ATLANTIS PRESS
Journal of Robotics, Networking and Artificial Life, Vol. 5, No. 2 (September 2018) 135-138

A Training Method for the Speech Controlled Environmental Control System Based on Candidate Word Discriminations

Taro Shibanoki[*], Masaki Watanabe[*], Go Nakamura[†], Takaaki Chin[†], Toshio Tsuji[‡]
[*]*Ibaraki University, 4-12-1, Nakanarusawacho, Hitachi, 316-8511, Japan*
[†]*Hyogo Rehabilitation Center, 1070 Akebonocho, Nishi-Ku, Kobe,651-2181, Japan*
[‡]*Hiroshima University, 1-4-1, Kagamiyama, Higashi-Hiroshima,739-8527, Japan*
e-mail: taro.shibanoki.ts@vc.ibaraki.ac.jp, 14t4070a@vc.ibaraki.ac.jp, g_nakamura@assistech.hwc.or.jp, chin@assistech.hwc.or.jp, tsuji@bsys.hiroshima-u.ac.jp
http://bs.cis.ibaraki.ac.jp

Abstract

This paper proposes a concept of a training system for the speech controlled environmental control system: Bio-Remote based on candidate word discriminations. The proposed system can provide three-types of voice signal training: (1) volume, (2) tempo/timing and (3) candidate word which are important for accurate speech recognition based on false recognition results. During the training, such three kinds of features are extracted from measured voice signals and visually and auditory fed back to the user in real time. This allows the user to train speech abilities even if false recognition results are extracted because of slurred speech. The efficacy of the proposed system was demonstrated through training experiments for slurred speech conducted with healthy participants. The results showed that the proposed system was capable for the training of speech abilities.

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

1. Introduction

The number of disabled people in Japan continues to increase annually and stands at 1.76 million. The population of severely disabled people in particular was around 760,000, and such patients associated with a speech disability reached 81,000 [1].

Against such a background, many speech-controlled environmental control systems (ECSs) have been developed [2], [3]. However, it is difficult for patients with dysarthria to use such systems, since the models used in these systems considers standard adults' speech. Although some studies have investigated the use of speaker-dependent models to support the learning of individual users' voices [4], and the authors' research group also proposed the voice signal-based manipulation method for ECS based on candidate word discrimination [5], fluctuating speech makes discrimination difficult.

The speech training system can support the recovery of as much of a user's speech as possible, and it has been widely discussed as motivation and long-term experiences for users [6], [7]. However, it can be difficult to fully recover verbal functioning because of individual differences and degrees of disability.

This paper proposed a training system for a speech-controlled environmental control system based on candidate word discrimination that can acquire the skill of fixed speech. After the training, the user's intention can be accurately discriminated, even if the user cannot fully recover his/her verbal functioning.

2. Speech Training System Based on Candidate Word Discrimination

Figure 1 shows the structure of the proposed speech training system based on candidate word discrimination. The proposed system provides training for patients with dysarthria to speak the same way every time, even with slurred speech. This training can be applied to control the training of a voice-controlled environmental control system [5] with slurred speech.

The proposed speech training system consists of a PC with a feedback display, audio processor, and microphone. During the speech training, the display

provides the extracted features of voice signals and current status of the users' abilities to improve their speech skills. The details of the system are described in the following subsections.

2.1. *Voice signal processing*

The structure of the voice signal processing is shown in Fig. 1. First, the amplitude and timing information of voice signals are extracted, and discrimination results are then obtained using the candidate words/phenomes $\boldsymbol{W}_h$/$\boldsymbol{M}_h$ and the log-likelihoods $T(\boldsymbol{W}_h)$ with the candidate word discrimination method [5].

2.1.1. *Extraction of voice signal features*

Voice signals are recorded using a microphone and sampled at 16 [kHz]. The amplitude information $v(t)$ of the measured voice signals with full-wave rectification and low-pass filtering (cut-off frequency: 1 [Hz]) is obtained based on the gains of the amplifier and microphone input levels.

Feature vector $\boldsymbol{X}$ used for speech recognition is then defined as the low-frequency components of Mel-frequency cepstrum coefficients (MFCCs) for each frame, and the output probabilities $P(\boldsymbol{X}|\boldsymbol{W})$ of a feature vector $\boldsymbol{X}$ from word $\boldsymbol{W} = \{w_1, w_2, \ldots, w_K\}$ (w_k: word, K: number of words) are calculated using an N-gram model and phoneme-hidden Markov model (phoneme HMM) dividing the words $\boldsymbol{W}$ into phonemes $m = \{m_1, m_2, \ldots, m_J\}$ (m_j: phoneme, J: number of phonemes) and matching phoneme HMM to $\boldsymbol{X}$. Subsequently, the top H words $\boldsymbol{W}_h$ ($h = 1, 2, \ldots, H$) with the maximum log-likelihood, their phonemes $\boldsymbol{M}_h$, and log-likelihoods $T(\boldsymbol{W}_h)$ are extracted using Julius [8].

2.1.2. *Intention estimation using candidate word discrimination*

The user's intention is discriminated using the learning-type look-up table (LUT) [5]. The user is instructed to utter C words (corresponding to the control commands of domestic appliances) multiple times, and top V words ${}^{c}W_v$ with their maximum log-likelihood and their phonemes ${}^{c}M_v$ and log-likelihood $T({}^{c}W_v)$ in H extracted words are corresponded to each discrimination class ($c = 1, 2, \ldots, C$; $v = 1, 2, \ldots, V$; $V < H$) in the learning stage. In the discrimination stage, a new set of H words are extracted, and the phenomes ${}^{(D)}M_u$ ($u = 1, 2, \ldots, U$; $U < H$) of top U words with their maximum log-likelihood are compared

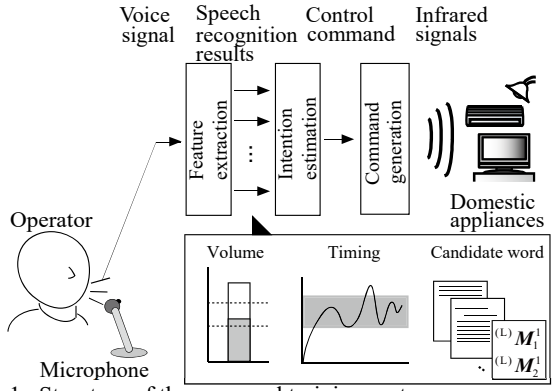
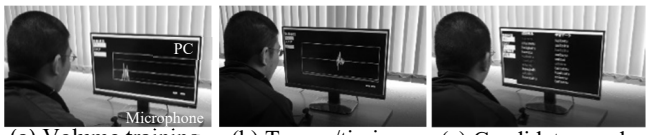


Fig. 1. Structure of the proposed training system



(a) Volume training (b) Tempo/timing training (c) Candidate word speech training

Fig. 2. Scenes from the training

to phoneme ${}^{(L)}M^{c}_i$ ($i = 1, 2, \ldots, I_c$; I_c: number of learning data for class c) of each discrimination class memorized in the learning-type LUT. The coincidence $s^{c}_{u,i}$ between ${}^{(D)}M_u$ and ${}^{(L)}M^{c}_i$ is then calculated, and a class with a maximum value of r^{c} representing the average of all $s^{c}_{u,i}$values (the number of coincident phonemes) is then taken as the discrimination result. To disambiguate discrimination, the difference between log-likelihoods $T({}^{(D)}W_u)$ and $T({}^{(L)}W^{c}_i)$ is used to determine the result when the same values for some classes are obtained.

2.2. *Speech training for candidate word discrimination method*

For speech therapy, the treatment of articulation, prosody and pitch range, speech rate, vocal volume, or resonance is important [6]. The proposed speech training is therefore composed of three stages (see Fig. 2), as explained in the following subsections.

Before the training, to make the learning data sets used in each training, the trainee is instructed to utter C words T times, and the maximum/minimum and average volumes of each word are determined. The durations of $C \times T$ words $[{}^{\text{start}}D^{c}_t, {}^{\text{end}}D^{c}_t]$ used in the timing/tempo control training are also determined using Julius [8].

2.2.1. *Volume control training*

In this training, the trainee practices adjusting the vocal volume level. During the training, the amplitude information extracted from a measured voice signal is presented on the display with the desired values (predetermined maximum/minimum and average). The trainee controls his/her voice signal so the extracted amplitude information follows the average value and falls within the min./max. range. The training result is evaluated using the following equation.

$$S_t = \begin{cases} 100 - (V_{\mathrm{ave}} - v(\mathrm{t})) & (V_{\min} \le v(\mathrm{t}) \le V_{\max}) \\ 0 & (V_{\min} > v(\mathrm{t}), v(\mathrm{t}) > V_{\max}) \end{cases}. \quad (1)$$

The closer the amplitude information is to the average, the higher the score, and if it exceeds the min./max. values, the score becomes zero. The average of the S_t value of speech duration is output at the end of each trial.

2.2.2. *Tempo/timing control training*

For accurate discrimination using the candidate word discrimination method, it is also important to control the timing and time from the start to the end of speech. In this training, the trainee controls the tempo/timing of his/her speech. During the training, the stored voice signals of each word are randomly shown in the display. The trainee is instructed to regulate his/her speech duration and timing according to the pre-specified timing shown in waveforms. The system evaluates the ratio of extracted speech duration to pre-specified duration in this training.

2.2.3. *Candidate word speech training*

Candidate word speech training is conducted so the trainee speaks approximately the same way every time. The candidate phonemes for each discrimination word and extracted trainee's phonemes are shown in the display during this training. The trainee practices to control his/her speech so similar candidate phonemes are extracted. The score of this training is defined as a ratio of the number of complete/ambiguous coincidence in candidate and extracted phonemes:

$$S_h = \begin{cases} (100 - S_{\mathrm{th}})/D_{\mathrm{cmp}} & (L({}^{(L)}M_i^c, M_h) = 0) \\ (100 - S_{\mathrm{th}})/D_{\mathrm{amb}} & (0 < L({}^{(L)}M_i^c, M_h) \le L_{\mathrm{th}}), \\ 0 & (L_{\mathrm{th}} < L({}^{(L)}M_i^c, M_h) \end{cases} \quad (2)$$

where $L(\cdot)$ represents Levenshtein distance.

3. Training Experiments

3.1. *Method*

To verify the efficacy of the proposed training system, training experiments were performed with three healthy males (subjects A–C, 22.3 $\pm$ 1.15 [year]). In the experiments, participants were instructed to speak with their tongue touching the maxillary central to simulate slurred speech. The parameters used in the experiments were set as C = 7, T = 10, H = 10, V = 10, U = 5, D = 10, $L_{\mathrm{th}} = 1$, D_{cmp} = 10, D_{amb} = 500, and S_{th}=50. The other parameters, K, J, and I_c, were adjusted based on the input voice signal durations and learning procedure results. Ten sessions of each training stage were performed in the training experiments, and discrimination experiments were also performed before and after training to verify the effectiveness of the proposed training method. In the discrimination experiments, participants were asked to utter each word three times without feedback.

3.2. *Results and discussion*

Figure 3 shows examples of experimental results. From this figure, the training scores are stable as the number of sessions increased, and the average discrimination rates before and after training have relatively high accuracy. Although participants simulated slurred speech, they could utter each word the same way from the beginning of the training.

Therefore, other training experiments were performed so the participants could mimic other participants' speech. In the experiments performed, Sub. A trained using Subs. B and C's learning data sets. These experimental results are shown in Fig. 4, and it is confirmed that Sub. A cannot follow other participants' speech during 10 training sessions. An additional 10 training sessions were therefore performed with words, the score of which was below 40. In the latter 10 sessions, each score was gradually increased as the number of training sessions increased. The average discrimination rates before and after 20 training sessions were 61.90 $\pm$ 48.56 [%] and 80.95 $\pm$ 39.27 [%], respectively, and the significant difference before and after training was confirmed at a level of 1 [%]. These outcomes indicated that the proposed training system was capable of speech training.

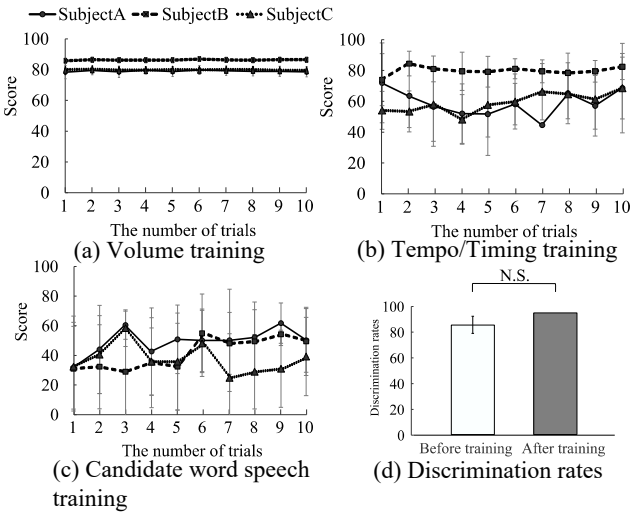


Fig. 3. Experimental results

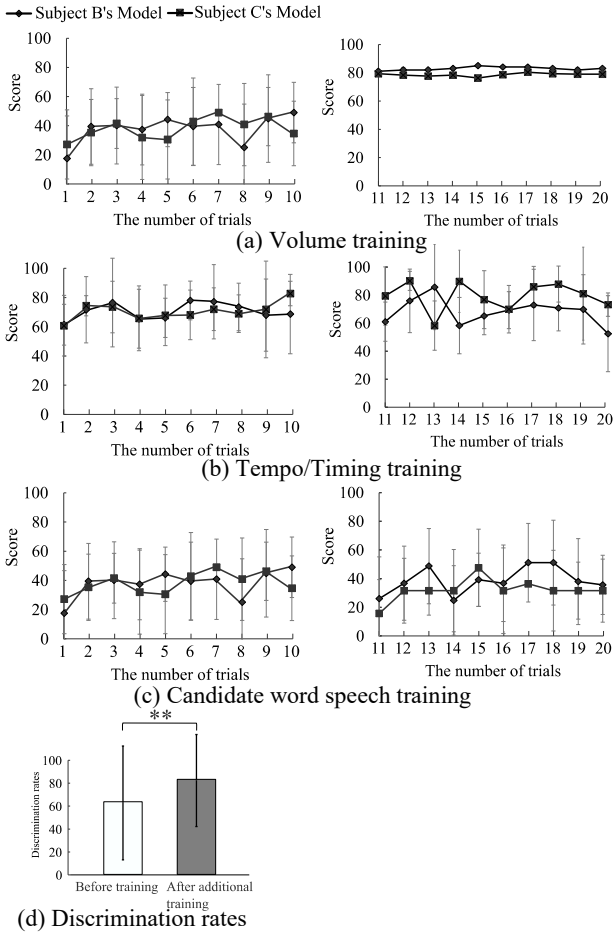


Fig. 4. Experimental results using the other subjects' learning data sets (Sub. A).

4. Conclusion

This paper proposes a speech training system for the voice signal controlled ECS based on candidate word discriminations. The proposed training system provides three types of speech training that are important to speak in the same way every time. In the training experiments, it could be confirmed that the trainees' speech skills were gradually improved through training using the proposed system.

In future work, the authors plan to perform training experiments for patients with dysarthria and establish an online tuning method of training levels for each stage.

Acknowledgements

This work was partially supported by JSPS/MEXT KAKENHI Grant Numbers 17K12723 and 26330226.

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/whitepaper/h25hakusho/zenbun/furoku_08.html* (accessed December 2017).
2. Glamo Inc., iRemocon Wi-Fi, *http://i-remocon.com/aboutiremoconwifi/* (accessed December 2017).
   Nature Inc., Nature Remo, *http://nature.global/* (accessed December 2017).
3. 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, pp. 586-593.
4. 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.
5. J. Mühlhaus, H. Frieg, K. Bilda,and U. Ritterfeld, Game-Based Speech Rehabilitation for People with Parkinson's Disease, *UAHCI 2017, Part III, LNCS10279* (M. Antona and C. Stephanidis Eds), 2017, pp. 76-85.
6. J. Tamplin, A Pilot Study into the Effect of Vocal Exercises and Singing on Dysarthric Speech, *Neurorehabilitation*, **23,** 2008, pp. 207-216.
7. Large vocabulary Continuous Speech Recognition Engine, Julius, *http://julius.sourceforge.jp/index.php* (accessed December 2017).