Presentation Sensei: A Presentation Training System using Speech and Image Processing

Kazutaka Kurihara
National Institute of Advanced Industrial Science and Technology (AIST)
1-18-13 Sotokanda, Chiyoda-ku, Tokyo, 1010021, Japan
+81-3-5298-4116
k-kurihara@aist.go.jp

Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki, 3058568, Japan
+81-29-861-5898
m.goto@aist.go.jp

Jun Ogata
National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki, 3058568, Japan
+81-29-861-5898
jun-ogata@aist.go.jp

Yosuke Matsusaka
National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki, 3058568, Japan
+81-29-861-5898
yosuke.matsusaka@aist.go.jp

Takeo Igarashi
Department of Computer Science, The University of Tokyo
17-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
+81-3-5841-4109
takeo@acm.org

ABSTRACT

In this paper we present a presentation training system that observes a presentation rehearsal and provides the speaker with recommendations for improving the delivery of the presentation, such as to speak more slowly and to look at the audience. Our system "Presentation Sensei" is equipped with a microphone and camera to analyze a presentation by combining speech and image processing techniques. Based on the results of the analysis, the system gives the speaker instant feedback with respect to the speaking rate, eye contact with the audience, and timing. It also alerts the speaker when some of these indices exceed predefined warning thresholds. After the presentation, the system generates visual summaries of the analysis results for the speaker's self-examinations. Our goal is not to improve the content on a semantic level, but to improve the delivery of it by reducing inappropriate basic behavior patterns. We asked a few test users to try the system and they found it very useful for improving their presentations. We also compared the system's output with the observations of a human evaluator. The result shows that the system successfully detected some inappropriate behavior. The contribution of this work is to introduce a practical recognition-based human training system and to show its feasibility despite the limitations of state-of-the-art speech and video recognition technologies.

Categories and Subject Descriptors

H5.2 **[User Interface]**: *Training, help, and documentation*.

General Terms

Human Factors.

ICMI'07, November 12--15, 2007, Nagoya, Aichi, Japan.

Keywords

Presentation, training, speech processing, image processing, sensei.

1. INTRODUCTION

Presentations play an important role in our society. They are commonly used to communicate the presenter's ideas to many people at once and many useful tools have been developed. However, giving a presentation successfully requires certain skills. One needs to prepare the presentation material carefully and deliver it to the audience appropriately. The necessary skills can be improved through training. Many books and tutorials have been written on this subject.

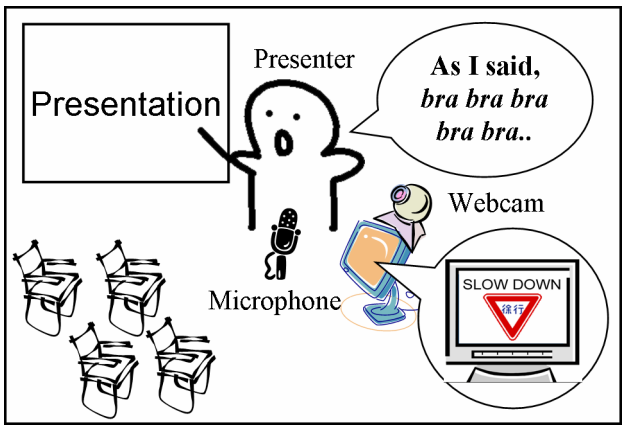

Figure 1. Presentation sensei system.

These books about describe two aspects of the presentation (see e.g. [20]). One aspect is the preparation of the materials in advance. The other is the way the speaker delivers the presentation to the audience. We can relatively easily refine the former by revising the presentation plan and the materials many times by ourselves or by others. On the other hand, it is difficult

to refine the latter. One needs to rehearse a presentation to improve it, but it is difficult for us to revise our own presentations objectively while giving them. One can use video recordings, but it does not give us any suggestions for improvements. In addition, instant feedback during presentations is not available. Consulting human advisors is the best solution, but it is costly and not always available.

In this paper we present a presentation training system that observes a presentation rehearsal and provides recommendations for improving the delivery of the presentation, such as to speak more slowly and to keep eye contact with the audience (Figure 1). We intentionally focus on the basic behavior patterns because high-level semantics is very difficult to analyze by computers in contrast to basic behavior patterns. Our goal is to help people improve their base-line presentation skills by reducing inappropriate behavior such as using many fillers (e.g. "er") or continuously looking down at the script.

In this work, we focus on the following five aspects of presentation delivery.

- The speaking rate should not be too fast.
- The speech should not be monotonous.
- The speech should not contain too many fillers.
- The speaker should look at the audience and avoid continuously looking down at a script or a screen.
- The speaker should finish the presentation within a certain time limit.

We selected these because they are emphasized in existing literature and they can be detected to some extent using current speech processing and image processing technologies. The *Presentation Sensei*[1] system visualizes the analysis result in real time communicating with a presentation tool. It can give the presenter both instant "online" feedback and post hoc "offline" feedback for improvements. The online feedback function shows the analysis result in real-time. When the system detects some inappropriate behavior, it alerts the presenter by showing a visual signal. The offline feedback function shows the visual summaries of the indices for the presenter's self-examinations.

It is often difficult to make speech processing and image processing work robustly in adverse environments. One advantage of our target application domain, personal presentation rehearsal, is that we can relatively freely configure the environment. It is realistic to rehearse in a silent room with no visual obstacles, where these technologies perform the best. This feature makes the Presentation Sensei system a highly practical application even with imperfect recognition technologies.

Our system is unique in that, while general multimodal systems help the user to control computers, it tries to help computers guide humans. This way of using a computer is relatively new. Heer et al. [6] investigated the design guidelines for this sort of systems and also introduced an experimental media capture system that acts as a film director. The contribution of our work is to introduce a practical application based on this approach and to show its feasibility using state-of-the-art speech and video recognition technologies.

[1] The word "Sensei" means a master or a teacher in Japanese.

2. RELATED WORK

In this paper, we report an extension for presentation tools, i.e. supporting rehearsals. Although many papers discuss new tools for editing and running a presentation, few have focused on rehearsals. Microsoft PowerPoint [3] has a rehearsal mode where the user can record the timing of slide transitions. Our work involves richer multimodal analyses using speech and image processing techniques.

There is some prior work using recognition technologies not to control a computer but to observe people and provide feedback. Ikari [15] and TalkMan [4] applied speech recognition technologies to support learning of a foreign language. A commercial video game called Shibaimichi [5] utilizes speech recognition technologies to train a role-playing skill. Heer et al. [6] investigated systems involving system-initiated direction of human action. Based on contextual interviews with a variety of experts in fields involving human-human direction, they constructed design guidelines for such systems. They also introduced a media capture system that acts as a film director. Our work is slightly different in that, while their system directs the user to make new actions, our system makes recommendations to improve the user's behavior.

One important aspect of our work is that we use recognition techniques in such a way that tolerates recognition errors. Traditional approach of using a recognition technique is to translate the user's multimodal behavior to commands a computer [21]. In these cases, recognition errors are basically not allowed. If they occur, the user needs to fix them via some mediation techniques. However, we use recognition techniques only for providing recommendations for the user and some recognition errors are acceptable. In this sense, Hindus et al. [7], Lyons et al. [9], and Kurihara et al. [8] have studied similar unobtrusive usages of recognition technologies. These systems also observe the user's natural behavior and provide support when possible.

The importance of non-verbal communications is frequently emphasized in empirical studies. Mehrabian [10] reported that only 7% is contributed by the verbal information when people try to convey their emotions, whereas the residual 93% is done by the non-verbal information such as voices or facial expressions. Based on this report [10], books of presentation methodologies [20, 19] describe the importance of utilizing such non-verbal information intentionally and suppressing unintended non-verbal information conveyed. One of our objectives is to help people improve these non-verbal communication skills.

Goto [17] proposed a series of voice interfaces that utilize non-verbal information obtained by speech recognition technologies. We apply the technologies introduced in [17] to detect non-verbal information in presentations.

3. EVALUATION INDICES OF PRESENTATIONS

Books on presentation methodologies (see e.g. [20, 19]) list several indices for evaluating presentations, such as the speaking rate, the tone of the speech, the frequency of fillers, the eye contact level with the audience, and the timing. We consulted the literature to determine thresholds for inadequate states for each index.

During presentations, people should try to speak slightly slower because they tend to speak faster than usual [20]. In addition, monotone voices might lose the attention of the audience [20]. To help presenters improve their skills, we decided to alert them when their current speaking rate[2] exceeds a predefined threshold, and when the standard deviations of their pitches (F0) fell below a predefined threshold. Kori [11] reported the analyses of various voice data gathered through decades using the speeds of speech, the pitches (F0), and their standard deviations. Referring to this, we set the criterion for the speeds of speech as 7.6 mora/sec, and for the standard deviations of the pitches as 10Hz (targeted for male users). The frequent fillers are annoying for the audience [20]. We therefore decided to alert the presenters when they make filled pauses, a kind of fillers, such as "er," or "um."

Besides the audio indices, we also list visual indices of presentations. One is the amount of eye contact with the audience. In [20] the eye contact ratio is defined by dividing the time that the presenter looks at the audience by the total time. According to [20], when the eye contact ratio is less than 15%, the audience perceives negative impressions such as apathetic, apologetic, and premature. On the other hand, it says that when the eye contact ratio is more than 80%, the audience perceives positive impressions such as confident, honest, affinitive, and practiced. Referring to this, we decided to give an alert when the eye contact ratio fells below 15%.

Lastly, it is important to be punctual in presentations. We decided to alert the presenter when 80% and 100% of the time limit are exceeded.

4. THE PRESENTATION SENSEI SYSTEM

We developed a prototype of a presentation training system "Presentation Sensei" based on the evaluation indices introduced in the previous section.

4.1 System Configuration

The Presentation Sensei system consists of several modules connected by a network (Figure 2). The audio analysis module continuously analyses the input signal from a microphone and provides the integration module with the results of the utterance duration detection, pitch (F0) detection, and filled pause detection. The speech recognition module also continuously provides the integration module with the mora-based speech recognition results. The image processing module continuously analyzes the input from a webcam and provides the integration module with the result of the face position/orientation detection. The integration module integrates all the provided information and gives feedback to the presenter using various monitors. These modules can be distributed over a local network for the load sharing purpose and they communicate with each other via RVCP protocol [17]. The system can also be connected to a third party presentation tool to achieve synchronization. We currently connect our system to an in-house presentation program to receive timing information and thumbnail images of slides.

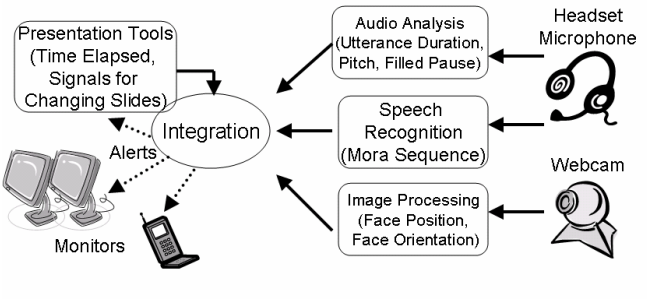

Figure 2. System configuration of presentation sensei.

4.2 User Interface

The Presentation Sensei system provides a presenter with feedback based on the indices of presentations accumulated in the integration module. We developed two kinds of feedback: instant "online" feedback and post hoc "offline" feedback. The online and offline feedback complement each other and users choose which to use to satisfy their needs.

Figure 3. Online feedback. (Left) Real time monitor. (Trafic signals) Visual Alerts.

4.2.1 Online Feedback

The online feedback provides the presenter with short term statistics of the indices during a presentation. "Real time monitor (Figure 3 left)" indicates the indices as they are. They are used by the presenter to check the recent status visually. "Alerts" are the notifications of 6 kinds of information shown in Figure 3 right. The alerts are visualized in two ways: one is with a pop up window rising from the bottom right of the displays for a few seconds, and the other is with a full-screen sized window. The presenter can choose where to show these real time monitors and alerts among the multiple monitors (the main screen, the presenter's laptop's, and sub-displays[3]) furnished freely in the

[2] In our system, we define the speaking rate as the number of moras per second in the Japanese language. "Mora" measures a phonological unit. It is intuitive for the Japanese language because the Japanese language is one of the mora-based languages. In contrast, the English language is one of the syllable-based languages.

[3] To have a good eye contact ratio, it is effective to show the visual feedback on the sub-displays furnished in the direction of the audience.

presentation rehearsal environment. The presenter can also use other modalities such as sounds and vibrations[4] to receive alerts.

The online feedback can distract the user during operation. However, this is natural, as when somebody is practicing presentations in front of a human teacher (*i.e.*, a trainer), instructions from the teacher sometimes need to catch the trainee's attention, otherwise, the teaching would not be effective.

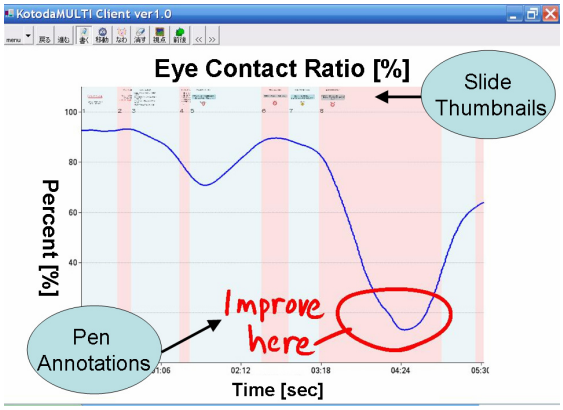

Figure 4. Offline feedback. An example of the generated charts. The presenter can annotate them with a pen.

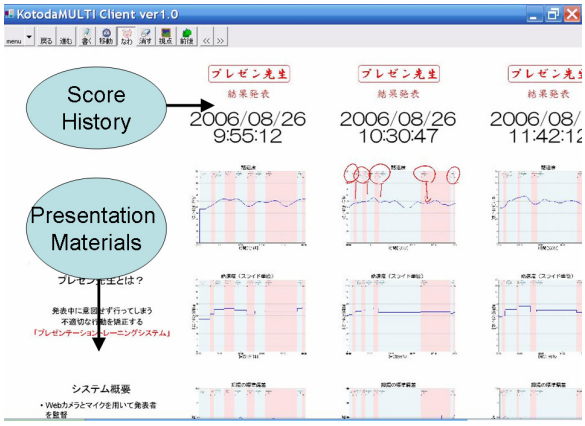

Figure 5. Offline feedback. The rehearsal histories are recorded in a plane with the presentation materials.

4.2.2 Offline Feedback

The offline feedback provides the presenter with statistics of the whole presentation using charts (Figure 4). The time series data of the indices are plotted with the thumbnail images of the slides. For each index, the system generates both a chart of the raw data and a chart of the average values for each slide. The latter allows the presenter to review the presentation from the macro level view and to identify parts to refine. The former allows the presenter to review the presentation from the micro level view and to analyze specific parts in detail.

[4] The current prototype uses a cell phone as a vibrating machine.

The various charts generated by the system are accumulated in a single file together with presentation slides provided by the presentation tool ([3] or [14]). After a presentation, the Presentation Sensei system pastes the charts with a time stamp on the right side of the presentation materials, allowing the presenter to add annotations on them with a stylus pen.

Each presentation rehearsal generates a series of charts pasted in the vertical direction. They can be interpreted as a history of the rehearsals. After several trials, the presenter can compare the performances by scanning the charts in the horizontal direction (Figure 5).

In the current prototype, the offline feedback only provides the presenter with the charts. One future direction of this research is to provide the presenter with advice to improve the presentation by comparing the indices with a data base obtained by a large number of presentation recordings. It also might be useful to provide advice in natural languages.

4.3 Audio Analysis (Utterance Detection, Pitch, Filled Pause)

The audio analysis module detects the duration of utterances, the pitch (F0), and the filled pauses from the presenter's voice input using a microphone for every 10 milliseconds. These data are continuously sent to the integration module. The utterance duration is detected by finding durations where the power of the voice is high. We use the F0 detection and the filled pause detection as proposed in [16]. One advantage of this method is that it works robustly against background noise. It detects the filled pauses with a bottom up signal processing of using two features of filled pauses: small F0 fluctuations, and small deformations of the spectrum envelope. It can detect any long utterance of a vowel regardless of the language spoken.

4.4 Speech Recognition (Speaking Rate)

The speech recognition module executes a mora-based speech recognition from the presenter's voice input using a microphone. The recognition results (a series of moras) and their corresponding utterance durations are sent to the integration module. The speaking rate excluding silence is calculated by dividing the number of moras by the duration.

We use *julian* [2] as the speech recognition engine. We extended it to send the output continuously to the integration module [18]. As a language model, we used a network grammar that allows arbitrary transitions between 121 kinds of moras in the Japanese language including silence.

The speech recognition module takes a few seconds to return recognition result but this performance is good enough for our purposes. Currently our speech recognition module only supports the Japanese language. However, it is easily applicable to other languages such as the English language as long as a speech recognition engine for the language is available and the speaking rate in the language is defined.

4.5 Image Processing (Face Position, Face Orientation)

The image processing module detects the position and the orientation of the presenter's face from the input using a webcam. These data are continuously sent to the integration module. For this process, we implemented two methods for the current

prototype. One is to use a visual marker and the other is to use a pure image processing technique. We currently use AR toolkit [1] for marker based recognition, and sub-space method and SVR (Support Vector Regression) [12] for marker-less (pure image) recognition. Each of the two methods can detect 6 degrees-of-freedom information of the position and the orientation of the presenter's face in real-time using a standard single eye webcam.

4.5.1 AR toolkit method

In the AR toolkit method, the presenter wears a special visual marker on the head (Figure 6). The marker is made of a foam polystyrene cube with 2D codes on each surface to avoid the occlusion problem. One advantage of this method is that it can be applied to new users without pre-training.

Figure 6. The visual marker for the AR toolkit method.

4.5.2 Sub-space and SVR Method

In the sub-space and SVR method, the system needs a series of image data capturing any angles and poses of the presenter's head in advance. The system applies the principal component analysis on the data and obtains the eigenvectors for each attitude of the face. Using the eigenvectors as the models, the system detects the best-fitted model on input images. It enables an adaptive tracking of the position of the presenter's face in various poses. In addition, the system applies a face orientation detection based on a SVR to the detected face region on the input image. One advantage of this method is that the presenter does not have to wear any visual markers.

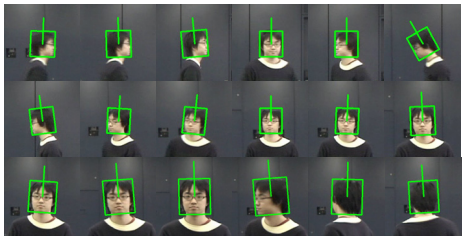

Figure 7. The sub-space and SVR method. It can detect and track the face position and the face orientation from any 360 degree views of 2D images.

4.5.3 Definition of "looking at the Audience"

The 6 degrees-of-freedom information about the presenter's face is converted into a binary value whether the presenter is looking at the audience or not. In our current prototype, we assume a rehearsal environment illustrated in Figure 8. The current simple algorithm outputs "looking" signals when the horizontal angle of the presenter's face is in a region shown in Figure 8, otherwise outputs "not looking" signals. In the future we will implement a calibration function to define the rehearsal environment in advance to achieve more precise and flexible detections.

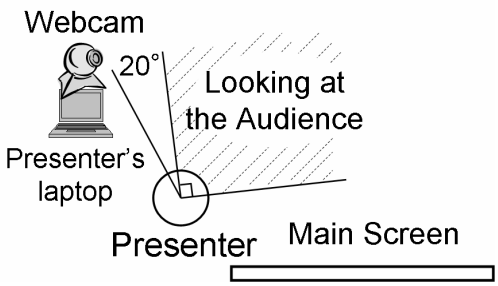

Figure 8. Assumption of presentation rehearsal environment. The presenter is on the right side of the main screen. The webcam is on the presenter's laptop.

5. USER STUDY

We conducted a small user study to evaluate our prototype system and to obtain early feedback from the users.

5.1 Method

Three male graduate school students A, B, and C participated in this study. They have given many presentations and all of them were interested in improving their presentation skills. The task was to give a presentation for eight minutes using our system[5] without the audience to simulate our target situation: a self-training in an empty room. They used their own presentation materials. We used the AR toolkit method for the image processing method, and gave the online feedback with the visual modality only. After each presentation, we let the participants review their presentations with the offline feedback. Then we collected their feedback on the overall impression of the system and suggestions for improvement by means of a questionnaire and interview.

In addition, we tried to evaluate the accuracy of the recognition techniques by comparing system output with observations by a human evaluator. We video-recorded the participants' presentations from outside of the system. A third party volunteer person manually analyzed the videos. The human evaluator counted the numbers of filled pauses and calculated the eye contact ratios. We also asked her for subjective impressions with respect to speaking rates and tones of speech.

5.2 Results

5.2.1 Effectiveness and Suggestions for Improvements

All the three participants answered that the system was useful because of the instant feedback. The system successfully notified unintended behavior to the presenters, which were difficult to be noticed by themselves. On the other hand, they commented that they wanted to customize the criteria of the alerts by themselves. We also obtained a suggestion that it is effective to alert the presenter when the time exceeds predefined durations for each slide rather than only alerting with respect to the overall duration.

[5] A function to import a Power Point file was used.

5.2.2 Comparing the System Output vs. the Human Observation

Table 1 shows the number of filled pauses detected by the system and the human evaluator. The system successfully created alerts when the speech contained too many filled pauses[6]. In addition, the system did not erroneously alert participants when they did not make filled pauses. Participant B, who unintentionally had the greatest number of filled pauses, commented that he would have preferred to be notified of all possible filled pauses during training, even at the risk of increasing recognition errors by the system (*i.e.*, failure to detect pauses, or false positives). The results were attributed to the current conservative setting, which is used to avoid recognition errors that might occur when a presenter utters long vowels, such as "volunteer".

Table 1. Number of fillers detected by the system and a human evaluator.

Participant ID	Total (counted by a person)	Detected successfully by the system	Failed to detect	False positives
A	18	1	17	0
B	73	12	61	0
C	1	0	1	0

Figure 9 is supplemental data showing our baseline results for the precision and recall rates obtained by applying our algorithm to a Japanese spontaneous speech corpus [13] while changing the parameter Th_{fp}. Although a simple comparison between the results in Figure 9 and Table 1 is difficult because the recording environments differed, the results do indicate that precision and recall can be adjusted to some extent by changing the parameter Th_{fp} to meet users' requests to be conservative or aggressive.

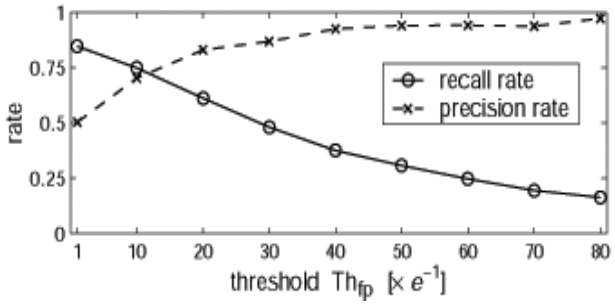

Figure 9. Precision and recall rates with changing Th_{fp} (the time-threshold parameter for filled pauses).

[6] Note that it was not our original goal to report all possible filled pauses.

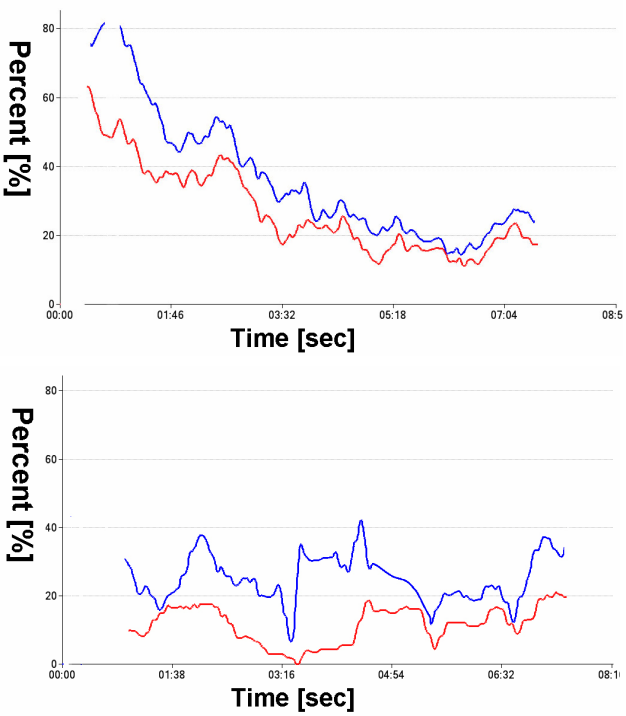

Figure 10. Comparison of eye contact ratios (blue: the system output, red: the human observation). (Top) Obtained a good match on Participant C. (Bottom) Obtained a not good match on Participant B.

Figure 10 shows the comparisons of the eye contact ratios obtained by the system and by the human evaluator. They roughly matched in the cases of Participant A and C. These two participants commented that these deviations were acceptable because they could still observe general tendencies of the index. On the other hand, Participant B had a habit to look at a direction between the audience and the main screen, where we had set the boundary of "looking at the audience" and "not looking at the audience." It caused a large deviation between the system output and the human observation. We hope that this can be easily fixed by allowing the user to customize the threshold.

Table 2 shows the comparison between the volunteer's subjective impressions and the frequencies of the system's alerts, dealing with the speaking rates and the tones of speech. The correlation in terms of the speaking rate was high. On the other hand, the correlation in terms of the tone of speech was low because the criterion for the alert was too low. We analyzed the F0 standard deviations of all the participants through the presentations and found that the system could have alerted the presenters correctly if we had set the threshold between 20Hz and 40Hz. Although currently we set the threshold by the absolute frequency (Hz), in the next version we will introduce logarithmic frequencies to apply the system to users with wide voice ranges, including females.

In this user study, the thresholds for the alerts were conservative in general. However, the participants commented that the fact that the system was continuously supervising them itself helps them avoid inappropriate behavior.

Table 2. Comparison of speaking rate and tone of speech.

Participant ID	Speaking Rate		Tone of Speech	
	Human Evaluation	Alert Frequency	Human Evaluation	Alert Frequency
A	Too Fast	Often	Vivid	None
B	Too Fast	Often	Vivid	None
C	Adequate	None	Monotone	None

6. LESSONS LEARNED

In this section, we summarize the important lessons learned from our prototype implementation and the results of the user study. We hope that these lessons are useful for others building similar systems.

6.1 Decomposing Human behaviors into Elements

Heer et al. point out that when a system captures human behavior for analysis, it is important to decompose the complex human behavior into less complex sub-components that could be recognized by existing technologies. Accordingly, we decompose low-level human behavior into several indices and process them independently with corresponding recognition technologies. This is possible because these indices are almost mutually independent in presentations, as well as independent from the higher-level semantic content. Our experience shows that such decomposition makes it easier for the user to understand the system's behavior and tolerate simple errors. Without proper decomposition, it is difficult for the user to understand what is happening and small errors can make the entire system unusable.

6.2 Setting Recognition-Friendly Environments

It is also important that the user can configure places and devices to get the most out of the existing recognition technologies. Presentation rehearsal is an ideal application in that sense. Another interesting observation is that it is acceptable to ask the users to adjust their behavior to be recognized correctly in our target application. For example, we can construct an acoustic model of a speech recognition engine based on sound data obtained by professional speakers' "good" pronunciations. We then can encourage the users to make their behaviors recognized correctly as part of the training [15]. This is somewhat different from standard use of recognition technique, where the user should be able to speak freely. These factors significantly contribute to the successful use of imperfect recognition techniques in our application.

6.3 Informed Consent of the User

Our Presentation Sensei system explores a unique field of interaction design: supporting self-motivated peoples' self-improvements. We learned that to make a system support self-motivated users' self-improvements effectively, it is important that the user clearly understands the system's (imperfect) behavior and can consent to work with it. This can be even more important than simply improving the recognition rates. With enough information on what kind of errors the system can make, the self-motivated users can make sensible decision on what to do with the system's possible wrong feedback.

As observed in the user study, self-motivated users are eager to configure the system parameters determining what to display and when to alert. It gives the users the impression that the system is under their control. We infer that providing interfaces to configure systems in detail is a possible way to obtain informed consent (i.e. they can feel that the system is under control) of these self-motivated users.

One interesting comment obtained in the user study is that a participant preferred to be notified of more filled pauses for his training even at a risk of increasing recognition errors. His comment implies a possibility that a system might not need to try to return perfect recognition results as long as it obtains informed consent of the users. Namely, the users can simply "accept" the situation where the recognition technologies do not perform perfectly as long as they can obtain general tendencies of their performances.

6.4 Designing the System Impression

The concept of "system impressions" is a system design strategy proposed in [6] stating that systems should adopt the appropriate tone and role for the context of interaction. Many existing tutorial systems involving recognition technologies introduce a human-like agent as its central interface. One advantage of this is that it can give a friendly impression to the user and can help to maintain the users' motivations. Another advantage is that it can give a plausible argument that unexpected behaviors of the agent (actually caused by misrecognitions) are caused by a fancy of the intellectual existence.

On the other hand, as discussed, self-motivated users are eager to configure the system. This means that they want the system not to teach them, but to act as a reliable tool working as they have configured it. This is an important clue to design the system impression. We therefore designed the system interface not as a fancy human-like agent, but as a conscientious tool so that the users may understand the state of system instantly. This design concept can be achieved by keeping the interface design simple and honest with respect to recognition errors. Our prototype system achieves this goal by introducing simple indicators (or meters) and traffic signals as the design motif. Positive feedback from test users indicates the success of this design.

7. CONCLUSION AND FUTURE WORK

We present a presentation training system that observes the user's basic behavior patterns during a presentation rehearsal and provides feedback for reducing inappropriate behavior, such as to speak too fast and to continuously look down at a script. A small user study demonstrated the effectiveness of our prototype and we obtained the users' suggestion for improvements. Our experience with this system gives us important design implications for building practical recognition-based human training systems.

There are five directions that we want to explore in the future. First, although we focused on building an environment in which users could confidently engage in self-training, we also need to evaluate how well the system improves the quality of users' presentations to real audiences as a result of online feedback as compared to offline feedback.

Second, the offline feedback could be improved by providing better recognition results refined by global-level analysis throughout a presentation, as compared to the instant analysis of online feedback (local-level).

Third, it is important to determine the most effective way to provide a presenter with online feedback from among various modalities and intensities.

Fourth, it is also important to establish a way for users to configure the thresholds for alerts. The thresholds in the current system were based on books on presentation methodologies. The first step is to allow users to configure the thresholds manually. Then, we will obtain a database of the various configurations required by different people, who have their own preferences, for presentation rehearsals. The database could be used to customize the criteria automatically for new users who input their preferences in advance.

Lastly, another promising idea is to introduce new indices of presentations, such as the stability of the presenter's position and face movement, gestures, or facial expressions, where the presenter is looking at the presentation materials, and prosodic analyses of the speech [22].

8. ACKNOWLEDGMENTS

This research was partially supported by Microsoft Institute for Academic Research Collaboration (IJARC) Core2 Project and grant from MSRA Mobile Computing in Education Theme program.

9. REFERENCES

1. AR-toolkit. http://www.hitl.washington.edu/artoolkit/
2. julian. http://julius.sourceforge.jp
3. PowerPoint. http://www.microsoft.com/office/powerpoint/prodinfo/
4. TalkMan. http://www.jp.playstation.com/scej/title/talkman/
5. Shibaimichi. http://www.jp.playstation.com/scej/title/shibaimichi/index.html
6. Heer et al. Presiding Over Accidents: System Mediation of Human Action. *In CHI'04*, pp.463.470, 2004.
7. Hindus et al. Ubiquitous Audio: Capturing Spontaneous Collaboration. *In CSCW'02*, pp.210.217, 1992.
8. Kurihara et al. Speech Pen: Predictive Handwriting based on Ambient Multimodal Recognition. *In CHI'06*, pp. 851.860, 2006.
9. Lyons et al. Augmenting Conversations Using Dual-Purpose Speech. *In UIST'02*, pp. 237.246, 2004.
10. A. Mehrabian. Silent messages, Implicit Communication of Emotions and Attitudes. 2nd Ed., *Wadsworth Pub. Co.*, 1981.
11. S. Kori. The Acoustic Characteristics of Styles Seen in Announcements and Narrations. *In 16th Conference of Acoustic Society Japan*, pp.151-156, 2002, in Japanese.
12. Y. Matsusaka. 2D Omni Directional Head and Head-Parts Tracking Technique Using Subspace Method and SVM. *IEICE Technical Report PRMU*, Vol.106, no.72, pp.19-24, 2006, in Japanese.
13. Itou et al. A Japanese Spontaneous Speech Corpus Collected using Automatically Inferencing Wizard of OZ System. *J. Acoust. Soc. Jpn. (E)*, Vol. 20, No. 3, 1999.
14. [anonimized]
15. Ikari et al. English CALL System with Functions of Speech Segmentation and Pronunciation Evaluation Using Speech Recognition Technology. *In ICSLP'2002*, pp.1229-1232, 2002.
16. Goto et al. A Real-time Filled Pause Detection System for Spontaneous Speech Recognition. *In Eurospeech '99*, pp.227-230, 1999.
17. Goto et al. Speech Completion: New Speech Interface with On-demand Completion Assistance, *In HCI International 2001*, Vol. 1, pp.198-202, 2001.
18. Kitayama et al. "SWITCH" on Speech. *IPSJ SIG Technical Report*, SLP-46-12, Vol.2003, No.58, pp.67-72, 2003, in Japanese.
19. I. Takeuchi. More Than 90% is Judged by Your Look. *Shincho-sha Pub. Co.*, ISBN: 4106101378, 2005 in Japanese.
20. H. Yahata. Perfect Presentation. *Seisansei Shuppan Pub. Co.*, ISBN: 4820115634, 1998 in Japanese.
21. Oviatt et al., Individual Differences in Multimodal Integration Patterns: What Are They And Why Do They Exist?. *In CHI'05*, pp.241-249, 2005.
22. Rosenberg et al. Acoustic/Prosodic Correlates of Charismatic Speech. *In Eurospeech'05*, pp.513-516, 2005.