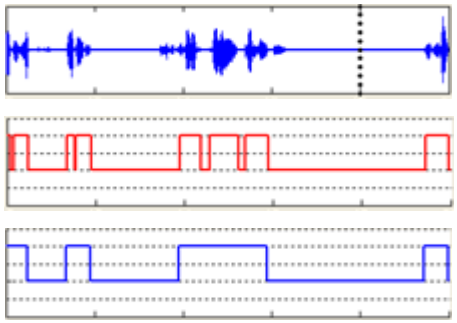

Figure: 3 a) Audio File. b) Voice Activity Detection. c) Speaker Segmentation

B. Graphical User Interface

With automatic analysis of measures, comes and obvious requirement of data representation in order to grasp the dynamics of the conversation. The audio file is played simultaneously with the plot of speaker segmentation; the user can easily observe an on-going discussion as well as analyse the accuracy of the system. For a given duration of recording, the GUI displays and continuously updates most of measures per second and the overall percentages at the end of the speech. The GUI is designed to accept a set of user inputs e.g. Discussion Time, Analysis Time (i.e. the period after which the analysis is performed), Volume, Audio Channel etc. as shown in the Fig.4.

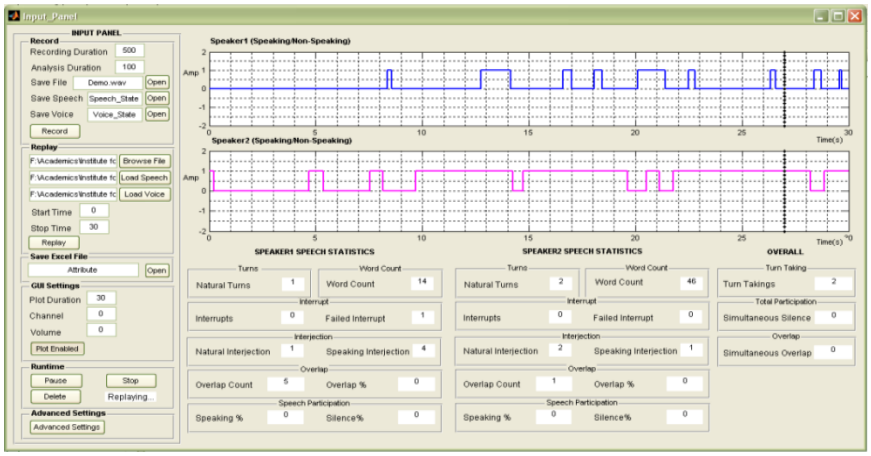

Figure 4: The Graphical User Interface. It displays Speaker Segmentation. Below each speaker signal are its different statistical measures. On the left hand side, user Input Panel is on the top and controls panel beneath it.

'Allowable Overlap at Start of Speech' refers to the allowed fraction of time before which the speaker 'B' can start his natural turn while speaker 'A' has not completely finished his turn. 'Allowable Time for Natural/Speaking Interjection/Interrupts' is the fraction of time which the speakers usually take to interject or interrupt respectively. 'Maximum Gaps within Natural Turns 'sets the maximum time of pause allowed for a speaker.

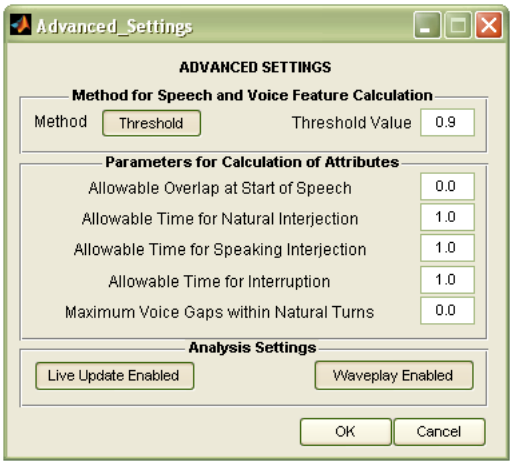

Figure 5: Advanced Settings GUI

C. *Sensing and Recording*

This section describes the recording system and equipment. To collect audio data, we prepared a static setup. It consists of two table top microphones, one per person. We used a ZOOM H4n portable voice recorder which allows us to record from multiple speakers simultaneously. It has flexible sampling rates and bit resolutions. 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. Recordings are directly saved on the laptop.

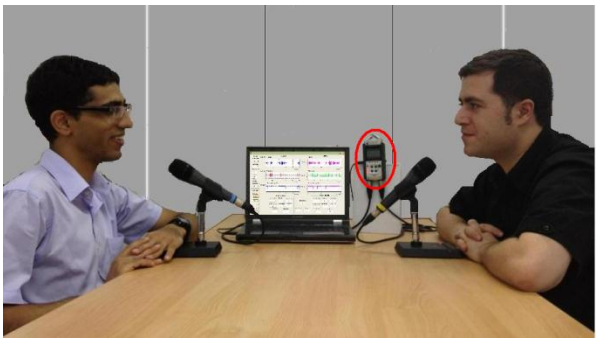

Figure 6: Recording setup using the Zoom H4n voice recorder (circled in red) with Sennheiser e845s microphones for two people conversation.

For online recording and real time feedback, recordings are required to be saved directly in laptop. With H4n recorder acting as interface, recordings from two separate, table-top microphones could be recorded synchronously without any delay. The setup with H4n recorder provides an easy, quick, cost effective, portable and most importantly undistorted, original signal recording. In [19], there is good overview of computer based audio recordings.

It is important to have the right microphones and connectors to get undistorted original speech signal. Microphones need to have flat frequency response to preserve original speech energy, optimally sensitive to allow talking from comfortable distance, directional to avoid capture of other signals and low handling noise. Connectors should be balanced to reject line noise interference. Also, they should not be imposing or making the speaker nervous. For our present recording, we use Sennheiser e845 s microphones with XLR connectors. The above solution is for two people recording. For small group interaction (>2), one may use professional audio interfaces, that allow simultaneously multi-channel recording.

D. *Visual Voice Activity Detection and Speaking Segmentation*

To cater for background noises 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 to make the system more robust towards non-speech sounds.

For this project we have implemented visual voice activity detection using optical flow [21-23]. The reason for selecting optical flow methodology is that it does not require exact lip extraction for each user, which makes it more flexible towards dealing with illumination issues and facial hair. Also as the lips deform slowly with time we do not need to continuously process each frame, by adding gap between frames the processing load is greatly reduced.

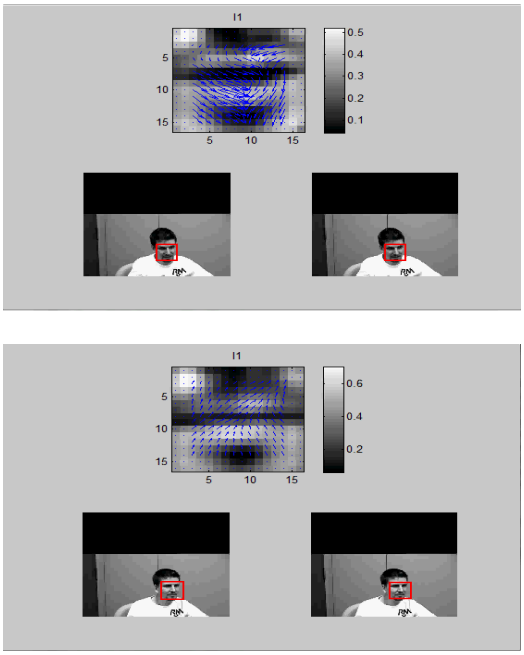

Figure 7: Face detection and optical flow plot for a) Speaking sequence and b) Silent sequence.

The algorithm works by taking video input continuously. Face is detected in sequential frames from the video using [20]. From the detected face the region of interest i.e. lip region is extracted. The optical flow of two sequential frames

gives us vertical and horizontal motion in the lip region as shown in figures using [24]. We then use covariance between these two movements as vertical motion is more relevant to lip movement. This measure of covariance shows clear differentiation between speaking and silent parts, thus enabling us to make a decision.

Fig.7.a shows the consecutive frames captured from a speech sequence and Fig.7.b shows the frames from a silent sequence. In the bottom half two frames are shown whereas the upper half shows the plot of optical flow vectors on the region of interest.

E. Accuracy

The accuracy of the non-verbal cues was satisfactory despite the static recording setup. The Table I illustrate the accuracy in the evaluation of audio and visual activity detection.

TABLE I. ACCURACY CALCULATION FOR VISUAL VOICE ACTIVITY DETECTION

session	Frames (25/sec)	Audio Accuracy	Video Accuracy
1	6000	96%	83%
2	6000	93%	77%
3	6000	94%	85%
4	6000	94%	82%

IV. Feedback

Once the initial examination of the offline data (recording from corpora) was carried out, the experimental setup was used to perform real-time analysis. The system was used to perform real-time analysis of the individual contribution of the speakers as well as the emerging group dynamics using the multimodal non-verbal cues.

The audio and visual non-verbal cues are displayed and updated after a scheduled period set by the user usually 5-10 minutes (Fig.4). This provides statistical analysis of on-going conversation while recording simultaneously. This setup of real time automatic analysis is the pre-cursor to provide real-time feedback to individual participant towards making conversations productive.

Each participant during an on-going conversation is fed back various data on the emerging group dynamics, his/her own individual contribution, behaviour etc. Such immediate feedback would provide opportunity not only to adapt personal behaviour within the group but also collaboratively help towards an efficient conversation. This will help to reach the objective of meeting sooner, resulting in increased productivity.

There can be various approaches to provide feedback to each participant. The two approaches currently being pursued include sending timely feedbacks via SMS on smart phones while the other is creating an animation which depicts roles of each individual in the discussion.

A. Sending Feedback via SMS on Smart Phones

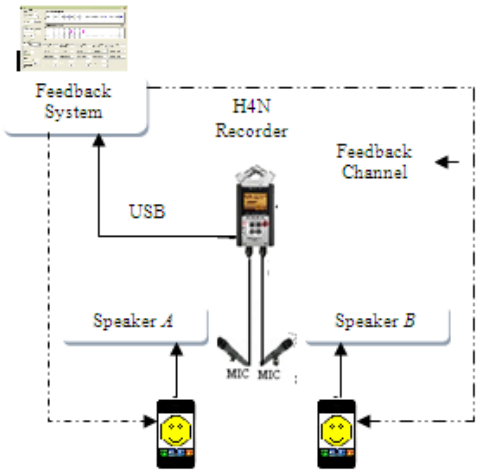

Figure 8: Real time individual feedback system. The Zoom H4n recorder acts as audio interface.

In this approach each participant will have a feedback device via smart phones etc. Every participant will be given timely feedbacks through the operator computer. The form of feedback is very critical. It should be such that it does not disturb the rhythm and concentration of the participant. An audio feedback could be disturbing to individual or the entire group. Graphical images can easily portray the behaviour of individuals as it is self-explanatory. In Fig.9, we present few samples of smiley for feedback that correspond to relevant behaviours.

Figure 9: Sample of Smileys. a) Interrupting, b) Sleeping, c) Boring, d) Aggressive, e) Emerging Leader

B. Creating Animation Based on Events of Discussion

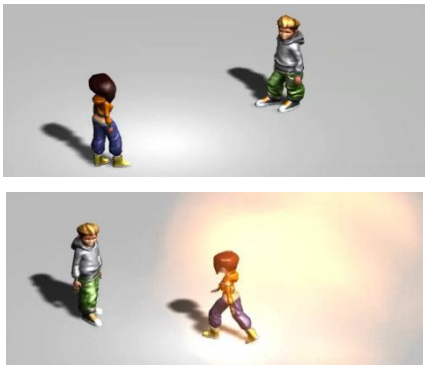

Figure 10: Real time feedback via animations. a) Turn Taking, b) Interrupt

In this approach each participant will be depicted as a character in the animation and based on the current events i.e. turn taking, silence, interruption etc. Various animations will be played in real-time to illustrate the current scenario and the direction of the discussions (Fig.10). The advantage of this approach is that it could present the many complex threads of information in a manner that is relatively easy for

participants to review and comprehend.

V. Conclusion and future works

In this paper, we described our experience in real time automatic analysis of conversations. We also presented a user-friendly GUI and provided a practical, portable, cost effective recording solution. We elaborated on the use of benchmarking for different contexts, which may help to better interpret the behavioural characteristics and eventually increase accuracy of group dynamics. We also introduced a system that provides feedback of speaking performance to individual participants during on-going conversations. Such system may yield greater efficiency, transparency and increasing productivity of discussions, thereby also leading to cost and resource savings. The system can be applied to various applications. For example, business meetings, coaching, lectures, therapy sessions etc.

In the future, we plan to extend our design in numerous directions including implementation of proposed feedback and gauging its performance and including the analysis based on gestures of the speakers. Also, we plan to develop our own corpora in specific contexts like coaching. External annotators will carry out manual annotation, to extract personality traits of participants in the corpora. We can then use those results as benchmark to test the accuracy of our system.

ACKNOWLEDGMENTS

This research project is supported by the Institute for Media Innovation (Seed Grant number M4080824) and the Nanyang Business School, at Nanyang Technological University (NTU), Singapore. We would like to thank Mr.Vincent Teo 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 we also 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, IEEE.

[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] SumitBasu, "Conversation Scene Analysis", PhD Thesis, MIT, Dept. of Electrical Engineering and Computer Science, 2002.

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

[17] 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.

[18] J. R. Curhan and A. Pentland, "Thin Slices of Negotiation: Predicting Outcomes from Conversational Dynamics within the First 5 minutes", Journal of Applied Psychology, vol. 92, no. 3, 802-811, 2007, American Psychological Association

[19] 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.

[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] S. Tamura, K. Iwano, and S. Furui, "Multi-modal speech recognition using optical-flow analysis for lip images," Journal of Vlsi Signal Processing Systems for Signal Image and Video Technology, vol. 36, pp. 117-124, Feb-Mar 2004.

[22] A. J. Aubrey, Y. A. Hicks, and J. A. Chambers, "Visual voice activity detection with optical flow," Iet Image Processing, vol. 4, pp. 463-472, Dec 2010.

[23] P. Tiawongsombat, M. H. Jeong, J. S. Yun, B. J. You, and S. R. Oh, "Robust visual speakingness detection using bi-level HMM," Pattern Recognition, vol. 45, pp. 783-793, Feb 2012.

[24] P. Dollar, "Piotr's Image and Video Matlab Toolbox (PMT)".