Laughter Detection for an Assisting Tool of Group Conversation

Myagmarbayar Nergui and Mihoko Otake
Mechanical Engineering Department, Graduate School of Engineering, Chiba University, Yayoi-cho 1-33, Inage-ku, Chiba-shi, Japan
Email: myagaa@chiba-u.jp otake@chiba-u.jp

***Abstract*—In super aged society, population with age related disease such as dementia increases, which is disorder of brain functions. Interactive group conversation or group activities play an important role of activating brain functions. In order to make interactive group conversation among older people for preventing or inhibiting the progress of mild cognitive decline and dementia, our research aims to develop a system assisting group conversation, which is capable of detecting laughter and emotions, and recognizing speech contents of the participants. This study is a part of the developing system to recognize laughter from acoustic signals of group conversation. Mel frequency Cepstral Coefficients (MFCCs) are calculated and vector quantized from the acoustic signals. Minimum distance classifier is applied for recognition of laughing and non-laughing states. Experimental results show very good performance. F-score and accuracy rate are 88% and 98%, respectively.**

***Index Terms*—laughter detection, group conversation, dementia, mel frequency cepstral coefficents (MFCC)**

I. Introduction

Nowadays, developed countries had become already and developing countries are supposed to become super aged societies in this century. Above all, highly developed countries are lacking real human interactions, and the human lives are too busy due to highly technological environment. It causes super aged people to stay alone. In a super aged society, there have been some problems of mental diseases, such as dementia. People with dementia are lacking memory due to the disorders of brain functions. In order to prevent and inhibit progress of mild cognitive impairment and dementia, older people have to activate their brain functions. For activating brain functions effectively, social interactions play an important role. Group conversation is one of the typical social interactions. Reminiscence and life review [1] have been studied for half a century as group or individual therapy. Recently, the coimagination method [2] was proposed for prevention of onset and progress of dementia and cognitive decline.

To make interactive group conversation among older people, there is a demand for a system which can handle the participants of group conversation to participate equally, and make involving environment. Interactive conversation means that the participants listen to each other well and respond by verbal and non-verbal communication such as facial expressions, nodding, laughing and so on.

Laughter is one of the important cues in human communication and social interactions. Laughter is a part of human behavior regulated by the brain, helping humans clarify their intentions in social interaction and providing an emotional context to conversations. Laughter is used as a signal for being part of a group — it signals acceptance and positive interactions with others. Laughter is sometimes seen as contagious, and the laughter of one person can itself provoke laughter from others as a positive feedback [3].

In this study, we develop a method of laughter detection using acoustic signals during group conversation among older people.

The method proposed in this study is intended to be used in a part of the system that assists group conversation among older adults. The system can be a talkative robot, which can listen to each participant well, respond to their talks by re-voicing some speech contents, detect laughter, and laugh at right time. Previously developed system towards this goal is described in [4]-[7].

The paper is organized as follows. Section II describes related studies of a system assisting group conversation and laughter detection. Section III represents main structure of methodology and laughter detection from continuous speech using acoustic signal. Section IV demonstrates experiments and its results. Finally, in section V, we discuss and conclude.

II. Related Studies

A. A System Assisting Group Conversation

For making an interactive group conversation among older people, we develop a system that handles duration of each speech, reads a mood of group conversation through visual and acoustic data of each participant in group conversation, and gives a feedback according to the result of the system. In particular, we have been focusing on smile and laughter as major responses for the purpose of estimating internal state of each participant.

Firstly, we developed system assisting group conversation of older adults which recognizes a smiling

Manuscript received August 19, 2013; November 20, 2013

©2013 Engineering and Technology Publishing

doi: 10.12720/ijsps.1.2.212-217

face using a conventional web camera, calculates smile degrees of the participants, and evaluates duration of speech using headset with a microphone for each participant [4].

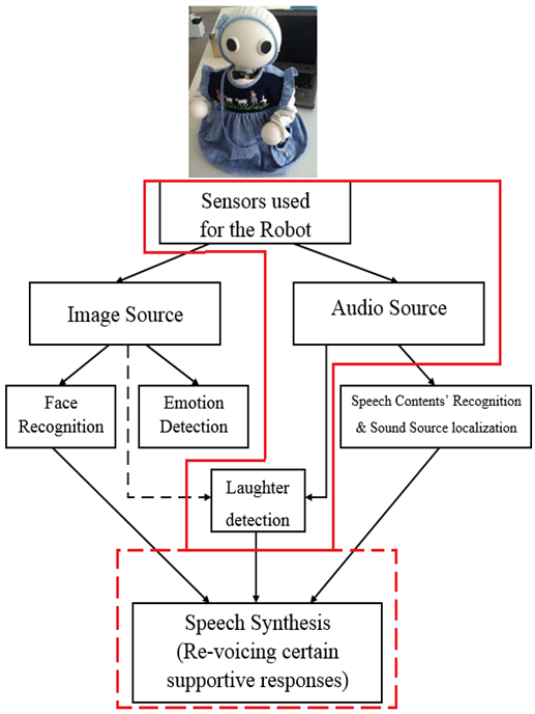

Figure 1. Main structure of the system that assists group conversation

Secondly, we developed system assisting group conversation which comprises following four functions using only a Kinect sensor [5].

1) face detection and recognition,
2) some certain speech contents recognition
3) sound source localization
4) speech synthesis (re-voicing some certain responses based on the results of speech contents recognition).

Thirdly, we preliminary studied facial expression detection and emotion detection using wireframe model of a face, and can be applied to a mood reader of group conversation by detecting the participants' emotions from their face [6].

Most recently, we proposed laugher detection algorithm, by combining respiratory data and the smiling degree obtained by facial images. In order to obtain respiratory data, we developed respiratory sensor using a dielectric elastomer which is an elastic non-skin contact sensor.

B. Laughter Detection

In literatures, there have been many studies related to laughter detection based on different classification algorithms using many signal sources received from many kinds of sensors, such as acoustic signals, video signals, and facial expressions. Different types of features such as spectral and prosodic for laughter detection were investigated using different classification techniques including Gaussian Mixture Models, Support Vector Machines, Multi-Layer Perceptron which are often used in language and speaker recognition [8]. Knox and Mirghafori detects laughter automatically using Neural Networks based on extracted features, such as Mel-Frequency Cepstral Coefficients, pitch and energy, from acoustic signals [9]. Gaussian Mixture Models were trained with Perceptual Linear Prediction features, pitch and energy, pitch and voicing, and modulation spectrum features to model laughter and speech [10]. While these studies have been using ICSI Meeting Recording Corpus[11] for comparing the performance of their methods, our study focuses on developing a method which is applicable to natural everyday conversations among older adults based on the previously developed methods [8], [9], [10].

III. Methodology of Assisting System to Group Conversation

A. A Main Structure of the Methodology

Main purpose of developing the system assisting group conversation is to make a more interactive system for all the participants. In the developed system, every participant can take part evenly in group conversation, and enjoy their talks supported by the robot, which is capable of looking at the participants, listening, speaking, and nodding with supportive responses. Main structure of the system assisting group conversation is shown in Fig. 1. In this study, we consider using a less number of sensors and less complex system. Because of this reason, we use a Kinect sensor, which consists of a color camera, an infrared emitter, a depth sensor, and four microphone arrays. Main structure of our developed system shown in Fig. 1 is based on the Kinect sensor. The Kinect sensor has a vision and a hearing functions for the robot which assists group conversation. In this paper, we focus on only laughter detection as a part of the developed system. This study describes laughter detection from acoustic sounds obtained by the microphone used for the robot, which is surrounded by the red line in Fig. 1. Other modules including speech contents' recognition, face recognition and Speech synthesis are developed in the previous studies [5], [6]. The dashed arrow shows a planned feedback of the system by applying some speech responses based on results of laughter detection. Following sections describe laughter detection from acoustic signals in more details.

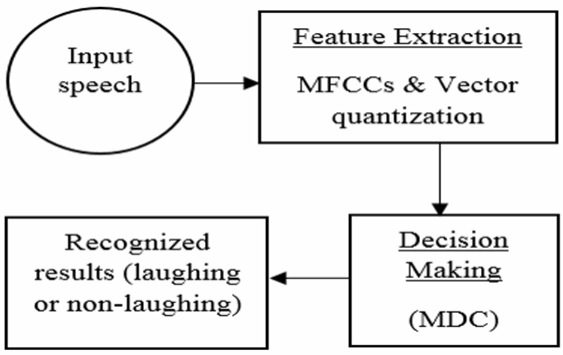

Figure 2. Block diagram of laughter detection

©2013 Engineering and Technology Publishing

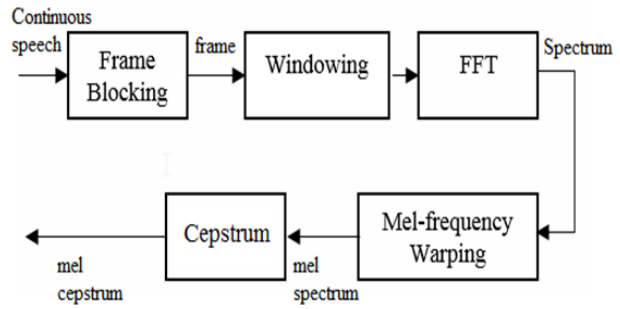

Figure 3. Feature extraction steps

B. Laughter Detection from Continuous Speech

In order to detect laughter from continuous speech signal, we have to obtain characteristics of laughter. For obtaining laughter characteristics, we used Mel Frequency Cepstral Coefficients. Then we applied a Linde–Buzo–Gray (LBG) algorithm, introduced by Yoseph Linde, Andrés Buzo and Robert M. Gray in 1980 [12], as a vector quantization algorithm on extracted MFCCs of the speech signals to derive a good codebook of laughter. For classifying laughter from the continuous speech signal or for decision making process, we have applied a minimum distance classifier, such as Euclidean distance classifier. Fig. 2 shows block diagram of laughter detection.

1) Mel frequency cepstral coefficents (MFCC)

Mel Frequency Cepstral Coefficents (MFCCs) are features widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since. Prior to the introduction of MFCCs, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) were the main feature type for automatic speech recognition (ASR).

Feature Extraction is used in both training and recognition phases. Feature extraction steps are as follows and shown in Fig. 3.

It comprises the following steps:

1. Frame Blocking
2. Windowing
3. FFT (Fast Fourier Transform)
4. Mel-Frequency Wrapping
5. Cepstrum (Mel Frequency Cepstral Coefficients)

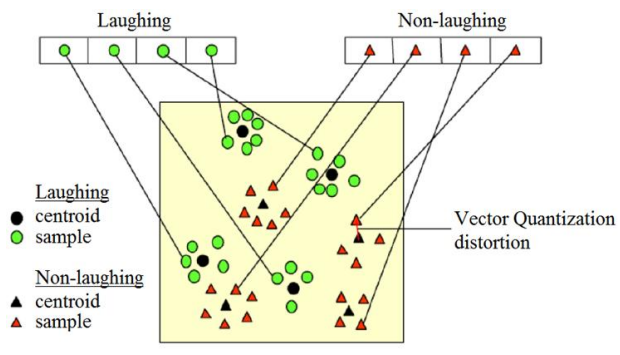

Figure 4. Conceptual diagram drawing vector quantization codebook

The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely what humans hear.

The formula for converting from frequency to Mel scale is:

$$mel(f)=2595*log10(1+f/700) \quad (1)$$

2) Vector quantization and minimum distance classifier

Vector quantization (VQ) is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a centroid. The collection of all codewords is called a codebook.

Fig. 4 shows a conceptual diagram to illustrate this recognition process. In the figure, only two states (laughing and non-laughing) and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the laughing while the triangles are from the non-laughing. In the training phase, laughing and non-laughing VQ codebook is generated from the acoustic speech signal. The result codewords (centroids) are shown in Figure by black circles and black triangles for laughing and non-laughing, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of group conversation is framed a certain sized window and is "vector-quantized" using each trained codebook and the total VQ distortion is computed.

The states of laughing or non-laughing corresponding to the VQ codebook with smallest total distortion are identified. The smallest total distortion means that minimum distance classifier, such as Euclidean distance classifier is applied in here. Laughing and non-laughing can be discriminated from each other based on the location of centroids.

IV. Experiments and Experimental Results

Experiments have two stages:

A. Experimental Stage I

We have conducted the experiment among three older sisters, who are famous for their healthy longevity in Japan. Their average age is 92 years old. They are physically and mentally healthy. The coimagination method was used for supporting group conversations [2]. Each participant talks one minute about their experiences using one photo. Each speaker has three topics with three images. They also have 3 minutes for questions and answers for each topic.

During experiment, we recorded visual and acoustic signals by using Sony Handycam video recorders located in the back and in front of them. We also recorded acoustic signals by using Zoom handy recorder put in front of each of them.

In this experiment, 15 min group conversation data was recorded. Sampling frequency of recorded acoustic signal is 44100Hz. For laughter detection, the recorded acoustic signal is converted to 16000 Hz sampling

©2013 Engineering and Technology Publishing

frequency. Overall process of laughter detection is written in Matlab programming language. We made hand transcribed note as laughing or non-laughing on the overall speech signal. Mel Frequency Cepstral Coefficients (MFCC) were extracted from several acoustic signals during laughing and non-laughing (speaking and other sounds) states. Then we applied LBG algorithm as a vector quantization algorithm on extracted MFCCs of the speech signals to derive a good codebook of laughter for identifying laughing or non-laughing from acoustic signal. With the above signal processing steps, Fig.5 shows a part of the acoustic signals during laughing and its calculated MFCCs and VQ signal.

B. Experimental Stage II

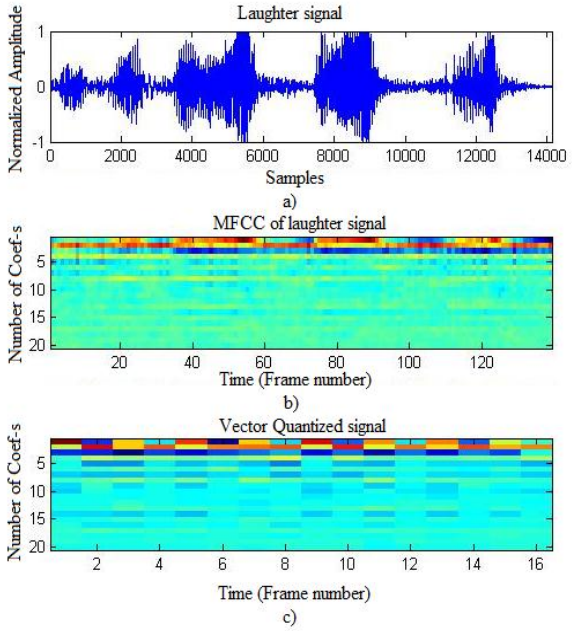

Figure.5 Signal processing

We also recorded all visual and acoustic signals during preparation of experiments. This experimental data includes basic data of more than six people, who are three older women, four experimenters. Acoustic signal comprises all conversation (speech and other voice sounds) among them. In this experiment, 20 minutes group conversation data was recorded.

For training phase, we extracted several frames of laughing and non-laughing signals from the experimental data, calculated MFCCs and generated their VQ codebook.

For recognizing phase, we applied all experimental data for recognizing laughing or non-laughing from continuous acoustic signals. Recognition phase has two stages, which are that continuous acoustic signals are framed by two different windows: 12000 samples and 14000 samples. Why we have chosen this two window size is that laughter signal average sampling periods were in between 12000 and 14000 samples in which signal is sampled by 16000Hz frequency.

By two window framed signals, we evaluated which window frame gives us better recognition results. In every frame of speech signals, each MFCCs and its vector quantization are calculated. Then minimum distance classifier, namely, Euclidean distance classifier was applied on vector quantized codebook for recognizing laughing or non-laughing states from acoustic signals.

Fig. 6 shows experimental results of laughing or non-laughing during group conversation. Fig. 6a shows real states of group conversation, which is marked laughing or non-laughing as states of conversation on the utterance signals. Laughing is painted by orange color, and non-laughing is painted by gray color.

Fig. 6b shows acoustic signals recorded during the experiment. Fig. 6c shows the recognized results of laughter detection. In Fig. 6c, the number 2 represents laughing state, while the number 1 represents non-laughing or other conversation state. From the recognition results, we could easily see that recognized states are almost similar to real states of group conversation.

In order to verify the recognition results, we calculated precision, recall, F-score and accuracy rate in each window frame. Table I shows the recognition results of laughing or non-laughing states in different window framing.

TABLE I. RECOGNITION RESULTS OF LAUGHING OR NON-LAUGHING STATES IN DIFFERENT WINDOW FRAMING

Frame Windows	12000 samples	14000 samples
Precision	0.81	0.85
Recall	0.97	0.93
F-score	0.88	0.88
Accuracy rate	0.98	0.98

From Table I, we can see that precision, recall, F-score and accuracy rate are very high values. Recall of framed window with 12000 samples is higher than that of framed window with 14000 samples. But precision of framed window with 12000 samples is lower than that of framed window with 14000 samples. This means that small size framed window is recognized laughing states well, but non-laughing states are possibility to recognize as laughing states. Finally, F-score and accuracy rate showed same results in both framed window of acoustic signal.

©2013 Engineering and Technology Publishing

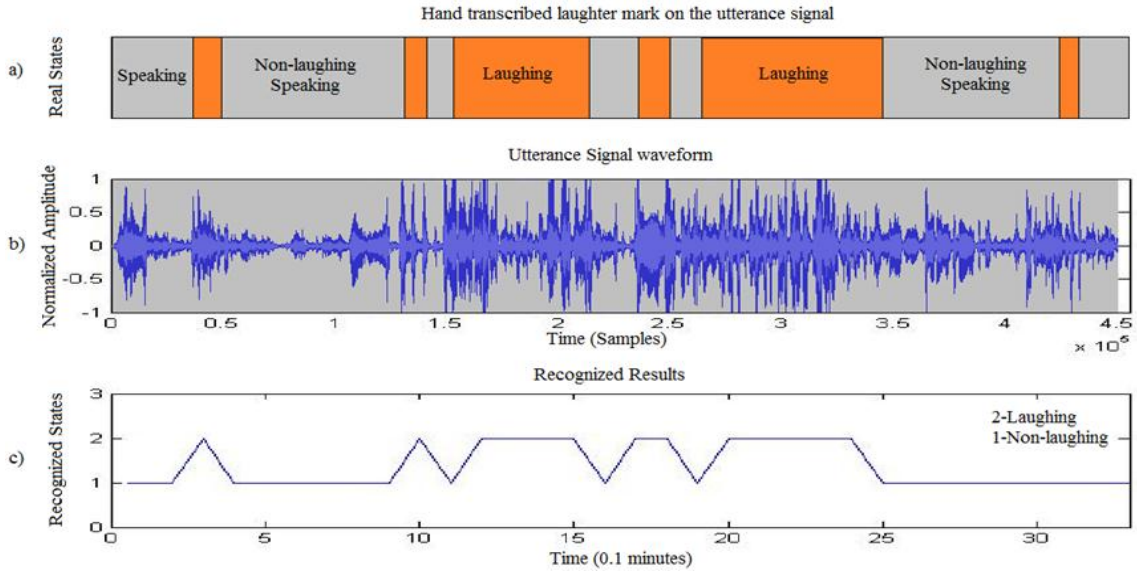

Figure.6 Experimental results

V. Discussion and Conclusion

We developed laughter detection from acoustic signals as a part of a system assisting group conversation. Our previous studies [4]-[7] constructed a main structure of system assisting group conversation. The developed system can assist the participants of group conversation among older people by handling duration of speech, reading a mood of group and giving feedbacks based on situations of group conversation. One of our previous studies developed automatic laughter detection using respiratory sensory data with smile degree data [7]. Respiratory sensor data is not robust during group conversation due to body motions of the participants. Therefore, in this study, we aimed to detect laugher from acoustic signals. We extracted some important features using MFCCs from acoustic signals recorded during group conversation, created codebook of laughing and non-laughing. Then overall speech signal of group conversation has been tested for recognition of laughing and non-laughing states. From continuous speech, acoustic signal is framed by windows for detecting laughing or non-laughing. Samples of framing window applied on this study are 12000 and 14000 samples. In order to know which one is good for laughter detection, we tested acoustic signal data for both window samples. Experimental results of both cases showed very good performance. We calculated precision, recall, F-score and accuracy rate. By small framed window with 12000 samples, recall was higher than that by large framed window with 14000 samples. Precision was vice versa: that by small framed window was lower than that by large framed window. Laughing state was recognized well but some of the non-laughing states were mistakenly recognized as laughing states by small framed window. . Finally, F-score and accuracy rate were calculated and showed same results with both framed window of acoustic signal. F-score and accuracy rate are 0.88 and 0.98, respectively. Even experimental data consists of multiple participants, and not using any headset microphones, F-score and accuracy were still very high. The result was successful compared to the previous studies whose maximum accuracy rate were 0.92 [8] to 0.98 [9]. The supposed reason for this success is that the laughter of the participants was very clear during everyday conversation compared to that in the meeting corpora. Further investigation is needed for group conversation among different groups of older adults and different age groups in order to test applicability of the method.

In this study, laughter detection was implemented in offline. We will implement it in online system for real-time feedback and combine it with other parts of developing system in future works.

Acknowledgment

This Research was supported in part by Grant-in-Aid for Scientific Research on priority area Founding a creative society via collaboration between humans and systems (#4101) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] B. K. Haight and I. Burnside, "Reminiscence and life review: Explaining the differences," *Archives of Psychiatric Nursing*, vol. 7, 1993.

[2] M. Otake, M. Kato, T. Takagi, and H. Asama, "The coimagination method and its evaluation via the conversation interactivity measuring method, early detection and rehabilitation technologies for dementia," in *Neuroscience and Biomedical Applications*, Jinglong Wu (Ed.), IGI Global, 2001, pp. 356 - 364.

[3] Camazine, Scott, ed. *Self-Organization in Biological Systems*. Princeton University Press, 2003.

[4] T. Yamaguchi, J. Ota, and M. Otake, "A system that assists group conversation of older adults by evaluating speech duration and facial expression of each participant during conversation," in *Proc. IEEE International Conference on Robotics and Automation*, USA, 2012, pp. 4481-4486.

©2013 Engineering and Technology Publishing

[5] M. Nergui and M. Otake, "Development of a tool for assisting group conversation by re-voicing supportive responses," in *Proc. ICEMA*, in Press, Singapore, 2013.

[6] M. Nergui and M. Otake, "Facial expression detection through a RGB-D sensor for an assisting tool of group conversation," in *Proc. ICMTSET*, in press, Dubai, 2013.

[7] M. Nergui, P. H. Lin, G. Nagamatsu, M. Waki, and M. Otake, "Automatic detection of laughter using respiratory sensor data with smile degree," in *Proc. IIENG*, Bangkok, Thailand, 2013.

[8] K. P. Truong and D. A. van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, no. 2, pp. 144-158, 2007.

[9] M. T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," *INTERSPEECH*, pp. 2973-2976, 2007.

[10] K. P. Truong and D. A. van Leeuwen, "Automatic detection of laughter," *INTERSPEECH*, pp. 485-488, 2005.

[11] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, and N. Morgan *et al*, "The ICSI meeting corpus," in *Proc. ICASSP*, Hong Kong, 2003

[12] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communication*, vol. Com-28, no. 1, 1980.

Myagmarbayar Nergui was born in Uvurkhangai province, Mongolia. She received complete secondary education and high education in Mongolia. She received a Bachelor of Science degree in radio communication engineering from School of information Technology and Communication, Mongolian University of Science and Technology in 2002, a master of technology degree in communication engineering from National Institute of Technology, Karnataka, India in 2009, and a Ph D in medical system engineering form Graduate School of Engineering, Chiba University, Japan in 2013. Myagmarbayar works as a specially appointed researcher at Mechanical Engineering Department, Graduate School of Engineering, Chiba University, Japan. She worked as a quality engineer and a marketing manager at Mongolian Telecommunication Company for 5 years, and as a patent engineer at Law Firm of Naren Thappetta, Bangalore India. Her research area is medical system engineering, mobile and communication robotics, and machine learning. Dr. Myagmarbayar Nergui was awarded as a special young researcher (DC-2) of Japan Science and Promotion Society, 2011-2013, and received Excellent International Students Scholarship of Chiba University, 2009-2012, and Indian Council and Cultural Relations Scholarship for master degree study. She is a member of IEEE EMBSociety and Science and Engineering Institute (SCIEI).

Mihoko Otake was born in Tokyo, Japan. She received her B.E., M.E., and Ph.D. in Mechano-Informatics in 1998, 2000, and 2003, respectively, all from the University of Tokyo, Tokyo, Japan. She has been appointed as an Associate Professor with Chiba University, Chiba, Japan. Prior to the current position, she was a Specially Appointed Assistant Professor, an Assistant Professor, an Associate Professor with the University of Tokyo. She published world first book on gel robots, "Electroactive Polymer Gel Robots - Modelling and Control of Artificial Muscles" (Heiderburg, Germany: Springer-Verlag, 2009). Her research topics include modeling and simulation of cognitive function of humans and electroactive polymers and thier applications to design and control of intelligent systems and services. Dr. Otake is a member of the IEEE Robotics and Automation Society, Society for Neuroscience, Gerontology Society of America, Information Processing Society of Japan, Japan Society for Artificial Intelligence, and Robotics Society of Japan. She was awarded as a JSPS Research Fellow in 2001, JST Fellow in 2004 and 2010.

©2013 Engineering and Technology Publishing