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

Taro Shibanoki, Go Nakamura, Takaaki Chin and Toshio Tsuji
Ibaraki University, Hyogo Rehabilitation Center, Hiroshima University
SHIBANOKI, Taro <taro.shibanoki.ts@vc.ibaraki.ac.jp>

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 words are extracted. Experimental results showed that the method accurately discriminate slurred words (average discrimination rate: 93.9±2.40 [%]), and that participants were able to voluntarily control domestic appliances.

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

1. Introduction

A variety of environmental control systems (ECSs) for disabled and bedridden elderlies have been developed to be self-sufficient and maintain an independent life in recent years. There are a 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, the process of attaching the necessary sensors to the skin is unpleasant and burdensome for the user because it involves the use of paste and medical tape. Additionally, long-time use of the system is difficult due to the effects of changes in skin impedance caused by factors such as perspiration. To overcome these problems, the authors focused on voice signals as an extension input for the Bio-Remote

A number of speech-controlled ECSs 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 using a speaker-independent acoustic model, and devices are controlled based on the results. It is difficult to accurately discriminate the words of patients with dysarthria who have difficulty speaking because the model used in these systems considers only standard adult speech. However, providing training on individual users' voices and using a speaker-dependent model enables accurate discrimination for the speech of such patients [8]. However, as speaker-dependent model training is time-consuming and requires large amounts of data, it may place a burden on the user.

This paper proposes a novel voice signal-based environment control system for patients with dysarthria. The system individual user voice features based on candidate words discrimination using a speaker-independent model, and can discriminate user intensions without large amounts of training data.

2. Speech controlled Bio-Remote

Figure 1 shows an overview of the proposed system, which involves the stages of voice signal measurement and feature extraction, operation estimation, and device control. The details of each stage are outlined below.

2.1. *Voice signal measurement and feature extraction*

Voice signals are recorded using a microphone and digitized using an A/D converter (sampling frequency: 16,000 [Hz]; default sampling rate in Julius [10]]). Mel-frequency cepstral coefficients (MFCCs) are then extracted when an inverse cosine transform is applied to the log power spectra of the sampled signals. The feature vector $\boldsymbol{X}$ used for speech recognition is defined from the low-frequency components of each frame of extracted MFCCs [9].

Next, the output probability P ($\boldsymbol{W}$) of word $\boldsymbol{W} = \{w_1, w_2, \ldots, w_K\}$ (w_k: word, K: number of words) is approximated using word N-gram language model [9]. Additionally, the output probability $P(\boldsymbol{X}|\boldsymbol{W})$ of a feature vector $\boldsymbol{X}$ from $\boldsymbol{W}$ is calculated using an acoustic model. A triphone hidden Markov model (triphone HMM) [9] with which context and time variation can be considered is used to calculate $P(\boldsymbol{X}|\boldsymbol{W})$ with the words $\boldsymbol{W}$ divided into phonemes $m = \{m_1, m_2, \ldots, m_J\}$ (m_j: phoneme, J: number of phonemes) and matching triphone HMM to $\boldsymbol{X}$. Then, the top H words $\boldsymbol{W}_h$ ($h = 1, 2, \ldots, H$) with the maximum log-likelihoods, their phonemes $\boldsymbol{M}_h$ and log-likelihood values $T(\boldsymbol{W}_h)$ are extracted.

2.2. *Operation discrimination using a learning-type lookup table*

The user's intention is discriminated using a learning-type lookup table (LUT). The user is instructed to utter words used in device control, and the relationships between control commands, and the extracted words $\boldsymbol{W}_h$ and phenomes $\boldsymbol{M}_h$ (which include false recognition results) as well as the log-likelihoods T ($\boldsymbol{W}_h$) are input into the LUT. The control command corresponding to the extracted word can be selected using the trained LUT.

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

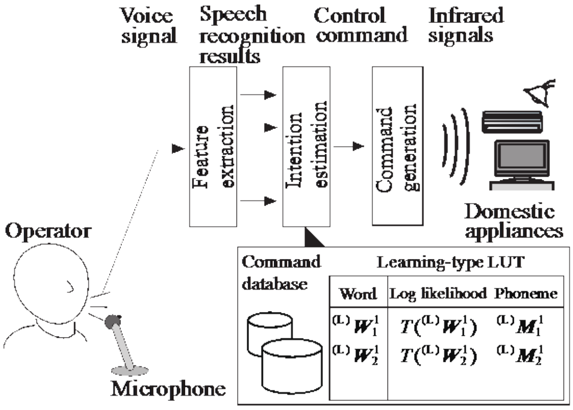

Fig. 1. Overview of the proposed system.

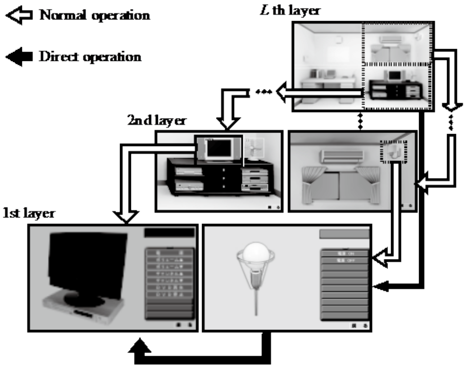

Fig. 2. Layer-based selections for the Bio-Remote

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

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, the difference between log-likelihoods $T({}^{(D)}W_u)$ and $T({}^{(L)}W^c{}_i)$ are used to determine the result.

2.3. *Device control*

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.

TABLE I Relationships between discrimination classes and control commands

Class number	Speech words	Control command
C_1	*shoumei*	Light
C_2	*terebi*	TV
C_3	*o-dhio*	Audio
C_4	*onn*	ON
C_5	*ofu*	OFF
C_6	*saisei*	Play
C_7	*teishi*	Stop

To operate a variety of the domestic appliances, the layer based selections is adopted for the Bio-Remote, and the related selections (control commands, devices and operating area) are grouped and arranged in a layer structure (see Fig. 2). The user can directly select appliances and their control commands with the proposed system. As an example, a user wishing to select the TV switch-on command can say “TV” to bring up the TV menu. Next, the user can say “Power” to 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*

To verify the efficacy of the proposed method, an experiment was performed for determination of discrimination accuracy. The participants are three healthy males and one patient with dysarthria. Three healthy males were instructed to speak with their tongue touching to maxillary central to simulate slurred speech. A directional microphone (Audio-Technica Corp., AT-9942) and an audio processor (ONKYO Corp., SE-U33GXV) were used to record voice signals. There were seven discrimination classes (C = 7, see TABLE I), and participants repeat to speak each word 50 times. The 50 sets of each class data are separated into 10 learning data sets and 40 discrimination data sets. The parameters used in the experiment were set as N = 3, H = 10, V = 10 and U = 5. The other parameters, K, J, I_c were adjusted based on the input voice signal durations and learning procedure results. The Julius [10] was used to record and extract the features of each speech.

3.2. *Results and discussion*

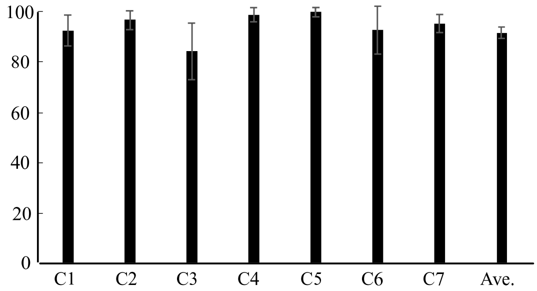

Fig. 3. Discrimination rates using the proposed method.

Figure 3 shows the discrimination rates for each class using the proposed method. 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 the proposed system is 93.9±2.40 [%]. Although speaker-dependent model training is time-consuming and requires large amounts of data, the proposed system can accurately discriminate slurred speech by only learning candidate words (including false recognition results) in advance. Additionally, when duplicative phenomes are extracted, the system can discriminate user intentions based on log-likelihood differences. To enable higher levels of discrimination performance, the authors plan to adjust appropriate discrimination parameters such as the number of extracted words U using discrimination.

4. ECS control experiment

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

Figure 4 shows a scene from the operation of the proposed system, and Fig. 5 shows examples of experimental results (from the top: input signals, extracted phenomes corresponding to extracted words with the maximum log-likelihood, discrimination results, selected devices and control commands). Figure 5 shows that the subject said ``*shoumei* (Light)” to move to the light operation layer (see Fig. 2), and then said “*onn* (ON)”. The operation command (“ON” menu

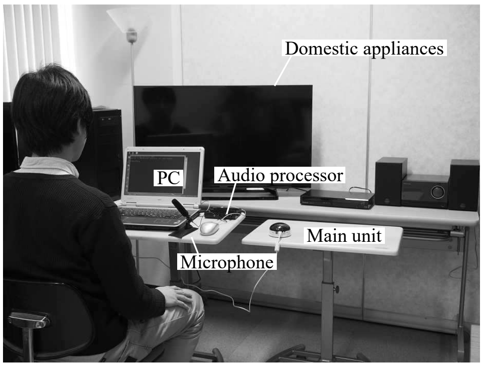

Fig. 4. Operation scene using the proposed system

of the light) was selected and the light was turned on in approximate 1.6 [s]. The outcomes here also show that the participant directly move from the light layer to the TV layer by saying ``*terebi* (TV)", and the DVD was played around 7.6 [s] after the TV was turned on. The user's intentions were thus correctly discriminated from slurred speech, and the devices were controlled based on the results.

5. Conclusion

This paper proposes a novel speech-controlled ECS for patients with dysarthria based on candidate word discrimination using a large-vocabulary speaker-independent model. In the experiments performed, the accuracy of speech recognition was evaluated. The outcomes showed that the proposed method enabled discriminate with 93.9 ± 2.40 [%] for three healthy males and one patient with dysarthria, and therefore supported correct discrimination of user intentions. An operation experiment conducted with the proposed system also showed that users could voluntarily control domestic appliances as intended.

In future work, the authors plan to perform operation experiments for other patients with dysarthria and further evaluate the effectiveness of the proposed system. To reduce the level of stress involved in Bio-Remote operation, the discrimination parameters will be discussed, and an online learning method will be incorporated.

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.

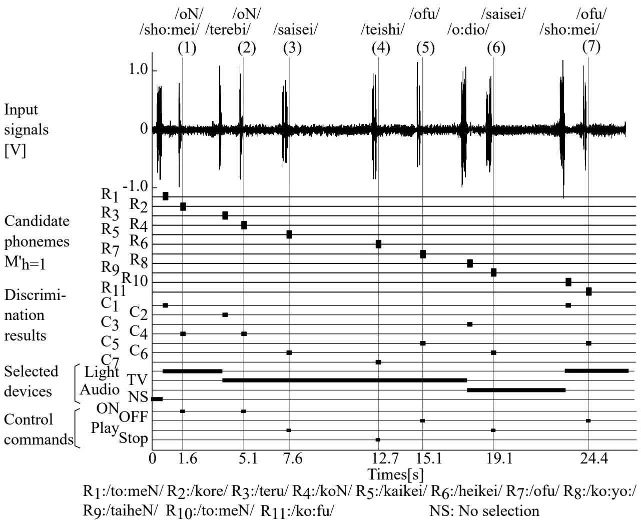

Fig. 5. An example of experimental results

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