Nguyen AT., Chen W., Rauterberg M. (2014). Mastering the art of persuasion - intelligent tutoring system for presenters. In: Proceedings of International Conference on Computer Vision Theory and Applications (VISAPP, vol. 3, pp. 83-90), IEEE press.

Mastering the Art of Persuasion
Intelligent Tutoring System for Presenters

Anh-Tuan Nguyen, Wei Chen and Matthias Rauterberg
Department of Industrial Design, Eindhoven University of Technology,
Postbus 513, 5600 MB Eindhoven, The Netherlands
{*a.nguyen, w.chen, g.w.m.rauterberg*}*@tue.nl*

Keywords: Body Motion Analysis, Depth Vision, Nonverbal Behavior, Social Signal Processing.

Abstract: Public speaking is a non-trivial task since it is affected by how nonverbal behaviors are expressed. Practicing to deliver the appropriate expressions is difficult while they are mostly given subconsciously. This paper presents our empirical study on the nonverbal behaviors of presenters. Such information was used as the ground truth to develop an intelligent tutoring system. The system can capture bodily characteristics of presenters via a depth camera, interpret this information in order to assess the quality of the presentation, and then give feedbacks to users. Feedbacks are delivered immediately through a virtual conference room, in which the reactions of the simulated avatars can be controlled based on the performance of presenters.

1 INTRODUCTION

Public speaking is the art of persuasion. It has the tremendous impact on the success of everyone (Seiler and Beall, 2004, p.102). Unfortunately, delivering an oral presentation is not as simple as computer data transmission. Instead, the audience simultaneously perceives the messages via various non-spoken channels, which are known as nonverbal behaviors. On one hand, the content of a presentation must be *clear, vivid and appropriate* (Rodman and Adler, 1996). On the other hand, the significant component of a presentation lies upon nonverbal cues, which *has the power to change the meaning assigned to the spoken words* (Seiler and Beall, 2004, p.241).

Nonverbal behaviors of public speakers are expressed via several channels such as voice, gesture and facial expression. They have been proven to have greater influence than verbal cues. For example, a research by (Argyle et al., 1971) showed that, nonverbal messages are twelve to thirteen times more powerful than verbal ones. Similarly, according to (D'Arcy, 1998), the audience receives more than half of information from body language. The same result was found during the study of (Seiler and Beall, 2004), in which most people unconsciously more believe in nonverbal than verbal communication.

Practicing to express the effective nonverbal behaviors is difficult due to the fact that, they are mostly expressed subconsciously. Thus, in order to achieve the positive learning results, learners must be provided with the appropriate feedbacks from skilled experts, which in most cases might be expensive to achieve. In parallel, the role of nonverbal behaviors in computing is becoming increasingly recognized by the development of the emerging fields, such as social signal processing (Vinciarelli et al., 2009) and affective computing (Picard, 2000). Therefore, computers have been equipped with the abilities to decode the complexity of humans non-spoken channels.

In the literature, there are several approaches toward the automatic recognition of nonverbal cues from presenters, such as (Hincks and Edlund, 2009; Kurihara et al., 2007; Pfister and Robinson, 2011; Silverstein et al., 2003). These approaches analyze some vocal and visual channels of presenters, thus can provide them with the information about their performance. For example, the system in (Silverstein et al., 2003) was built solely on vocal cues, by analyzing the physical characteristics of voice such as pitch or tempo. It was similar to the approach of (Pfister and Robinson, 2011), which was originated from a vocal emotion detection module. The authors applied the support vector machine (Duan and Keerthi, 2005) to analyze one presentation based on a set of 6 qualities and achieved the accuracy of 81%. The approach introduced in (Hincks and Edlund, 2009) might be the simplest. By relying on the importance of pitch variance in oral presentations, the system measured the changes in vocal pitch, and then give visual feedbacks

to promote pitch variation.

On the other hand, there are three systems that include visual cues in the analysis. In (Kurihara et al., 2007), the authors added face position and orientation as the approximation of eye contact, together with utterance, pitch, filled pauses and speaking rate. In contrast, (Gao et al., 2009) introduced the method that only based on visual information. Similar to (Kurihara et al., 2007), face orientation was used as an indication for eye contact. The authors tracked the trajectories of global body movement and head position. This information helped their system to rank the performance of the whole presentation using the RankBoost algorithm (Freund et al., 2003), achieved promising results. However, they did not consider the complexity of body parts. To the best of our knowledge, the system that was presented in (Nguyen et al., 2012) is the only one that included the configurations of single body parts.

The common drawback of most existing systems is that, they were not implemented based on empirical research of nonverbal behaviors. Moreover, although most of them provided mechanisms to deliver feedbacks to the presenters (except (Pfister and Robinson, 2011)), the forms of feedbacks are rather simple. They are text/images (Kurihara et al., 2007), sound (Silverstein et al., 2003) or lightning (Hincks and Edlund, 2009). These methods can only provide users with solely assessment information, without concerning the entertaining aspect of the system, which might be valuable for educational purposes.

This paper presents our progress in developing a tutoring system for public speaking, which assesses presentations based solely on the visual behaviors of presenters. Firstly, an empirical study was performed in order to investigate on the nonverbal cues that impact a presentation, serving as the ground truth. Next, a Microsoft Kinect was implemented for capturing skeletal representations of the presenters' body as input data for the analysis. The recognition process can detect if the behaviors appeared in real-time. Multi-class support vector machine was used to classify the quality of presentations into a four-degree scale with the recognition rate of 73.9% on a training/test database that includes 76 presentations. For the feedback, the system allows presenters to review their presentation, together with the analysis results. In parallel, we developed a simulated conference room as the real-time feedback mechanism.

In the next section, we will explain our empirical results from the recorded presentations. The current development status will be introduced afterward. The last section is for conclusions and future works.

2 NONVERBAL BEHAVIORS OF PRESENTERS

In order to gather the ground truth as the guidance for our system, an observation was performed. We collected data from a training class about public speaking skills for postgraduate students. Learners were asked to give short presentations (about one minute) in front of the audience, which includes about ten other learners and one or two coaches. The content of the presentations was freely chosen by the presenters. In fact, all presenters chose to talk about their own research, in the ways that it can be understood by all of the audience that might came from the different fields. After each presentation, the audience gave feedbacks and suggestions on how the presentation should be improved, in terms of nonverbal expressions. We set up a regular camera to record the presentations. In parallel, a Microsoft Kinect was used to capture the whole-body movement for our further signal processing, as well as behavioral studies (Figure 1). Data from Kinect was stored as the *.ONI files using the OpenNI SDK (http://www.openni.org/). Finally, after removed the unsatisfied videos (e.g. presenters moved out of the camera range), 39 presentations of 11 presenters (four females, seven males) were collected.

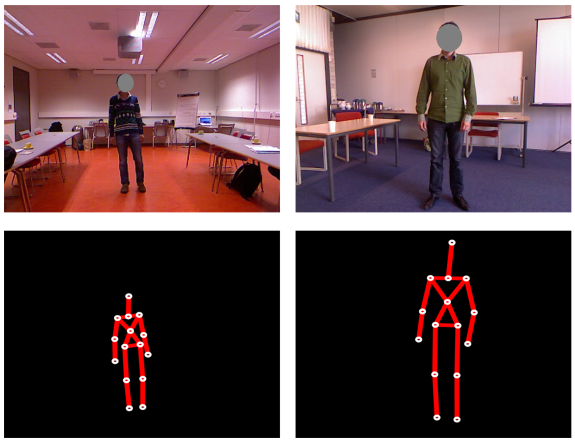

Figure 1: Two samples from the database, with the color images (top row) and the skeletal representations of presenters' bodies, which were extracted and stored using Microsoft Kinect and the OpenNI SDK (bottom row).

Regular videos were used for behavioral analysis. This task was done through the collaboration with an expert in public speaking. The role of the expert was to review the recorded videos, and then specifying the nonverbal cues that affected the performance of the speakers, together with the durations that they appeared. Thus, for each video, a set of behaviors was created. We collected the nonverbal cues and

Table 1: The list of observed nonverbal cues; each behavior is marked with either (+) or (-) if it improves or worsens a presentation. Point and State events are provided with rate of occurrences and percentage during observation, respectively.

#	Behaviors	Event Type (S/P)	No.	Rate of occurrences (times/minute)			Percentage during observation of the occurrences (%)		
				M	SD	Range	M	SD	Range
Postural behaviors									
1	(-) Shoulders too tight	S	19				60.94	23.80	12.67 - 98.50
2	(-) Standing ramrod style	S	12				73.02	36.44	5.15 - 100
3	(-) Legs too stretch	S	3				61.42	11.33	19.18 - 100
4	(-) Weight in one leg/foot	S	20				65.42	28.69	5.20 - 100
5	(-) Chin too high	S	14				64.94	23.80	12.67 - 98.50
6	(-) Hands in pockets/at back	S	3				11.85	4.89	12.76 - 96.20
7	(+) Lean forward	S	19				32.50	28.66	3.70 - 82.78
8	(-) Lean backward	S	17				62.80	28.07	12.73 - 96.20
Vocal behaviors									
9	(-) Speak too fast	S	19				45.88	36.55	7.32 - 100
10	(-) Start too fast	P	18						
11	(-) Energy decrease at the end	P	23	2.88	1.77	0.53 - 6.31			
12	(+) Vocal emphasis	P	33	5.51	4.51	0.59 - 17.50			
13	(+) Suitable pause	P	33	4.63	3.16	0.53 - 12.5			
14	(-) Unsuitable pause	P	20	1.73	1.14	0.53 - 5.19			
15	(-) Monotone	S	20				92.49	13.08	56.29 - 100
16	(-) Fillers	P	34	5.17	4.22	1.44 - 19.03			
17	(-) Stuttering	P	12	1.72	0.83	0.53 - 3.42			
Behaviors of eye contact									
18	(+) Make eye contact	S	39				93.81	8.24	75.00 - 100
19	(-) Eye contact avoidance	S	28				9.98	8.47	1.12 - 25.00
19.1	(-) Look up to ceiling	S	14				4.23	2.95	1.12 - 9.61
19.2	(-) Look down to floor	S	19				7.67	4.67	2.84 - 14.17
19.3	(-) Look at hands	S	11				10.24	3.15	4.40 - 13.15
Behaviors related to facial expression									
20	(+) Facial mimicry	S	30				39.31	25.97	4.50 - 91.81
21	(+) Smile	S	22				13.62	11.54	3.54 - 41.08
22	(-) Flat face	S	8				80.61	24.16	40.41 - 100
Behaviors related to whole body movement									
23	(-) Too much movement	P	11				42.21	25.87	4.68 - 89.32
24	(-) Too little movement	P	23				50.62	29.21	10.05 - 100
25	(-) Step backward	P	31	1.83	1.27	0.36 - 4.36			
26	(+) Step forward	P	34	2.06	1.04	0.59 - 4.61			
Behaviors related to hand gesture									
Amount of hand gesture									
27	Hand gesture occur	P	38	16.83	7.15	0.93 - 28.42			
28	(-) Hand gesture too little	S	20				69.55	34.64	17.21 - 100
29	(-) Hand gesture too much	S	10				61.49	31.82	27.34 - 96.10
Quality of hand gesture									
30	(-) Hand gesture bounded	P	30	6.75	5.33	1.00 - 19.77			
31	(+) Hand gesture relaxed	P	29	7.41	4.95	1.15 - 15.79			
32	(-) Hand gesture casual	P	10	5.16	3.14	1.56 - 10.28			
33	(-) Hand gesture uncompleted	P	27	3.23	2.78	0.93 - 10.27			
34	(+) Gestural emphasis	P	20	4.43	4.05	0.36 - 11.99			
35	(-) Hand Gesture repeated	P	31	6.57	2.49	1.09 - 12.31			
Synchronization with content									
36	(+) Gestural mimicry	S	25				4.81	3.04	1.19 - 11.05
37	(-) Gesture of being unsecured	S	11				28.3	33.17	6.25 - 99.30
Synchronization with voice									
38	(-) Stop speaking, quick gesture	S	1				9.91	0	9.91 - 9.91
39	(+) Synced with vocal emphasis	P	9	6.53	3.91	1.09 - 10.50			

then annotated their appearance using the commercial software Noldus Observer XT (Zimmerman and Bolhuis, 2009). Behaviors were categorized into either *State event* if their duration is necessary to be studied, or *Point event* otherwise. The software provided us with the statistical analysis on the appearance of these behaviors, including the number of presentations that contain the behaviors, the rate that they appeared (point events) and the percentage of time that they accounted for (Table 1).

The observed behaviors can be separated based on the nonverbal channels that they were generated: (1) Posture (the static configuration of body), (2) Voice (concerning the paralinguistic characteristics), (3) Eye contact, (4) Facial Expression, (5) Globe body movement, (6) Hand gesture. This method of categorization is similar to the literature of public speaking skills (Rodman and Adler, 1996). Due to the limited amount of space, we could not describe all of the observed behaviors in detail in this section. Only the behaviors that support our current development will be further explained in the next section. On the other hand, although we aimed to observe all of the available nonverbal cues, the contributions of each individual to the success of a presentation are unequal. From our observation, as well as advices from the expert, the following aspects are the most important:

- *Eye Contact*: Similar to social interaction, maintaining good eye contact is the first thing the presenters must keep in mind. It initiates and strengthens the connection between them and the audience (#18, 19 in Table 1). It might have the first and foremost influence to the performance of a presentation, as well as regular communications (Kleinke, 1986).
- *Amount of energy*: This aspect concerns the dynamic characteristics of a presentation, thus can reflect the internal state of the presenters. It has impact in most behaviors that we have found (except posture as the static channel). For example, the amount of whole body movement (#23, 24), the amount of hand gesture (#28, 29), vocal behaviors (partly via tempo, emphases) and most features of hand gesture.
- *Variety*: The presentations with strong variations significantly increase the attention of the audience. Lacking variation results in monotone (#15), flat face (#22), and hand gesture repeated (#35). In fact, variety can be separated as one single measurement to analyze a presentation. It takes the role as rhythm in music. Even a beautiful piece of music, without changes in rhythm will steadily lose attention from the audience.

3 INTELLIGENT TUTORING SYSTEM

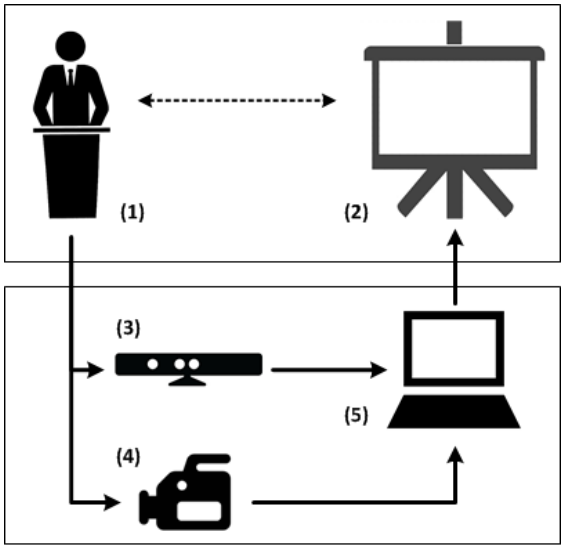

Figure 2: Setup of the system: (1) Presenter, (2) Large screen or projector displays the virtual audience, (3) Microsoft Kinect provides input for processing task, (4) Regular camera records presentations for later playback, (5) Computer handles the analysis and recording tasks.

In order to support presenters with an effective solution that can help them self-practice even at home, we aimed to implement the system with the following functions: (1)Automatic analyzes presenters performance; (2)Provides immediate feedback during the presentation; (3)Provides overall analysis about the whole presentation; (4)Lets users review their performance together with the analyzed results, thus allows them to keep track of their practicing progress. In order to achieve these purposes, we set up a Microsoft Kinect to extract body's skeletal representation as input for the analysis task. In parallel, a regular camera or webcam is positioned to simulate the audiences point of view. The automatic analysis, as well as recording is processed in real-time using a regular PC. The result is visualized on the PC or an external screen/projector (Figure 2). Users also have the chance to review their presentations, together with in-depth analysis about nonverbal cues in the end.

3.1 Recognition of Nonverbal Cues

In order to implement a system that mostly takes visual nonverbal channels into account, together with conclusions from our observation, we currently focused on the four aspects: (1)Eye contact; (2)Posture; (3)Gesture; (4)Whole body movement. For each as-

pect, several related behaviors were recognized, using solely data captured from the Kinect device. They are listed as follows, readers can refer to Table 1 to link to the behavior associated to each number:

- Eye Contact: #18, 19.
- Posture: #2, 3, 4, 7, 8.
- Gesture: #28, 29.
- Movement: #23, 24, 25, 26.

In the next subsections, we will explain the algorithms that were employed to detect these behaviors.

3.1.1 Eye Contact

Eye contact has significant impact on the creation and maintenance the connection between a presenter and the audience. Unfortunately, one common mistake was found in our observation is *eye contact avoidance*, which appeared in 28 over 39 presentations (Table 1, #19). Our observation found that, losing eye contact with the audience is the quickest way to ruin a presentation. Thus, we aimed to recognize the moments that the visual attention of the presenters is shifted away from the audience.

Visual focus of attention has been widely employed for human-computer interaction and usability research (Jacob and Karn, 2003). However, these applications require several constrain, regarding the minimum resolution of input data, or users are required to wear specific devices. Furthermore, in public speaking, eye contact does not really means knowing exactly the eye movements. Instead, it more concerns whether the audience can perceive that they are being addressed. In many applications, instead of localizing gaze exactly, head/face orientation was used as the effective approximation for subjects focus target (Sheikhi and Odobez, 2012; Ba and Odobez, 2009). The experiment in (Stiefelhagen, 2002) also proved that, head orientation was the reliable indication of the visual focus of attention in 89% of the time.

Our system relies on faces 3D orientation, which is provided as part of the Kinect SDK as the measurement for visual attention. In our observation, when not making eye contact, presenters mostly looked up to ceiling or down to floor. When these behaviors appear, the pitch angle of their faces drops to lower/raise to higher than some specific ranges. These ranges can be set manually based on the observation. In our implementation, we set the range of face's pitch angle that according to having good facial orientation is within [-15^0; 10^0]. Being outside this range is assumed as *eye contact avoidance* appears.

3.1.2 Amount of Global Movement and Hand Gesture

Amount of movements and hand gesture can visually indicate the amount of energy that presenters use, also can reveal some hints about their internal states. Our observation results suggested that, this amount should be kept at the appropriate intensity to avoid negative impressions for the audience (Table 1, #23, 24, 28, 29). The next issue is, *how much movements/gesture is suitable?* In order to answer this question, we extracted from the annotation data the durations that movements and hand gesture were annotated as too much/little. Next, the EyesWeb XMI (Camurri et al., 2000) was used on the Kinect-recorded database to extract the skeletal movement in these durations. We used central position of upper body to measure the amount of whole body movement, and total distance of two hands for the amount of hand gesture. Finally, the average results were computed (Table 2).

Table 2: Average velocity of body movements and hand gestures when being too much or little (in meter/s).

Behavior	N	Mean	Sd
Too much movements	42	0.34	0.08
Too little movements	67	0.11	0.03
Too much hand gesture	36	1.87	0.32
Too little hand gesture	53	0.17	0.09

We computed the average velocity of upper body and two hands in a moving time window of 5 seconds. Next, the mean values in Table 2 were used as hard thresholds to detect whether presenters' whole body and hands travelled too much/little.

3.1.3 Direction of Global Movement

Direction is one important characteristic of movement (Table 1, #25, 26). Moving small steps forward brings presenters closer to the audience, thus expresses the willing to make connection. In contrast, small steps backward can be seen in awkward presenters, when they subconsciously retreat from the stage.

We aimed to detect these behaviors via analyzing the trajectories of the skeletal joint *Spine*, being projected to the ground floor. At each frame, we calculated the direction of displacement. This direction was compared with the orientation of the upper body that was determined via shoulders positions. The comparison output the decision whether the presenter shifted backward/forward in this frame. We accumulated the number of these frames in a moving window of 1 second. All of frames inside a moving window is regarded as positive if such window contains more than 80% positive frames.

3.1.4 Body Posture

Posture is the static configuration of the presenters′ body. We focused on the two postural aspects, including foot position and the focus of body weight. Foot position relates to the behaviors #2 and #3 in Table 1. It can be easily measured via Euclidean distance of *Foot_Left* (F_C) and *Foot_Right* (F_R). This number was normalized by the distance of *Shoulder_Left* (S_L) and *Shoulder_Right* (S_R) to eliminate the effect of body size (Equation 1).

$$Ratio_{FootDistance} = \frac{distance(F_L, F_R)}{distance(S_L, S_R)} \quad (1)$$

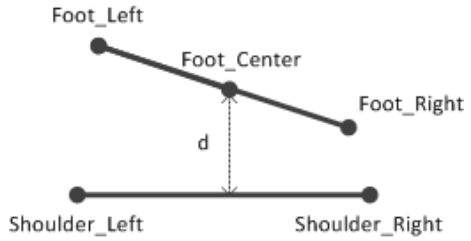

Figure 3: Determine whether the presenter is leaning forward or backward via the distance *d*.

Body weight relates to the behaviors #7, 8 in Table 1. It was measured via the the positions of shoulders, in comparison with foot, being projected to the ground floor (Figure 3). The distance *d* determines the degree of leaning, while the leaning backward or forward of body is determined via whether *Foot_Center* - (F_C) is in the left or right side of the vector created by *Shoulder_Left* and *Shoulder_Right*.

Another behavior relates to body weight is whether a presenter is standing in one leg/feet. This degree of leaning is determined via comparing the distance between *Hip_Center* - (H_C) with *Foot_Left* and *Foot_Right* (Equation 2).

$$Ratio_{Weight} = \frac{distance(H_C, F_L) - distance(H_C, F_R)}{distance(F_L, F_R)} \quad (2)$$

Again, we relied on our Kinect-recorded database to find the thresholds for the above postural metrics. The method similar to the one used in Section 3.1.2 was used, this time for every single frames that the behaviors appeared. The process produced the results as shown in Table 3.

3.1.5 Assessment of the Whole Presentation

After the separated behaviors can be recognized, we aimed to produce the final assessment for the whole presentation. Each presentation is assessed based on the four qualities, accordingly to the four nonverbal visual channels: (1) Posture; (2) Gesture; (3) Global Movement; (4) Eye contact. Additionally, one single quality is measured accordingly to the general performance. This is similar to what learners are given in most training classes about public speaking skills.

Table 3: Average values for the postural metrics.

Postural metric	N	Mean	Sd
Foot distance - closed (ratio)	3982	0.48	0.18
Foot distance - stretch (ratio)	1246	1.23	0.35
Weight forward (meters)	2108	0.31	0.12
Weight backward (meters)	2732	0.09	0.04
Weight on one feet (ratio)	4120	0.88	0.19

Firstly, in order to enrich the database, we asked for the permission to copy the DVDs of presentations from other students who attended the training class (each student is given one DVD that contains all of their presentations during the course in the last session). These video clips are not appear previously in our recorded database. Finally, we achieved extra 37 presentations (without Kinect recordings), in addition to our existing 39 videos that were used for the observation (which contain Kinect recordings). The expert was asked to assess the presentations based on a 4-level scale for each single aspect: (1) Posture; (2) Gesture; (3) Global Movement; (4) Eye contact, plus one score for the overall assessment. This data will be used for the later classification.

Based on the previous recognition process, a feature vector is created for each aspect, in which each component stands for the percentage that one single behavior appeared in the whole presentation. For example, feature vectors that represent hand gesture contain two dimensions for gesturing too much/little. Similarly, the number of dimensions of the feature vectors of movement, posture and eye contact are: 4, 5 and 1, respectively. Additionally, the whole presentations are represented via 12-component vectors, is the serialization of all features.

Amongst them, eye contact is represented via one single number and can be easily classified using hard thresholds. For the rest, feature vectors were used as the input for a multi-class Support Vector Machine system (Duan and Keerthi, 2005), using one-versus-all method and winner-take-all strategy. The algorithm was used to classify four classes, which represented the performance from [1 - Not good] to [4 - Excellent].

We applied 70-30% train/test split on the total of 76 presentations in order to evaluate the classification of the whole presentation, and then calculate the confusion matrix (Table 4). As the result, the system achieved the recognition rate of 73.9%.

Table 4: Confusion matrix for the classification of the whole presentation.

Truth	Classified as			
	1	2	3	4
1 - Not good	**0.826**	0.043	0.087	0.044
2 - Good	0.087	**0.696**	0.174	0.043
3 - Very good	0.087	0.130	**0.652**	0.131
4 - Excellent	0.043	0.043	0.043	**0.783**
	Average sensitivity: **0.739**			

3.1.6 The Two Methods of Giving Feedbacks

The system provides two ways of delivering feedbacks to the audience. The first one shows users their recorded presentation, the appearances of each behaviors (Figure 4) and results on the four nonverbal aspects, plus the overall result. In parallel, with purpose to give presenters the helpful feedbacks, also aim to provide them the experience as presenting for the real audience, we developed a virtual conference room as one method to deliver feedbacks. The environment was built using the Unity3D engine, simulates the classroom that we collected data for observation. Avatars are able to perform several animations that may bring either positive or negative feeling for presenters. These animation clips are sorted based on the increase of negative feeling: (1) Nodding; (2) Sitting still; (3) Sleeping; (4) Yawning.

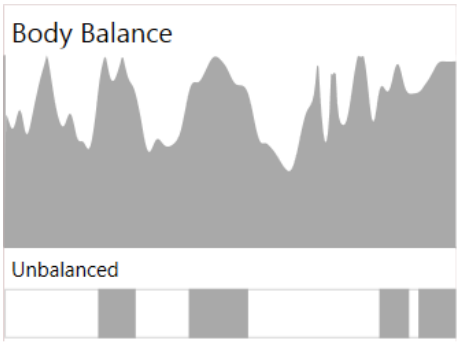

Figure 4: Example of one behavior after being analyzed and displayed to presenters as off-line feedback. The upper graph visualizes the change of the metric used to measure body balance. The lower bar marks the durations that the behavior appeared.

In order to achieve real-time assessment, we applied a moving window to assess the overall quality for the most recent 5 seconds. This process outputted a single value from [1 - Not good] to [4 - Excellent]. The value then was used to manually adjust the distribution of the animation clips. Higher quality presentations resulted in higher proportion of positive animations and vice versa.

Figure 5: The simulated conference room as the method to deliver immediate feedbacks. The avatars can perform different animations based on the performance of presenters.

4 CONCLUSIONS AND FUTURE WORKS

We presented the current progress of our intelligent tutoring system for presenters. Based on the recognition of visual cues from movement, gesture, posture and eye contact, the system can assess the performance of a presentation based on a 4-degree scale, with the accuracy of 73.9%. The behavioral cues and training data were collected from our observation in a class about public speaking skills. Additionally, an adaptive simulated environment was built as one way to give immediate feedbacks, in which avatars can react differently based on the performance of presenters.

The first task to consider in the future development is equipping the system with the ability to recognize more nonverbal cues (as listed in Table 1). The variety of presentations, one of the three most important aspects as explained at the end of Section 2, will be focused. In parallel, we consider the improvement of the classification method, together with collecting data for training/testing. Finally, quality of the simulated conference room will be improved by varying the number of animation, as well as investigate a more affective mechanism to control the avatars.

ACKNOWLEDGEMENTS

This work was supported in part by the Erasmus Mundus Joint Doctorate in Interactive and Cognitive Environments (ICE), which is funded by the Education, Audiovisual and Culture Executive Agency of the European Commission under EMJD ICE FPA 2010-0012.

REFERENCES

Argyle, M., Alkema, F., and Gilmour, R. (1971). The communication of friendly and hostile attitudes by verbal and nonverbal signals. *European Journal of Social Psychology*, 1:385–402.

Ba, S. O. and Odobez, J.-M. (2009). Recognizing visual focus of attention from head pose in natural meetings. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, 39(1):16–33.

Camurri, A., Hashimoto, S., Ricchetti, M., Ricci, A., Suzuki, K., Trocca, R., and Volpe, G. (2000). Eyesweb: Toward gesture and affect recognition in interactive dance and music systems. *Computer Music Journal*, 24(1):57–69.

D'Arcy, J. (1998). Communicating with effective body language. In *Technically Speaking*, chapter 14. Battelle Press.

Duan, K. and Keerthi, S. (2005). Which is the best multiclass SVM method? An empirical study. *Lecture Notes in Computer Science*, 3541:278–285.

Freund, Y., Iyer, R., Schapire, R., and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research*, 4:933–969.

Gao, T., Wu, C., and Aghajan, H. (2009). User-centric speaker report: Ranking-based effectiveness evaluation and feedback. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1004–1011. IEEE.

Hincks, R. and Edlund, J. (2009). Promoting increased pitch variation in oral presentations with transient visual feedback. *Language Learning & Technology*, 13(3):32–50.

Jacob, R. and Karn, K. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Work*, 2(3):573–605.

Kleinke, C. L. (1986). Gaze and eye contact: a research review. *Psychological Bulletin*, 100(1):78–100.

Kurihara, K., Goto, M., and Ogata, J. (2007). Presentation sensei: a presentation training system using speech and image processing. *Proceedings of the 9th international conference on Multimodal interfaces*, pages 358–365.

Nguyen, A., Chen, W., and Rauterberg, G. (2012). Online Feedback System for Public Speakers. In *IEEE Symposium on e-Learning, e-Management and e-Services*, pages 1–5.

Pfister, T. and Robinson, P. (2011). Real-Time Recognition of Affective States from Nonverbal Features of Speech and Its Application for Public Speaking Skill Analysis. *Affective Computing, IEEE Transactions*, pages 1–14.

Picard, R. (2000). *Affective Computing*. The MIT Press, 1st edition.

Rodman, G. and Adler, R. B. (1996). Style: Delivery and Language Choices. In *The New Public Speaker*. Wadsworth Publishing, 1st edition.

Seiler, W. J. and Beall, M. L. (2004). *Communication - Making connections*. Allyn&Bacon.

Sheikhi, S. and Odobez, J. (2012). Recognizing the visual focus of attention for human robot interaction. *Lecture Notes in Computer Science*, 7559:99–112.

Silverstein, D. A., Tong, Z., and Tong Zhang (2003). System and method of providing evaluation feedback to a speaker while giving a real-time oral presentation. *US Patent 7,050,978*.

Stiefelhagen, R. (2002). Tracking focus of attention in meetings. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 273–280. IEEE Comput. Soc.

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759.

Zimmerman, P. and Bolhuis, J. (2009). The Observer XT: A tool for the integration and synchronization of multimodal signals. *Behavior Research Methods*, 41(3):731–735.