Towards the Understanding of Gestures and Vocalization Coordination in Teaching Context

Roghayeh Barmaki
Department of Computer Science
University of Central Florida
barmaki@cs.ucf.edu

Charles E. Hughes
Department of Computer Science
University of Central Florida
ceh@cs.ucf.edu

ABSTRACT

Nonverbal behaviors such as facial expressions, eye contact, gestures, postures and their coordination with voice tone and prosody have strong impact on the process of communicative interactions. Successful employment of nonverbal behaviors plays an important role in interpersonal communication in the classroom between students and the teacher. Student teachers need to improve their teaching skills, from communication to management, and prior to entering the classroom. To support these aspects of teacher preparation, we developed a virtual classroom environment, TeachLivE^{TM} for teacher training, reflection and assessment purposes. In this work we investigate the connections between gestures and vocalization characteristics of participants in a teaching context for two settings within the TeachLivE environment.

We have developed an immediate feedback application that is presented to the participants in one of the study settings. It provides visual cues to the participant in front of the tracking sensor any time that she exhibits a closed stance. Identification of these type of connections between acoustic and gestural components of communication provides an added dimension that could assist us in using machine learning methodologies to extract multimodal features as teaching competency measures.

Keywords

gesture; vocalization; nonverbal behavior; Microsoft Kinect; virtual teaching rehearsal environment.

1. INTRODUCTION

Interpersonal communication involves a variety of modes and components in communication. We might think that actual words are the primary part of communication; however, the majority of interaction between individuals, including students and teachers, is nonverbal, encompassing between 65 and 93 percent of what occurs related to learning [7]. These nonverbal elements include both nonvocal (e.g. body language) and vocal components (e.g. voice pitch and intonation). Body language by itself include several aspects: facial expressions, eye contact, posture or stance, gestures, touch and appearance. This research investigates the connection of postures and/or gestures with acoustic components of the nonverbal communication in the teaching context.

Multimodal analysis co-processes two or more parallel input streams (modes) from human-centered interactions that contain rich high-level semantic information [9]. Teaching and learning have always been multimodal as both are unified with speech, gesture, writing, image and spatial setting [12]. Multimodal data analysis in a teaching context helps us to have an informed understanding of the performances of the teacher participants.

TeachLivE is a simulated classroom setting used to prepare teachers for the challenges of working in K-12 classrooms. Its primary use is to provide teachers the opportunity to rehearse their classroom management, pedagogical and content delivery skills in an environment that neither harms real children, nor causes the teacher to be seen as weak or insecure by an actual classroom full of students. TeachLivE uses its underlying multi-client-server architecture called AMITIES- Avatar Mediated Interactive Training and Individualized Experience System [8]. A human-in-the loop (called an interactor) orchestrates the behavior of the virtual students in real-time based on each character's personality and backstory, a teaching plan, various genres of behaviors and the participant's input. The virtual classroom is displayed on a large TV screen to the participant and the view of the virtual classroom scene dynamically changes based on the participant's movements in front of the tracking sensor. We have developed a real-time gesture recognition application for nonverbal communication skill training, based on the Microsoft Kinect SDK [1] as part of ReflectLivE, the TeachLivE integrated reflection tool [3]. The hypothesis is that our developed feedback application has positive impact on the participants' body language, leading to more open and fewer closed stances. The open stance has arms and legs not crossed in any way. To explore the validity of this hypothesis and system usability evaluation, we report the results from the conducted case study with two settings using the feedback application (section 2.1).

We are also interested in looking at the connections between the participant's gestures and acoustic characteristics in different situations in the classroom, such as while asking questions from virtual students, conversation turn-taking after students' responses, introducing a new topic, etc. The analysis of the recorded sessions from a gesture-voice aspect is another motivation for this research that seeks a broader understanding of communication practices that reflect and support teaching competency.

Investigating the related research, there have been a number of prior attempts to develop social skill training and

feedback applications using interactive environments. Presentation Trainer [10] collects multimodal data using the Microsoft Kinect and provides immediate cues about the trainee's body posture, embodiment and voice volume during her presentation. Similarly, Dermody and Sutherland [5] present a multimodal prototype for public speaking purposes that uses the Kinect sensor. Their system provides real-time feedback on gaze direction, body pose and gesture, vocal tonality, vocal dysfluencies and speaking rate.

At first glance, gesture and speech may be coupled less directly than, e.g., prosody and speech, as both originate in very different physiological systems. However, some views and findings suggest a close connection between both, especially in production. This mutual co-occurence of speech and gesture reflects a deep association between the two modes that transcends the intentions of the speaker to communicate [11].

2. APPROACH

We present our research to understand the gesture and vocalization connections in the following two separate subsections since most of our currently reported research has been done independently with our effort to fuse the collected multimodal data still under development.

2.1 Gesture

This research evolved based on the existing literature expressing the importance of open body gesturing in successful interactive teaching (teaching competency) [2]. Reviewing the existing recordings of teaching sessions in TeachLivE gave us a baseline about the way teachers use their body in the virtual classroom. In our observations, most of the teachers were not thoughtful of their body movements and many of them exhibited closed stances most of the time in their teaching sessions. The recognized frequent closed postures (or closed gestures) were hands folded in front and back, hands on hips, and crossed arms. These gestures are noted as closed or "not-recommended" gestures. We are interested in detecting these closed gestures and reminding the trainees about their closed body language. In social skill training, the impact of immediate and real-time feedback in the rehearsal process has been reported as very positive in comparison to other types of feedback provision such as delayed feedback [10]. The developed feedback application is capable of providing visual or haptic (vibration wrist band) prompts in real-time for targeted closed gestures. The effectiveness of the implemented visual feedback application was evaluated by conducting a user study. It was a single-time within-subjects, counterbalanced study with two settings (TeachLivE with and without feedback application) and each session was 7-minute long. Participants (N=30, 6M, 24F) were asked to attend both of the settings, and complete pre and post questionnaires (the total recruitment time was approximately 45 minutes per participant). We randomly assigned the participants into two groups A and B, where group A (N=15, 3M, 12 F) experienced TeachLivE with feedback setting in their second session and group B had this experience in their first session. The collected full-body tracking data from the participants was processed [3] to extract the percentage of time that a subject exhibited closed gestures (CGP) in the recorded sessions. Our expectation based on the hypothesis (section 1) was that we would observe a considerable difference between groups A and B in the first session and a slight difference between the two groups in the closed body gesture employment in the second session. To evaluate the impact of our proposed feedback application on body language thoughtfulness, we calculated CGP for 60 recorded clips from 30 participants. The boxplot in Figure 1 presents the distribution of CGP between two groups of participants.

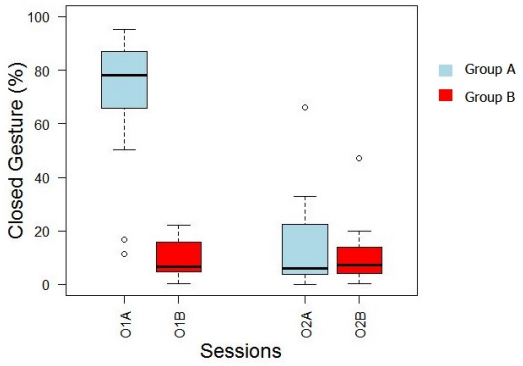

Figure 1: Medians and interquartile ranges of CGP exhibition in two sessions (observations) among groups A and B. Circle represent outliers.

Figure 1 shows some of key findings from this study. It presents the wide range (from 95% to 16%) of closed gesture employment for group A in the first session. It also indicates the median of CGP for group B participants is lower than group A (6.4 % and 7.2% for two sessions for group B and 78% and 5.9% for group A). As Figure 1 indicates, the hypothesized statement is supported for the participants of the study. The average time that all of the participants in group A exhibited closed gestures reduced significantly from their first session to their second session. Most interestingly, the participants in group B exhibited open gestures most of the time even in the second unaided session.

2.2 Vocalization

In this study, we recorded video, audio, full body tracking data and event logging information (including virtual students' talk-time and behaviors) from the TeachLivE system. The reader can find further recording details in [3].

After collecting the data, we processed the recorded audio from video sessions using Audacity software to extract the Waveform Audio File Format from recorded avi files. We opened the .wav files in the Praat tool [4] and extracted some basic vocal characteristics (pitch and intensity objects) from the audio files. Praat is a free computer software package for the analysis of speech. Voice pitch is the perceptual correlate of vocal fundamental frequency and voice intensity indicates voice loudness in db. A PitchTier object represents a time-stamped pitch contour (hereby feature), i.e. it contains a number of (time, pitch (Hz)) points, without voiced/unvoiced information. An IntensityTier object represents a time-stamped intensity contour, i.e., it contains a series of (time, intensity) points [4]. Pitch and intensity tier associated with our recorded sessions were exported for multimodal analysis purpose to the ANVIL [6]. ANVIL is a video annotation tool that offers multi-layered annotation

based on a user-defined coding scheme. Figure 2 shows the ANVIL tool.

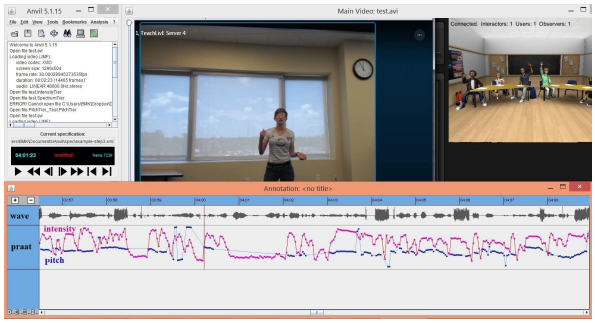

Figure 2: TeachLivE video sessions (including the participant front view and virtual classroom scene) within the ANVIL annotation tool [6]. Three acoustic contours waveform, pitch (blue) and intensity (pink) [4] are imported to the annotation project.

We intend to add our gesture recognition application output as an extended contour in the ANVIL. This will automatically present the types and timing for different closed gestures during the recorded session. The current version of the ANVIL does not support the exported (closed) labels of frames from the Kinect V2 gesture recognition tool as a contour, so we are working on this open-source tool to develop our desired contour structure. As mentioned earlier, our goal of using ANVIL is to understand the correlations of acoustic features with gesturing in these three main cases: 1) when the participant teacher asks a question from virtual classroom, 2) when the teacher listens to the responses from the class (conversation turn taking between students and teacher), and finally 3) when the teacher introduces a new or abstract topic or is summarizing the discussion. Literature supports that teachers gesture more in the mentioned cases [2]. We will annotate the recorded videos based on the teaching plan, conversational cases, open/closed, and affirmative gesture employment. The automatically generated vocalization information would be exported in conjunction with manual annotation data for further analysis.

3. CLOSING REMARKS

The study reported here fills a gap in multimodal research for education. In this paper, we first explained the impact of nonverbal behaviors in teaching competency. We then reported a case study to evaluate the performance of our developed feedback application for nonverbal communication skill training. We used the Microsoft Kinect sensor and its full-body tracking data stream to develop our real-time gesture feedback application. The results from the recorded body tracking data indicated the positive impact of informed body language and gesture in communication proficiency. We also introduced relevant tools and techniques for multimodal feature extraction for teaching competency, and we expect to report the results after developing an appropriate coding scheme framework and the annotation procedure.

For future research, we are looking forward to uncovering additional teaching evaluation insights with the analysis and evaluation of multimodal recorded data, as multimodality is an integral part of teaching.

Acknowledgments

The authors acknowledge the support of the Bill & Melinda Gates Foundation (OPP1053202) and the National Science Foundation (CNS1051067, IIS1116615). We also wish to express our gratitude to the entire TeachLivE team, especially the interactors who give life and authenticity to our avatars. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

4. REFERENCES

[1] Visual gesture builder: A data-driven solution to gesture detection. http://aka.ms/k4wv2vgb, July 2014. Retrieved 3/10/2016.

[2] M. W. Alibali and M. J. Nathan. Teachers' gestures as a means of scaffolding students' understanding: Evidence from an early algebra lesson. *Video research in the learning sciences*, pages 349–365, 2007.

[3] R. Barmaki and C. E. Hughes. Providing real-time feedback for student teachers in a virtual rehearsal environment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 531–537, New York, NY, USA, 2015. ACM.

[4] P. Boersma and D. Weenink. Praat: doing phonetics by computer [computer program] version. 6.0.17, 2016. Accessed 5/07/2016 from http://www.praat.org/.

[5] F. Dermody and A. Sutherland. A multimodal system for public speaking with real time feedback. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 369–370, New York, NY, USA, 2015. ACM.

[6] M. Kipp. Anvil: The video annotation research tool. In *The Oxford Handbook of Corpus Phonology*. Oxford University Press, 2014.

[7] A. Mehrabian. *Silent Messages: Implicit Communication of Emotions and Attitudes*. Wadsworth, 1972.

[8] A. Nagendran, R. Pillat, A. Kavanaugh, G. Welch, and C. Hughes. A unified framework for individualized avatar-based interactions. *Presence: Teleoper. Virtual Environ.*, 23(2):109–132, Aug. 2014.

[9] S. Oviatt and P. R. Cohen. *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces*. Morgan & Claypool Publishers, 2015.

[10] J. Schneider, D. Börner, P. van Rosmalen, and M. Specht. Presentation trainer, your public speaking multimodal coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 539–546, New York, NY, USA, 2015. ACM.

[11] P. Wagner, Z. Malisz, and S. Kopp. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232, 2014.

[12] M. Worsley, K. Chiluiza, J. F. Grafsgaard, and X. Ochoa. 2015 multimodal learning and analytics grand challenge. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 525–529, New York, NY, USA, 2015. ACM.