Onomatopoeia Expressions for Intuitive Understanding of Remote Office Situation

Kyota Higa, Masumi Ishikawa, and Toshiyuki Nomura

Information and Media Laboratories, NEC Corporation, 1753 Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa, 211-8666, Japan
k-higa@ah.jp.nec.com, m-ishikawa@bq.jp.nec.com, t-nomura@da.jp.nec.com

Abstract. This paper proposes a system for intuitive understanding of remote office situation using onomatopoeia expressions. Onomatopoeia (imitative word) is a word that imitates sound or movement. This system detects office events such as "conversation" or "human movement" from audio and video signals of remote office, and converts them to onomatopoeia texts. Onomatopoeia texts are superimposed on the office image, and sent to the remote office. By using onomatopoeia expressions, the office event such as "conversation" and "human movement" can be compactly expressed as just one word. Thus, people can instantly understand remote office situation without watching the video for a while. Subjective experimental results show that easiness of event understanding is statistically significantly improved by the onomatopoeia expressions compared to the video at 99% confidence level. We have developed a prototype system with two cameras and eight microphones, and then have exhibited it at ultra-realistic communications forum in Japan. In the exhibition, the concept of this system was favorably accepted by visitors.

Keywords: onomatopoeia, audio/video signal, remote office situation, collaborative work.

1 Introduction

A collaborative work between remote offices requires various communication tools such as telephone, videophone, e-mail, or online-chat. These tools frequently interrupt one's work, which greatly decreases work productivity [1]. This problem is caused because people cannot understand remote office situation such as occurrence and level of conversation and human movement (e.g. walking or desk work), and thus cannot infer how busy the fellow worker is on the other side. Understanding remote office situation is important for a smooth communication between remote offices.

Methods have been proposed for estimating a busyness level of a remote office worker by using biological sensors [2] or PC operation records [3]. However, the scopes of these methods are limited to the users wearing sensors or using PCs.

Another approach is to watch a video of the remote office by using a surveillance camera. However, this requires people to monitor the video of remote office for a while for understanding remote office situation.

R. Shumaker (Ed.): VAMR/HCII 2013, Part I, LNCS 8021, pp. 351–358, 2013.

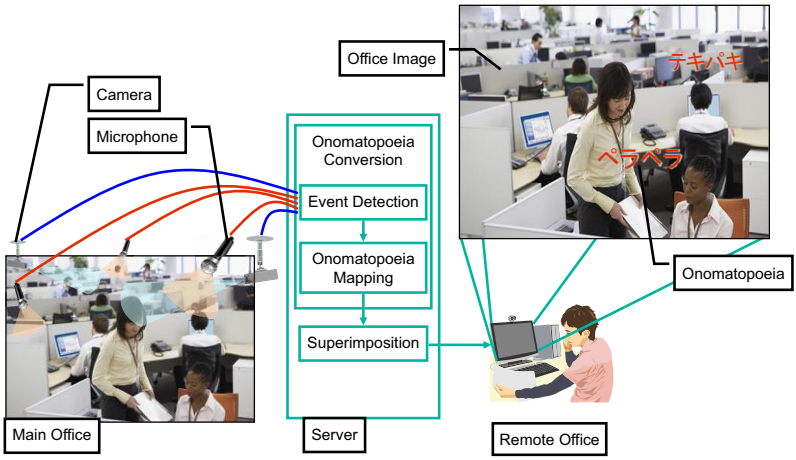


Fig. 1. System Concept

This paper proposes a system using onomatopoeia (imitative word) expressions, which enables people to instantly understand remote office situation without watching the video for a while.

2 System Concept

Figure 1 shows a concept of this system. The server detects office events such as "conversation" and "human movement (walking or desk work)" from audio and video signals captured by multiple microphones and cameras, and converts the office events to onomatopoeia texts. Onomatopoeia texts are superimposed on the office image, and sent to the remote office. A remote office worker can intuitively understand office situation by looking at the office image with onomatopoeia texts.

Onomatopoeia is a word that imitates the source of the sound or psychological states or bodily feelings (e.g. "whoosh" or "pitter-patter"). The onomatopoeia is often used in Japanese comics to explain details of various situations such as a sense of tension or a mental state of character.

By using onomatopoeia expressions, the office event such as "conversation" and "human movement" can be compactly expressed as just one word. Thus, people can instantly understand remote office situation without watching the video for a while. Furthermore, personal privacy is protected because details of the conversation contents are not presented to the fellow worker.

3 Onomatopoeia Conversion

Figure 2 shows the block diagram of onomatopoeia conversion. The audio and video signals are analyzed to detect office events, which are then mapped to onomatopoeia texts in the database.

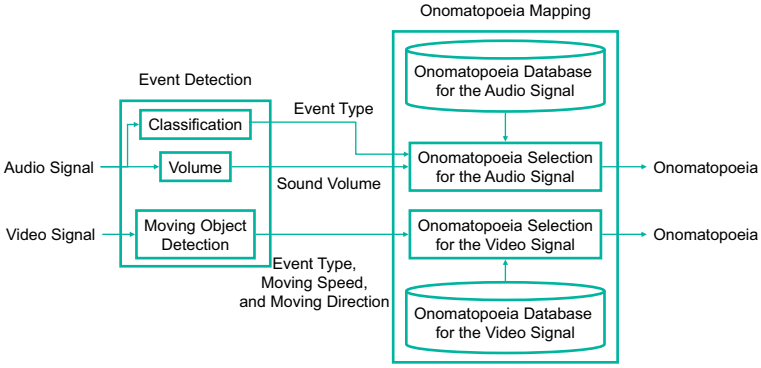


Fig. 2. Block diagram of onomatopoeia conversion

3.1 Onomatopoeia Conversion for Audio Signal

The audio signal is used for detecting "conversation" and its level. Onomatopoeia texts are selected based on sound volume and event type extracted from the audio signal.

The sound volume is calculated by integrating amplitude energy of audio frames. The event type is determined by classifying audio frames based on temporal change of amplitude energy and shape of frequency spectrum [4]. Each audio frame is classified into "non-conversation" (silence) or "conversation," and then the class with highest votes in N audio frames is selected as the event type.

The sound volume and the event type are mapped to onomatopoeia text as depicted in Fig. 3. "Non-conversation" is mapped to "Shiin (describing absolute silence)." "Conversation" is mapped to "Hiso hiso (describing talk in a dim voice)," "Boso boso (describing talk in a low voice)," "Pera pera (describing fluent talk at a normal voice)," or "Gaya gaya (describing talk in a loud voice)" based on the sound volume in ascending order. These onomatopoeia texts are heuristically selected for describing "conversation."

3.2 Onomatopoeia Conversion for Video Signal

The video signal is used for detecting "human movement (walking or desk work)" and its level. Onomatopoeia texts are selected based on moving direction, moving speed, and moving area of moving objects extracted from the video signal.

The moving objects are detected based on motion vectors of keypoints in the video frames. To calculate the motion vectors between adjacent frames, the keypoints are detected and tracked by Harris Corner Detector and Kanade Lucas Tomasi Tracker [5]. The keypoints with large motion vectors are grouped as a moving object.

The objects are tracked based on intersection ratio in adjacent frames to estimate the moving direction, moving speed, and moving area of moving objects. The moving

direction and speed are determined from magnitude and direction of motion vectors belonging to the objects, respectively. The moving area is a bounding rectangle including trajectory of the moving object center.

Event Type	Low → Sound Volume Level → High			
Non-Conversation	シーン (Shiin: describing absolute silence)			
Conversation	ヒソヒソ (Hiso hiso: describing talk in a dim voice)	ボソボソ (Boso boso: describing talk in a low voice)	ペラペラ (Pera pera: describing fluent talk at a normal volume)	ガヤガヤ (Gaya gaya: describing talk in a group)

Fig. 3. Onomatopoeia mapping for the audio signal

Event Type	Slow → Moving Speed → Fast	
Walking	テクテク (Teku teku: describing walk at normal pace)	スタスタ (Suta suta: describing quick straight walk)
Desk Work	ゴソゴソ (Goso goso: describing subtle movement)	テキパキ (Teki paki: describing crisp moevment)
	Narrow → Moving Area → Broad	

Fig. 4. Onomatopoeia mapping for the video signal

The moving direction, moving speed, and moving area are mapped to onomatopoeia text as depicted in Fig. 4. “Walking” or “desk work” are selected based on the moving direction. “Walking” is mapped to “Teku teku (describing walk at normal pace)” or “Suta suta (describing quick straight walk)” based on the moving speed in ascending order. “Desk work” is mapped to “Goso goso (describing subtle movement)” or “Teki paki (describing crisp movement)” based on the moving area in ascending order. These onomatopoeia texts are heuristically selected for describing “walking” and “desk work.”

Fig. 5. Examples of images with onomatopoeia texts using this experiment

4 Evaluation

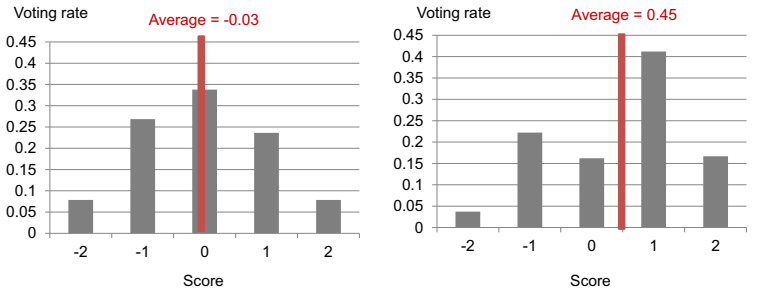
4.1 Experimental Conditions

We evaluated an effectiveness of the onomatopoeia expressions using 18 office events. The events consist of nine "conversation" and nine "human movement." Each event is presented to 12 subjects in two ways: a short video (4 to 13 seconds) and an image with onomatopoeia texts extracted from the video. The subjects compared correctness and easiness of event understanding between two ways of presentation. The subjects answered a score from -2 to 2 on five-levels where the higher score means better rating of the onomatopoeia expressions. The subjects are divided into two groups. We present the video and the image in a different order with respect to each group.

The length of audio frame is 10 ms, and onomatopoeia texts for audio and video signal are superimposed on a center of the upper on the image and a position of the moving object, respectively. Figure 5 shows examples of the images with onomatopoeia texts using this experiment. Onomatopoeia texts representing "conversation" or "human movement" are superimposed on each image.

4.2 Results

Figure 6 shows results of subjective experiments. Figure 6 (a) shows the voting rate on correctness of event understanding, and Figure 6 (b) shows the voting rate on easiness of event understanding.



(a) Correctness of event understanding (b) Easiness of event understanding

Fig. 6. Results of subjective experiments

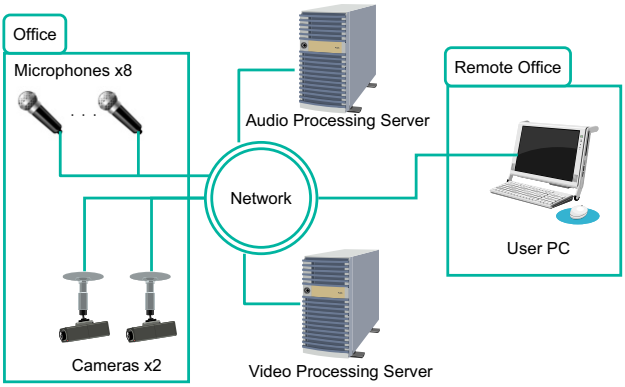


Fig. 7. System configuration of the prototype

The average score on correctness of event understanding is -0.03, which means office events can be equally understood with both ways of presenting. According to the subject comments, the onomatopoeia expressions are highly evaluated because of two advantages: (1) detecting of small events which are hard to notice by watching the video, such as slight human movements and hush conversations, and (2) correctly finding of a place where each event occurs. On the other hand, two disadvantages of onomatopoeia expressions are extracted: (1) misunderstanding of office events when onomatopoeia texts are superimposed on wrong positions, especially conversation, and (2) strange feeling from mismatch between office events and onomatopoeia texts.

The average score on easiness of event understanding is 0.45, which indicates office events can be more quickly understood by the images with onomatopoeia texts than the videos. The difference between both ways of presenting is statistically significant at 99% confidence level according to the t-test. The subjects favorably accept that the onomatopoeia expressions enable to instantly understand “walking” which requires a few seconds to see whole of the event.

Table 1. Specification of equipment for prototype system

Audio/Video Processing Server	HP xw6600 Workstation, Intel® Xeon® CPU E5450 @ 3.00GHz, 3.25 GB RAM, Windows XP Professional Service Pack 3
Microphone	Sony ECM-C10
Network Camera	Axis 2100
User PC	Windows XP Professional Service Pack 3, Intel® Core™2 Duo CPU E6850 @ 3.00GHz, 1.96 GB RAM

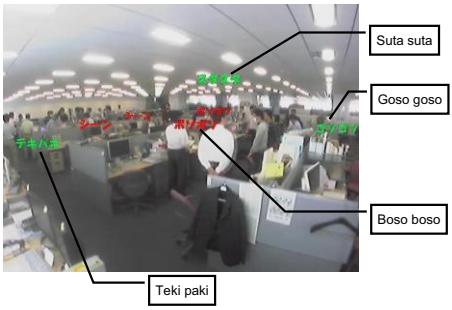


Fig. 8. Office image with onomatopoeia texts

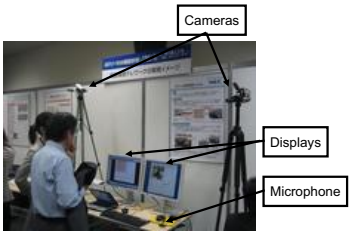


Fig. 9. Scene of exhibition

The onomatopoeia expressions are effective for intuitive understanding of office situation while the correctness of event understanding is comparable. To improve the onomatopoeia expressions, the following functions should be implemented: (1) detection of appropriate superimposing positions to prevent misunderstanding of office events and (2) selection of onomatopoeia texts suitable for the office events depending on user's preference.

5 Prototype System

We developed a prototype of the proposed system shown in Fig. 7. This system uses eight microphones and two cameras to monitor approximately 16 people. The specification of the system equipment is shown in Table 1. Onomatopoeia texts for audio and video signal are superimposed on a position of microphone in the office image and the moving object, respectively.

Figure 8 shows an office image with onomatopoeia texts of a scene that people get together for a short conversation. Red texts represent "conversation" and green texts represent "human movement." Onomatopoeia text such as "Suta suta (its meaning is shown in Fig. 4)" or "Goso goso" for describing human movements, and "Boso boso" for describing conversation are superimposed. The remote office worker can instantaneously understand that people are gathering and having a talk.

We have exhibited this system at ultra-realistic communications forum in Japan. As shown in Fig. 9, the system presented the events in the conference room. In the exhibition, the concept of this system was favorably accepted by visitors. We also got comments which recommend applying the onomatopoeia expressions to entertainment applications.

6 Conclusion

We have proposed a system for intuitive understanding of remote office situation using onomatopoeia expressions. This system detects office events such as "conversation" or "human movement" from audio and video signals of remote office, and converts them to onomatopoeia texts. Onomatopoeia texts are superimposed on the office image, and sent to the remote office. By using onomatopoeia expressions, the office events can be compactly expressed as just one word. Thus, people can instantly understand remote office situation without watching the video for a while. Subjective experimental results showed that easiness of event understanding is statistically significantly improved by the onomatopoeia expressions compared to the video at 99% confidence level. We have developed a prototype system with two cameras and eight microphones, and then have exhibited it at ultra-realistic communications forum in Japan. In the exhibition, the concept of this system was favorably accepted by visitors.

Acknowledgments. This work is partly supported by National Institute of Information and Communications Technology (NICT), Japan.

References

1. Mark., G., Gonzalez., V.M., Harris, J.: No Task Left Behind? Examining the Nature of Fragmented Work. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems, pp. 321–330 (2005)
2. Chen, D., Hart, J., Vertegaal, R.: Towards a Physiological Model of User Interruptability. In: Baranauskas, C., Abascal, J., Barbosa, S.D.J. (eds.) INTERACT 2007. LNCS, vol. 4663, pp. 439–451. Springer, Heidelberg (2007)
3. Tanaka, T., Fujita, K.: Interaction Mediate Agent Based on User Interruptibility Estimation. In: Human-Computer Interaction International, pp. 152–160 (2011)
4. Recommendation ITU-T G.720.1: Generic Sound Activity Detector. ITU-T (2010)
5. Shi., J., Tomasi., C.: Good Features to Track. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 593–600 (1994)