SmartVoice: A Presentation Support System for Overcoming the Language Barrier

Xiang Li
Interfaculty Initiative in Information Studies,
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-033, Japan
kular.me@me.com

Jun Rekimoto
Interfaculty Initiative in Information Studies,
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-033, Japan
Sony Computer Laboratories, Inc
3-14-13 Higashigotanda, Shinagawa-ku,
Tokyo 141-0022, Japan
rekimoto@acm.org

ABSTRACT
In most cases, speeches or presentations at an international event are required to be given in a common language (e.g. English). However, for people who are not proficient in that common language, delivering presentations fluently is very difficult. Simultaneous translation seems to be a solution, but besides its high cost, simultaneous translation undermines the nature of the presentation by substituting the real voice of the lecturer as well as his/her emotions. In this paper, we propose "SmartVoice", a presentation support system, which aims to overcome language barriers. By tracking the lip motion of the lecturer, SmartVoice controls the playback of the narration, which is a sound data prepared in advance or created automatically using a voice synthesizer. SmartVoice also controls the intonation of the sound based on the position and shape of the lecturer's mouth. As the lecturer can talk at his/her own pace with the voice automatically following, it appears as if he/she talks in his/her own voice. In our user evaluation, we confirmed that audiences find it difficult to distinguish between the narration generated by SmartVoice and that by a real voice. We also discuss the possibility of applying SmartVoice to fields other than multi-language presentation support, such as Automated Dialogue Replacement and language study.

Author Keywords
User Interface; Lip Sync; Presentation Support; Face Tracking; Facial Actions; Language Barrier.

ACM Classification Keywords
H.5.m. Information interfaces and presentation: [HCI]

INTRODUCTION
In events such as international conferences, lecturers are usually asked to use a common language (e.g., English). However, this is difficult for a person who has a different native language. Even though reading a manuscript is permitted in some cases, significant effort is still needed to pronounce words correctly and speak fluently. There is also the need to deliver the presentation in other languages when the majority of the audience does not understand the common language. For example, a presentation given in English in Asian countries may not be well understood by the audience with no translation. In both cases, the language barrier is a serious problem for presentations.

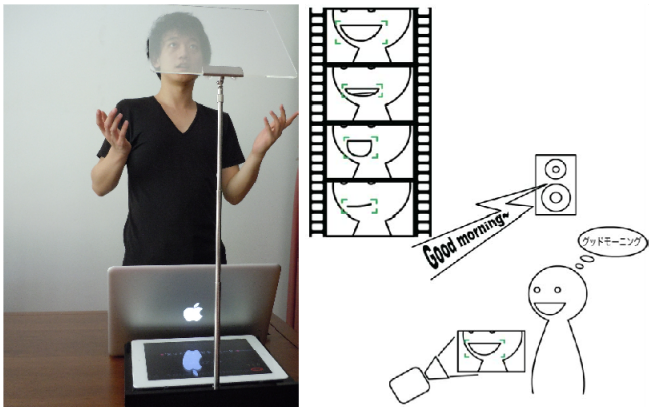

Figure 1. Presentation using SmartVoice

Simultaneous translation is one possible solution for such situations in which a lecturer and audience do not share the same native language. However, in addition to the considerable cost of hiring interpreters, there will be a delay between the voice of the lecturer and that of the simultaneous translation, which not only sounds unnatural but also fails to convey non-linguistic information such as the tempo, pauses, stresses, and intonation of the speech.

It might also be possible for the lecturer to perform a lip sync presentation if he/she records the narration of the manuscript in advance. However, this means that the lecturer will have to practice in order to be familiar with the speed of playback, and when the presentation begins, the lecturer could easily lose control of it.

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.

http://dx.doi.org/10.1145/2556288.2557161

With the help of a voice synthesizer, text in any language can be converted to sound data. Although a voice synthesizer can pronounce each word correctly, the narration of long sentences sounds artificial and robotic. A human lecturer will make a pause or speak louder to emphasize that something important is about to be mentioned, but a voice synthesizer speaks with the same tone consistently all the time. Due to the lack of intonation and emotional expression, voice synthesizers are still not a suitable substitute for a human narrator.

As a solution to the problems above, this paper proposes a support system based on a method with the name "smart lip-syncing." A smart lip-syncing system automatically controls the playback of sound data based on the motions of the lecturer's lip. The sound data can be either a narrated recording or created by a voice synthesizer. With this method, the lecturer can take full control of the presentation; he/she can pause or change the speech speed during the presentation to emphasize important points or gain the attention of the audience. In addition to the playback control, the change in volume as well as intonation can also be controlled in realtime by natural facial gestures. As a result, the audience recognizes that the lecturer standing right before them is delivering a talk with his/her own voice. We call the system based on this smart lip-syncing method "SmartVoice."

To give a better explanation of SmartVoice, we compare it with three known concepts, which is looping, pre-scoring and lip sync (see Figure 2).

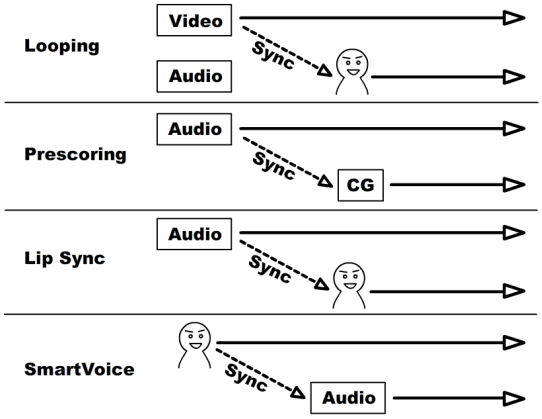

Figure 2. Comparison among Looping, Pre-scoring, Lip sync and SmartVoice

In the case of looping, video is recorded previously with no sound. Then, dubbing artists will narrate the script while watching the video in order to match their voice to the lip motion of the characters in the video.

Contrary to looping, in pre-scoring, which is widely used in American cartoon production, sounds are recorded in advance. Then, graphics are created according to the timing of the sound. Lip sync can be considered as a variety of pre-scoring if the graphics are substituted with a real person.

In all the cases discussed above, humans try to adjust themselves to either un-editable video or un-editable audio. However, in SmartVoice, although the audio file is created beforehand, its playback fits the pace of the lecturer giving the presentation, so that it is the human being that takes all the control in SmartVoice.

PRESENTATION USING SMARTVOICE

To give a presentation with SmartVoice, a user needs to prepare a text file with the presentation's manuscript and then create the sound data of the same manuscript. As mentioned above, the sound data can be a recording of the narration by the user himself or another narrator, or be generated by a voice synthesizer.

After the user feeds both the manuscript text file (.txt) and the sound file into the system, SmartVoice semi-automatically matches the text and the sound on a phrase-to-phrase basis. Phrases in the manuscript that are correctly matched to the sound will then be displayed on the screen of a half-mirror prompter.

Ensuring that SmartVoice is successfully tracking the user's face, he/she can control the sound playback of each phrase shown on the prompter by lip sync. If the user closes his/her mouth after reading one phrase, the playback will pause and that phrase will slide up on the prompter while the next phrase is emphasized. By opening his/her mouth again, the user can resume playback with the corresponding sound. Between every phrase, a user can pause as long as he/she likes. In addition, playback speed and intonation can be controlled in realtime based on the user's lip sync speed and the position of his/her mouth and eyebrows, respectively.

SMART LIP-SYNCING IN SMART VOICE

By mapping the playback control, pitch, speed and volume to the lecturer's lip motions and other facial actions in realtime, SmartVoice sends out the sound correctly and properly at the lecturer's pace of lip sync. Figure 1 shows a scene of giving a presentation with SmartVoice.

Playback Control

We introduce a prompter to our system to act as a SmartVoice interface for the user. With the prompter showing the phrase to read, the user can always confirm the progress of the presentation. To achieve this, we need to correctly match the manuscript with the sound. In SmartVoice, this matching is semi-automatic.

Generally, a lecturer will pause during a presentation to maintain the correct tempo. In order to ensure that the audience will not feel confused with incongruous pauses during the presentation, our system first extracts phrases from both the manuscript text and the sound data and then matches them. For the manuscript text, we extract phrases that are separated by commas and periods. For the sound data, we extract phrases separated by silent sections, which are characterized by low volume levels (see Figure 3).

To be specific, the silent section is determined as the region whose audio levels is less than 10% of the average audio level on the entire sound data and last at least 100 milliseconds. The results will be cached in an array of raw silent sections. Noise or breath sound, usually consisting of very short bursts of high volume, may exist in the sound data. In order not to regard noises as silent sections, we add an algorithm, which loops the array of the raw silent sections to find pairs of raw silent sections whose distance is less than 200 milliseconds, and then appends them to a single silent section, to identify and ignore noise or breath sound to guarantee that there are no silent sections separated by noises.

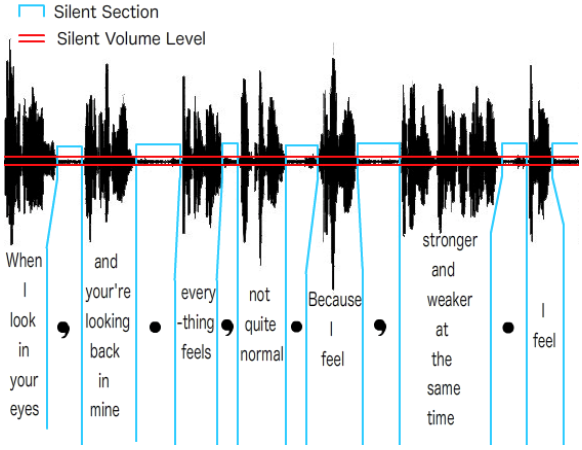

Figure 3. Phrase extraction from sound and text

With the phrases extracted from both the manuscript text and the sound data, our system semi-automatically matches them using a process we explain in detail below. Since the phrases extracted from both sides are matched on a phrase-to-phrase basis, SmartVoice guarantees that each phrase shown on the prompter is correctly matched to its sound and that after one phrase is read, the next phrase slides up and is ready with its corresponding sound so that the correct phrases in the sound data will be played in time with the user's lip sync.

We intend not to match the lip motion to exactly at each syllable, because not all syllables or words can be detected from the lip motion. Instead, the user can merely start a phrase by opening his/her mouth to indicate the system to start or resume. As long as the playback does not encounter a silent section with the user's mouth closed, SmartVoice will not take into account whether the lip motion/shape matches the syllable or word being played within a phrase.

Pitch, Speed & Volume Control

With SmartVoice, not only the playback but also the pitch, speed, and volume can be controlled by the user's facial gestures in realtime. This results in a more natural presentation experience, although the manuscript text is narrated by a voice synthesizer.

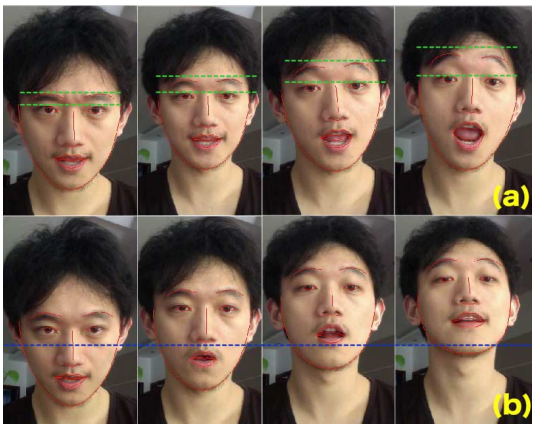

Figure 4. (a) Volume control (b) Pitch shifting control

We take into consideration the physical constitution of human beings when speaking and developed a method for controlling pitch, speed, and volume. Pitch is shifted up or down with the position of the user's head (see Figure 4(b)). If the user performs a rapid lip sync, the playback speed will be accelerated and vice versa. We attempted to control the volume by the openness of the eyes, but we eventually changed the controller to follow the height of the eyebrows (see Figure 4(a)). We return to this issue later.

SYSTEM ARCHITECTURE

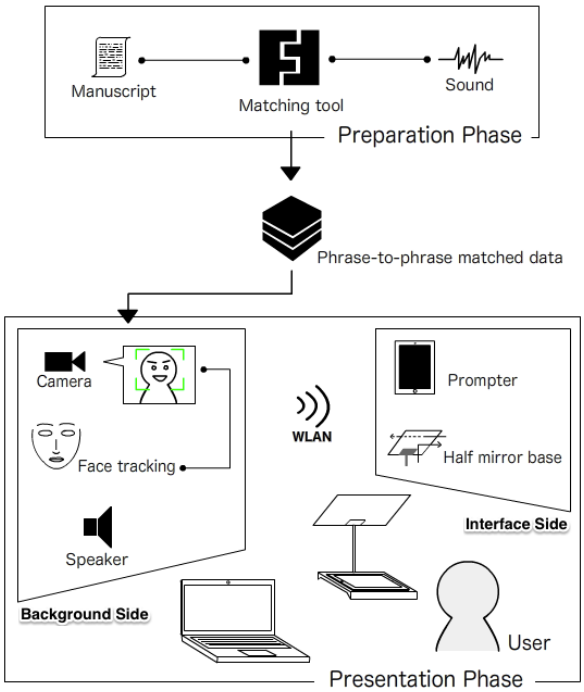

Figure 5. System architecture

In this section, we present details of the prototype of our presentation support system.

There are two main parts to our system. One is on the background side and the other is on the interface side. The

background side consists of a camera, which is used for detecting and tracking the user's face, a speaker, and a computer used for processing. The interface side basically consists of a prompter containing a tablet PC and a prompter base. The two sides are connected via wireless LAN (see Figure 5).

Background side

Since we want to achieve a natural presentation experience and at the same time free the user from complicated operations, everything is completely controlled only by the user's lip motions and facial actions. Not only can the user determine the pauses and resumption of the playback, but also can make the sound generated by a voice synthesizer more vivid by altering the volume, pitch and speed in realtime.

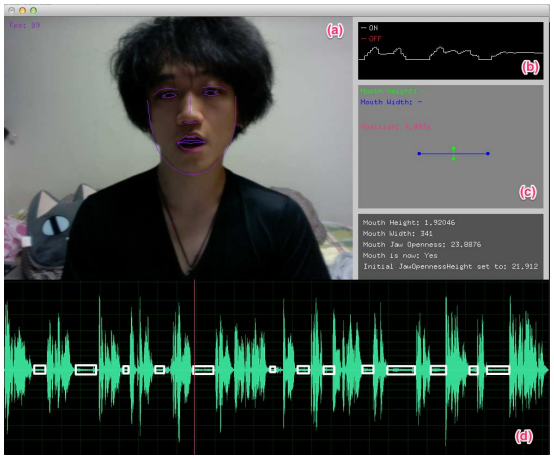

Figure 6. Background side program. (a) Real-time frame captured with camera with ofxFaceTracker markers as annotation, (b) mouth openness pattern, (c) indicator of the mouth height and width, and (d) sound waveform with silent sections shown as white rectangles

Face Detecting & Tracking

When in use, SmartVoice tracks the user's face position and lip motions. This is achieved by adopting the openframeworks add-on "ofxFaceTracker" for our system. Openframeworks is a C++ based multi-platform open source tool kit [1]. OfxFaceTracker is one of the many add-ons of openframeworks and has features to detect the position, inclination, and shape of a human face as well as all the parts on the face, such as mouth, eyes, nose, and eyebrows [2]. C++ ensures the high performance of our system because any delay will sound unnatural and may confuse the audiences during a presentation. Compared to other computer vision implementation for face tracking, ofxFaceTracker is easy to attach to our system because it is well moduled as an add-on with outstanding performance on tracking the human face as well as facial parts (The empty sample of ofxFaceTracker runs at over 100fps on a laptop). In our prototype, we track the position and shape of the user's face and other facial parts in the video captured by a laptop with a built-in web camera and display this information as annotations or charts in realtime (see Figure 6).

Phrase-to-phrase Matching

A pause within a single word may confuse the audience; this should be avoided. As we mentioned in the previous chapter, our system conforms to a phrase-to-phrase matching principle. In order to guarantee that the text in the manuscript correctly matches the sound at the phrase level, we introduce a semi-automatic matching tool.

First, the matching tool analyzes the sound data to find all silent sections. To be specific, the background side program will search all frames in the sound data (a wav file) for areas that both are significantly low in volume and last for over 100 milliseconds. We call these areas silent sections, which are assumed to be the intervals between adjacent phrases. However, since noise exists, usually consisting of very short bursts of high volume, especially when recording, our system can identify and ignore it to ensure that there are no silent sections separated only by noise.

Second, the matching tool generates temporary sound data with a voice synthesizer for each text phrase separated by punctuation in the manuscript file and caches the duration of each phrase.

Since the duration of each phrase generated by a voice synthesizer may be quite different from the length of each phrase in the sound file, especially when the sound is a recording of a human narrator, we choose not to match the two targets directly. We find that the ratio of the phrase duration to the full presentation length is a stable value in both cases; hence, the matching tool is based on this ratio.

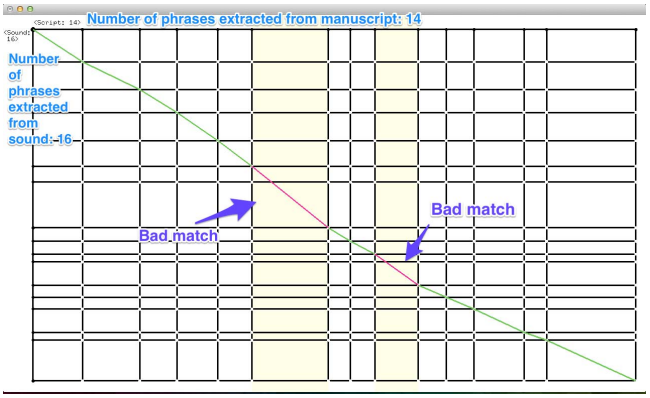

Figure 7. Matching tool conducts DP matching with the ratio of the phrase duration to the full presentation length respectively.

We refer to the DP Matching algorithm to match the two patterns and Figure 7 shows the results. The dots connected with green line segments are good matches that indicate that the phrase-to-phrase matching is successful, while red line segments indicate bad matches that will need to be fixed manually by the user.

There are two main kinds of bad matches. One occurs when the sound pauses where there is no punctuation in the manuscript text (see Figure 8). In this case, the user can choose to either add punctuation to the text or tell SmartVoice to ignore the silent sections within that particular "long" phrase. The other one occurs when the narrator ignores punctuation such as a comma or period so that the number of phrases in the manuscript is different from that in the sound data (see Figure 9). In this case, the user will have to delete the punctuation ignored by the narrator. For both cases, our matching tool provides the user an interface showing the target phrase(s) and waveforms to easily fix any mismatches manually.

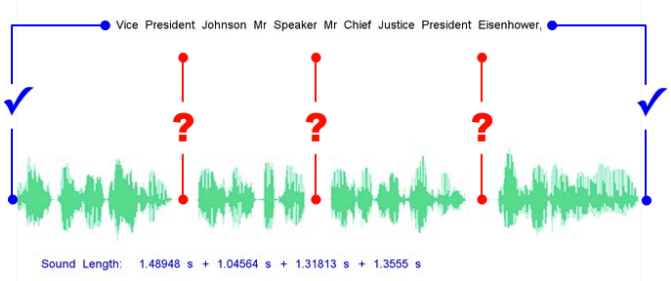

Figure 8. Bad match in which number of phrases in the text is less than that of phrases in the sound (blue line segments are automatic matching results; red line segments require manual fixing)

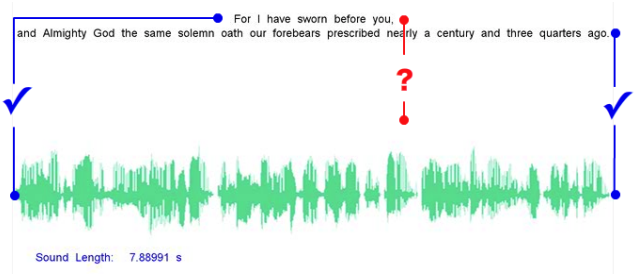

Figure 9. Bad match in which number of phrases in the text is more than that of phrases in the sound (blue line segments are automatic matching results; red line segments require manual fixing)

Playback Control, Intonation & Expression

In SmartVoice, we assign the authority to determine the duration of each pause to the user. Every time the playback pauses at a silent section, the duration of the silent section will be skipped so that when the user opens his/her mouth, the playback continues immediately with no lag.

As we discussed above, due to lack in intonation, the narration generated by a voice synthesizer may sound artificial even if the pronunciation of each word is perfect.

SmartVoice makes it possible to make the narration generated by a voice synthesizer sound vivid and natural by altering the volume, changing the playback speed, and shifting the pitch in realtime. This feature is implemented with the help of Dirac3, an open source C/C++ based multi-platform time and pitch manipulation library. [3] Based on the experimental results and the nature of human beings when speaking, we set the fluctuation band for pitch shifting to +/- one semitone in increments of 0.2 semitones and we constrained the speed variation to a range of 0.7−1.6 times normal. Volume was set to 70% of full loudness initially and could be altered between 50–100%.

Mapping

To allow the user to concentrate on the presentation, all operations are mapped to the user's face and facial parts that are detected and tracked with our system.

Pause and resume are mapped to a user's mouth openness. If the user closes his/her mouth when the current playback position is within any one of the silent sections, the playback will pause. Since the playback position will then be set to the start of the next phrase, when the user opens his/her mouth again, the playback continues with the phrase immediately. One thing to note is that when the playback position is within a phrase, the playback will not be paused until the next silent section even if the user closes his mouth. With the basic functions of SmartVoice detailed here, the audience is made to believe that a real presentation in another language has been given.

SmartVoice can also change the volume, pitch, and playback speed in realtime. As mentioned in Section 2, the pitch shifting is mapped to the position of the user's face. By calculating the y-direction difference between the current frame and the initial frame captured with the camera, SmartVoice shifts the pitch up when the user looks slightly up and vice versa. We initially mapped the volume to the openness of the user's eyes, but the rate of change of eye openness is so small that we finally chose to map the volume to the average height of both eyebrows, which moves in conjunction with the openness of eyes, but have a bigger rate of change. Like the mapping principle for pitch shifting, the difference between the current status and initial status determines the current volume. As for playback speed modification, we map it to the speed of lip sync. By comparing the difference in the mouth's area between the current and previous frame and calculating the rate of change, we can quickly grasp whether the user is speaking rapidly or not, so that our system will adjust the playback speed to the speed of lip sync.

Interface

Prompter

In our system, we introduce a prompter to provide an interface that interacts with the user during a presentation. In our prototype, we use a tablet PC (iPad) as the prompter screen, which shows the manuscript, and a half-mirror prompter base. With this prompter, it is possible for the user to read the manuscript while facing the audience naturally.

Co-operate with Background side

The main run environment on the interface side is an iPad that communicates with the background side program over WLAN. The reason for this is that we want to keep the prompter as simple as possible and keep all preparations completely on the background side computer. For instance, the manuscript shown on the prompter display is actually a text file loaded on the background side program sent to the iPad via WLAN before the presentation begins.

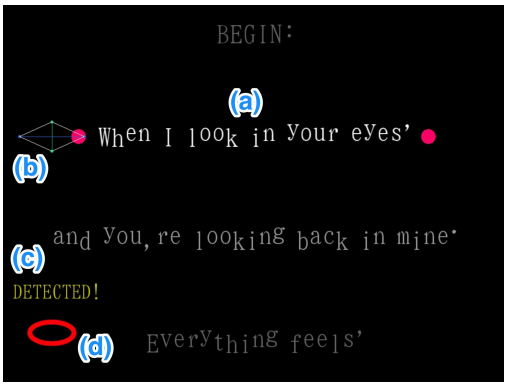

Figure 10-1. Prompter Screen

Because the user actually only interacts with the prompter during the presentation, we make it possible for the user to confirm all system behaviors from the prompter display. As shown in Figure 10-1(c), there is a label saying "DETECTED!" to notify the user that SmartVoice is successfully tracking his/her face. (The label changes to "???" if the user's face is not being tracked.)

As mentioned above, the manuscript and sound data have been correctly matched by the matching tool on a phrase-to-phrase base; hence we also break the manuscript into phrases to display them on the prompter. As Figure 5 shows, either the phrase currently being read or to be read after the current pause is emphasized and the phrase slides up after being read. (see Figure 10-1(a)).

Compared to the effects of changing the volume or pitch, the change in playback speed is difficult to immediately notice. To inform the user of the current playback speed, we add to the prompter screen an animation of a mouth whose frame rates alter with current playback speed. The mouth will be closed if the playback pauses (see Figure 10-1(d)). In addition, mouth height and width, the two parameters that determine playback speed, are represented as two orthogonal bars that change their length dynamically as the user lip syncs (see Figure 10-1(b)).

We realize that knowing the manuscript is insufficient for the user to lip sync if the manuscript is in a language that is totally strange to him/her, so we also designed another version of prompter screen, which is shown in Figure 10-2. In this version of the prompter screen, we add the translation to each phrase of the manuscript in the native language of the user (see Figure 10-2(a)) and suggest the sound level by displaying the waveform (see Figure 10-2(d)), and a pink bar is displayed as an indicator to let the user know the progress of the playback (see Figure 10-2(e)).

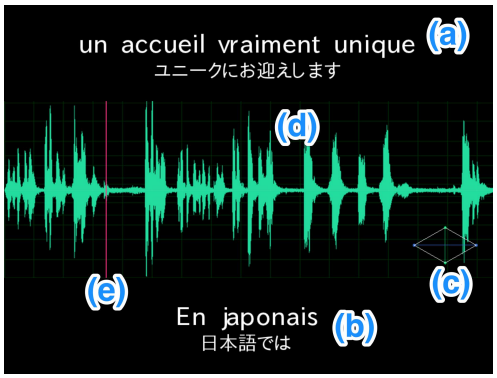

Figure 10-2. Prompter Screen ver. 2

(a) Current phase of the manuscript and its translation in native language of the user (b) next to read phrase (c) orthogonal bars showing the height and width of mouth dynamically (d) waveform (e) playback progress indicator

Performance

The phrase-to-phrase matching between manuscript text and sound data is conducted semi-automatically by the matching tool. In the case where the narration is generated by a voice synthesizer, the correct rate is 100% because the voice synthesizer always obeys the punctuation in the text file. However, if the punctuation is ignored or inserted when a human narrator reads the manuscript, there will be mismatching results that need to be fixed manually.

The background side laptop is a MacBook Pro running OSX 10.8.4 with a 2.4GHz Intel Core i5 CPU and 4GB DDR3 RAM. The tablet PC running the prompter program is a 4th generation iPad. The built-in FaceTime HD camera on the background side laptop can capture up to 720p videos at 30fps at maximum, and we set the resolution to 640x480 manually.

The background side program maintains a minimum of 55fps; hence no noticeable delay occurs with any user operations in a presentation.

EVALUATION

We conducted a user evaluation experiment to confirm the effectiveness of SmartVoice. In the user test, we showed volunteers a video that was edited with clips of real speech and speech moderated by SmartVoice. During the video playback, volunteers were asked to click and hold their mouse button whenever they thought the speech was lip synced. In order to avoid the effect of different vocal effects, the sound data used for SmartVoice was the same voice as that used for the real speech. Eight volunteers participated, all with a background in computer science.

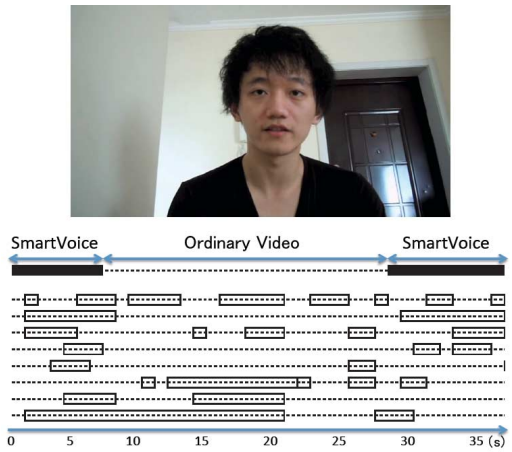

Figure 11. Results of the evaluation experiment (white rectangles in each of the eight lines indicate the sections of speech that each tester thought were generated by SmartVoice.)

Figure 11 illustrates the results. The white rectangles in each of the eight lines indicate the sections of speech that each tester thought were generated by SmartVoice. There were instances that were real speech but thought to be SmartVoice and instances that were SmartVoice speech but considered to be real. Hence, we conclude that there is no significant noticeable difference, even when we compare SmartVoice with a real presentation.

DISCUSSION

Related Work

Using facial actions as a controller is not novel. The Mouthesizer is a system that uses a mini head-mount CCD camera to track the shadow area inside the mouth using color and intensity thresholding to map it to MIDI control and instrument playing [4]. The researchers showed that although there is no difference between the Mouthesizer and the piano pedal with respect to a non-hand controller, the Mouthesizer is more intuitive and easier to learn, also suggesting the possibility of machine interfaces driven by facial actions. [5]

Facial actions are not only used for making sounds. There is also research involving a user's facial actions to control the facial expressions of CG avatars in realtime. [6]

Future Work

Integrate Voice to Video

Sometimes when we have to add narration to a demo video, especially when the narration is in a foreign language, a voice synthesizer seems to be the best choice. However, the narration generated by a voice synthesizer is too artificial to be satisfactory. By adding intonation to the synthesized voice in realtime, SmartVoice may be practical for integrating voice data generated by a voice synthesizer to videos and animations.

Ad lib

As we discussed above, the sound data used by SmartVoice needs to be prepared in advance, but it is common for a lecturer to talk about something not in the manuscript. If we are able to hide the incongruity when switching between SmartVoice and a real voice, ad-libs during a presentation may be possible.

Looping

In video production and filmmaking, video and dialogues are usually recorded separately and the dialogue is added during the automated dialogue replacement process by voice casters. The video input source is not limited to the realtime frames captured with the camera; SmartVoice is able to track the face of characters in a video as well. As a result, we may be able to use SmartVoice to do automatic looping.

Singing and Foreign Language Learning

Speeches and presentations should not be the only targets that SmartVoice aims to support. For example, SmartVoice has a prompter that is similar to that of a Karaoke system. Since both pitch and speed can be controlled in realtime with SmartVoice, our system may be able to help people who have a bad sense of pitch and rhythm to sing in the right key and tempo with practice.

For people who have to give a presentation in a foreign language, SmartVoice works as tool for practicing with a native voice. As they practice, they can memorize the pronunciation and intonation and finally get ready for the presentation with their own voice. Foreign language learners who wish to speak a language like a native may find it difficult to habitually speak with correct pronunciation and intonation, even though they often mimic native speakers. By adding a feature to compare the voice of the user to the native one for practice, we believe it may be possible for foreign language learners to develop native-like intonation with the help of SmartVoice. If the precision of detecting facial muscles is precise enough, a pronunciation correction feature is also feasible.

CONCLUSION

We proposed a presentation support system, SmartVoice, with the aim of overcoming language barriers. By matching the manuscript and sound data on a phrase-to-phrase basis, we guarantee that the sound of each phrase is correctly played according to the corresponding lip sync speed. We also attempted to solve the artificial intonation of narration generated from a voice synthesizer by changing it dynamically according to the user's facial actions. To evaluate the effectiveness of our system, we conducted a user experiment and concluded from the results that there is no significant difference between SmartVoice and real presentations. We believe that our system will enable a variety of new applications and may be the basis for substantial follow-up research.

ACKNOWLEDGMENTS
We thank all the volunteers who participated in our evaluation experiment for their help and precious advice.

REFERENCES

1. openframeworks. http://www.openframeworks.cc/.
2. ofxFaceTracker. http://github.com/kylemcdonald/ofxFaceTracker/
3. Dirac3L. http://dirac.dspdimension.com/
4. Lyons. M. J., Haehnel. M, and Tetsutani. N. The Mouthesizer: A Facial Gesture Musical Interface. In *Conference Abstracts*, *SIGGRAPH 2001*, ACM (LA, 2001), 230.
5. Lyons. M. J., Tetsutani. N. Facing the Music: A Facial Action Controlled Musical Interface. In *Proc, CHI 2001, Conference on Human Factors in Computing Systems 2001*, ACM (Seattle, 2001), 309-310.
6. T. Weise, S. Bouaziz, H. Li, and M. Pauly (2011). Realtime Performance-Based Facial Animation. *ACM Trans. Graph.*, *30*(4), 77.