A Voice Signal-Based Manipulation Method for the Bio-Remote Environment Control System Based on Candidate Word Discriminations

Taro Shibanoki
Ibaraki University, Japan
Go Nakamura, Takaaki Chin
Hyogo Rehabilitation Center, Japan
Toshio Tsuji
Hiroshima University, Japan

Abstract

This paper proposes a voice signal-based manipulation method for the Bio-Remote environment control system. The proposed system learns relationships between multiple candidate words' phonemes extracted by a large-vocabulary speaker-independent model and control commands based on a self-learning look-up table. This allows the user to control various devices even if false recognition results are extracted. Experimental results showed that the method accurately discriminate slurred words (average discrimination rate: 94.4±2.53 [%]), and that the participant was able to voluntarily control domestic appliances.

Keywords: environment control system (ECS), speech recognition, candidate word, learning-type look-up table

1. Introduction

A variety of environmental control systems (ECSs) for people with disabilities and bedridden elderlies have been developed to be self-sufficient and maintain independent life in recent years.

There are number of studies to develop such systems using biological signals [2] - [5]. As an example, the Bio-Remote – a new ECS developed by our research group [4][5] – has the distinctive features of (1) various input systems such as biological signals, keyboard and mouse input to meet user requirements, (2) flexible adaptation to individual users depending on their capabilities, and (3) learning function to enable adaptation to variations among individuals.

The Bio-Remote has been proven effective in support for the everyday lives of people with spinal injuries through its usefulness in the operation of domestic appliances [5]. However, sensors must be attached to skin surface with paste and medical tapes, it is user's unpleasantness and burden because the procedure to wear electrodes is complicated. Additionally, because of the effectiveness of changes of skin impedance such as perspiration, long-time use of the system is difficult. To overcome these problems, we focus voice signals as an extension input of the Bio-Remote.

A number of speech controlled ECS have been developed, such as Voicecan (Voicecan co., ltd) [6] and Lifetact (Asahi kasei technosystem co., ltd) [7]. These systems discriminate users' intensions from recorded voice signals based on a speaker independent acoustic model and control devices corresponding to discrimination results. It is, therefore, difficult to accurately discriminate speech of patients with dysarthria who speak difficult, since the model used in these systems consider to standard adults' speech. On the other hand, training each user's voice and building a speaker dependent model, it is possible to accurately discriminate speech of patients with dysarthria [8]. However, it takes a lot of time and needs a large amount of data to train a speaker dependent model, it is possible to be a burden to a user.

This paper proposes a novel environment control system for patients with dysarthria using voice signals.

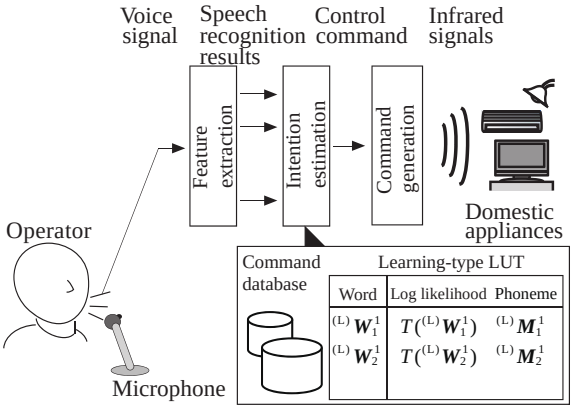

Fig. 1. Overview of the proposed system.

The system trains individual user's features based on the false recognition results using a speaker independent model and can discriminate user's intensions without large amount of training data.

2. Speech controlled Bio-Remote

Figure 1 shows the overview of the proposed system. The proposed system consists of voice signal measurement and feature extraction, operation estimation, and device control stages. The details of each stage are outlined as follows.

2.1. *Voice signal measurement and feature extraction [10]*

Voice signals were recorded using a microphone and digitized using an A/D converter (sampling frequency: f_s [Hz]). Mel-frequency cepstrum coefficients (MFCCs) are then extracted as the inverse cosine transform is applied to the log power spectra of the sampled signals. Feature vector $\boldsymbol{X}$ used for speech recognition is defined as the low-frequency components of each frame of extracted MFCCs [9].

Next, output probabilities $P(\boldsymbol{W})$ of word $\boldsymbol{W} = \{w_1, w_2, \ldots, w_K\}$ (w_k: word, K: number of words) is calculated approximately using N-gram model. Additionally, output probabilities $P(\boldsymbol{X}|\boldsymbol{W})$ of a feature vector $\boldsymbol{X}$ from $\boldsymbol{W}$ is calculated using the acoustic model. A phoneme hidden Markov model (phoneme HMM), which can consider a context and a time variation, is used to calculate $P(\boldsymbol{X}|\boldsymbol{W})$. $P(\boldsymbol{X}|\boldsymbol{W})$ is calculated dividing the words $\boldsymbol{W}$ to phonemes $m = \{m_1, m_2, \ldots, m_J\}$ (m_j: phoneme, J: number of phonemes) and matching phoneme HMM to $\boldsymbol{X}$. Then, the top H words $\boldsymbol{W}_h$ (h = 1, 2, …, H) with maximum log-likelihood, their phonemes $\boldsymbol{M}_h$ and log-likelihoods $T(\boldsymbol{W}_h)$ are extracted.

2.2. *Operation discrimination using the learning-type look-up table*

The user's intension is discriminated using the learning-type look-up table (LUT). The user is instructed to speak some words used in device control, and relationships between control commands, and extracted words $\boldsymbol{W}_h$ and phenomes $\boldsymbol{M}_h$, which include false recognition results, and log-likelihoods $T(\boldsymbol{W}_h)$ are learned to the LUT. Then, the control command corresponding to the extracted word can be selected using the trained LUT.

In the learning stage, the user speaks C words, which is used in device control, some times and top V words ${W^c}_v$ with maximum log-likelihood, and their phonemes ${M^c}_v$ and log-likelihood $T({W^c}_v)$ in H extracted words are corresponded to each discrimination class (c = 1, 2, …, C; v = 1, 2, …, V; $V < H$). In the discrimination stage, extracted phenomes of top U words with maximum log-likelihood in a new H words are used. First, extracted phoneme ${}^{(D)}M_u$ (u = 1, 2, …, U; $U < H$) is compared to phoneme ${}^{(L)}M^c_i$ (i = 1, 2, …, I_c; I_c: number of learning data for class c) of each discrimination class memorized in the learning-type LUT. The coincidence between ${}^{(D)}M_u$ and ${}^{(L)}M^c_i$ is then calculated as follows:

$$s^c_{u,i} = \begin{cases} 1 & ({}^{(D)}M_u = {}^{(L)}M^c_i) \\ 0 & (otherwise) \end{cases}. \qquad (1)$$

A class with a maximum value of r^c representing the average of all $s^c_{u,i}$ values is then taken as the discrimination result. When the values for some classes are same, difference between log-likelihoods $T({}^{(D)}W_u)$ and $T({}^{(L)}W^c_i)$ are used to determine the result.

2.3. *Device control*

The domestic appliances are controlled based on discrimination results using the Bio-Remote. The Bio-Remote consists of a sensor unit to measure biological signals and a main unit to control the target device using the measured signals.

The user can directly select appliances and their control commands with the proposed system. As an example, if

the user wishes to select the command to switch on the TV, speaks "*TV*" command to select the TV menu. Next, the user speaks "*power*" command so that the operator can finally choose the ultimate target – the "Power" menu option. The control instruction corresponding to the previously learned TV power menu item is then sent from the computer to the main unit, and an infrared signal is transmitted.

3. Speech recognition experiment

3.1. *Method*

In order to verify the efficacy of the proposed method, experiments were performed to demonstrate the accuracy of discrimination. The participants are three healthy males; and assuming slurred speech, they are intended to speak with their tongue touching to maxillary central. A directional microphone (audio-technica corp., AT-9942) and an audio processor (ONKYO copr., SE-U33GXV) are used to record voice signals. A number of discrimination class is seven (C = 7), and participant repeat to speak each word 50 times. 50 sets of each class data are separated into 10 learning data sets and 40 discrimination data sets. The parameters used in the experiment are set as f_s = 16 [kHz], N = 3, H = 10, V = 10 and U = 5. The other parameters, K, J, I_c are adjusted based on durations of input voice signals and results of learning procedure. The Julius [10] is used to record and extract the features of each speech, and recognition results using it are compared to the results using the proposed method.

3.2. *Results and discussion*

Figure 2 shows the discrimination rates for each class using the proposed method and Julius. It plots the average discrimination rates for each class while the set of learning data and discrimination data were changed and discriminated at each data set for 10 times. From this figure, the average discrimination rate for all classes using Julius is 3.52 ± 5.08 [%], and those using the proposed system is 94.4 ± 2.53 [%], respectively. These results indicate that the proposed system can accurately discriminate slurred speech, which is difficult to discriminate a large vocabulary speaker independent model, learning false recognition results in advance. Additionally, when duplicative phenomes are extracted, the system can discriminate user's intensions based on the difference of log-likelihood. To realize higher level of discrimination performance, we plan to adjust appropriate discrimination parameters such as number of extracted words U using discrimination.

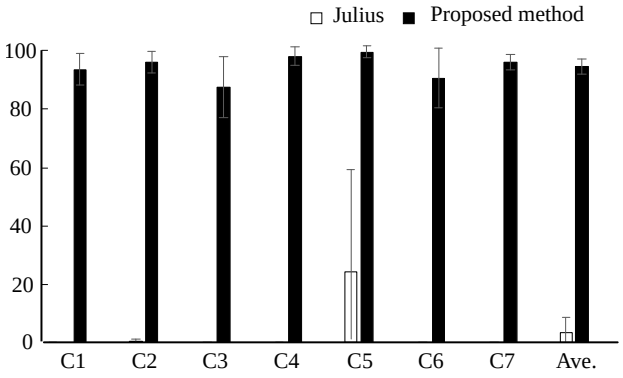

Fig. 2. Comparison of discrimination rates using the proposed method and Julius [10]

4. ECS control experiment

Assuming operation in real life, an experiment was performed using the Bio-Remote with the proposed method introduced. In the experiment, the participant A instructed to (1) turn on the light, (2) turn the TV on, (3) play DVD, (4) stop DVD, (5) turn the TV off, (6) play the audio player and (7) turn off the light. The parameters used in the experiment are as same as the Section 3.

An scene showing the operation of the proposed system is given in Fig. 3, and examples of the experimental results are shown in Fig. 4. This shows input signals, extracted phenomes (corresponding to extracted words with the maximum log-likelihood), discrimination resutls, selected devices and control commands in order from the top. From Fig. 4, the subject spoke ``*shoumei* (Light)", thereby moved to the light operation layer. Then, speaking "*onn* (ON)", the light was turned on after approximate 1.6 [s] (see Fig. 4 (1)). The outcomes here also showed that the subject directly move from the light layer to the TV layer as speaking ``*terebi* (TV)", and DVD was played after the TV was turned on around 7.6 [s] (see Fig. 4 (2)). It was therefore indicated that users' intensions were correctly discriminated from slurred speech to learn features of individual speech and devices were controlled based on discrimination result.

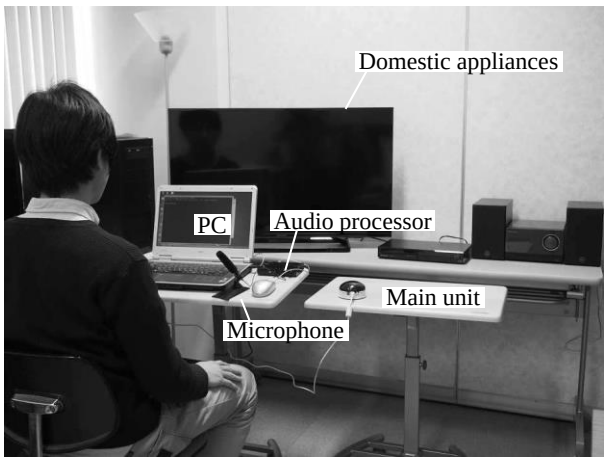

Fig. 3. Operation scene using the proposed system

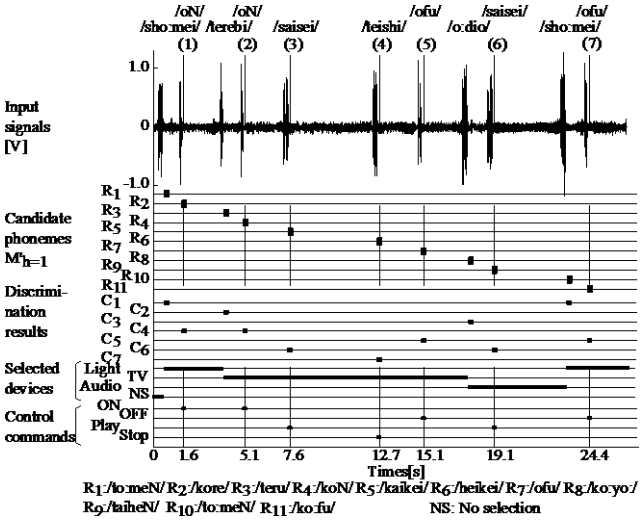

Fig. 4. An example of experimental results

5. Conclusion

This paper proposed a novel speech controlled ECS for patients with dysarthria based on a false recognition results for slurred speech using a large vocabulary speaker independent model. In the experiments performed, the accuracy of speech recognition was evaluated comparing to results using Julius. The results showed that the proposed method was able to discriminate with an accuracy level of 94.4 ± 2.53 [%] for three healthy males. The method was therefore able to correctly discriminate user intensions. Further, operation experiments using the proposed system showed that the subject can voluntarily control domestic appliances.

In future work, we plan to perform operation experiments for patients with dysarthria and validate the effectiveness of the proposed system. In order to reduce the level of stress involved in Bio-Remote operation, we also plan to discuss the discrimination parameters, and to incorporate an online learning method into it.

Acknowledgements

The authors would like to cordially acknowledge and express appreciation to Mr. K. Harada for his assistance in the implementation of this research. This work was supported by JSPS KAKENHI Grant Number JP26330226.

References

[1] Ministry of Health, Labour and Welfare, "Ministry of Health, Labour and Welfare Fact-Finding Investigation of Fisically Disabled," *http://www8.cao.go.jp/shougai/data/datah23/zuhyo09.html* (accessed December 2016).

[2] A. Craig, P. Moses, Y. Tran, P. McIsaac, L. Kirkup, The Effectiveness of a Hands-Free Environmental Control System for the Profoundly Disabled, *Archives of Physical Medicine and Rehabilitation*, **83** (10) (2002) 1455-1458.

[3] X. Gao, D. Xu, M. Cheng, S. Gao, A BCI-based Environmental Controller for the Motion-disabled, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **11** (2) (2003) 137-140.

[4] T. Tsuji, K. Shima, A. Funabiki, S. Shitamori, K. Shiba, O. Fukuda and A. Otsuka, A New Manipulation Method for Environment Control Systems, *The Society of Life Support Technology*, **18** (4) (2006) 5-12 (in Japanese).

[5] T. Shibanoki, G. Nakamura, K. Shima, T. Chin and T. Tsuji, Operation Assistance for the Bio-Remote Environmental Control System Using a Bayesian Network-based Prediction Model, *Proceedings of 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Milan, Italy, 2015), pp. 1160-1163.

[6] Voicecan Co., Ltd, VOICECAN, *http://www.voicecan.ecweb.jp/* (accessed December 2016).

[7] Asahi Kasei Technosystem, Co., Ltd, LIFETACT, *http://www.asahi-kasei.co.jp/ats/hukushi final.html* (accessed December 2016).

[8] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O. Neill and R. Palmer, A Speech-controlled Environmental Control System for People with Severe Dysarthria, *Medical Engineering & Physics*, **29** (5) (2007) 586-593.

[9] A. Lee and T. Kawahara, Recent Development of Open-Source Speech Recognition Engine Julius, *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (Hokkaido, Japan, 2009), pp. 131-137.

Large vocabulary Continuous Speech Recognition Engine, Julius, *http://julius.sourceforge.jp/index.php* (accessed December 2016).