T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title	HMM Based Continuous EOG Recognition for Eye-input Speech Interface
Authors	Fuming Fang, Takahiro Shinozaki, Yasuo Horiuchi, Shingo Kuroiwa, Sadaoki Furui, Toshimitsu Musha
Citation	Proceedings of INTERSPEECH 2012, , , , Tue.P3c.04
Pub. date	2012, 9
Copyright	(c) 2012 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

Powered by T2R2 (Science Tokyo Research Repository)

HMM Based Continuous EOG Recognition for Eye-input Speech Interface

Fuming Fang[1], Takahiro Shinozaki[1], Yasuo Horiuchi[1], Shingo Kuroiwa[1],
Sadaoki Furui[2], Toshimitsu Musha[3]

[1]Division of Information Sciences, Chiba University, Chiba, Japan
[2]Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan
[3]Brain Functions Laboratory Inc., Kanagawa, Japan

Abstract

To provide an efficient means of communication for those who cannot move muscles of the whole body except eyes due to amyotrophic lateral sclerosis (ALS), we are developing a speech synthesis interface that is based on electrooculogram (EOG) input. EOG is an electrical signal that is observed through electrodes attached on the skin around eyes and reflects eye position. A key component of the system is a continuous recognizer for the EOG signal. In this paper, we propose and investigate a hidden Markov model (HMM) based EOG recognizer applying continuous speech recognition techniques. In the experiments, we evaluate the recognition system both in user dependent and independent conditions. It is shown that 96.1% of recognition accuracy is obtained for five classes of eye actions by a user dependent system using six channels. While it is difficult to obtain good performance by a user independent system, it is shown that maximum likelihood linear regression (MLLR) adaptation helps for EOG recognition.

Index Terms: electrooculogram, hidden Markov model, amyotrophic lateral sclerosis, continuous speech recognition, maximum likelihood linear regression

1. Introduction

Amyotrophic lateral sclerosis (ALS) is an intractable motor neuron disease that causes significant decrease in the mass of muscles of the whole body [1]. The patients eventually lose the ability of breathing due to paralysis of respiratory muscles and are required to use mechanical ventilation, which makes it impossible to produce a speech sound. Other popular communication methods such as hand writing are also disabled and only eye motion is specifically kept. On the other hand, there is usually no damage to the brain and consciousness is clear. Therefore, establishing a communication method through eye motions is essential for the patients.

A conventional communication method uses a transparent alphabet board [2]. A patient and a carer face each other through the board. The carer slowly moves the board and reads a character that he thinks the patient is gazing at. The patient sends pre-decided yes or no sign for the read character. By repeating this, arbitrary sentences can be transmitted. However, a problem is that it requires the cooperation of a skillful carer. Therefore, several automated systems have been developed using a computer. The basic idea is to use a software keyboard projected on a computer display with a mechanism that allows patients to control the keyboard with their eyes.

There are mainly two types of such systems. One is based on a sweeping cursor interface, in which a cursor on a software keyboard repeatedly moves from one side to another. The patient selects a column of the keyboard by sending a sign when the cursor crosses the column. Once a column is selected, then a row is selected based on a similar process. This approach is simple and works by detecting only a single sign. However, its input speed is slow since the long waiting period to wait for the cursor is indispensable. As a detection method of the eye sign, electrooculogram (EOG), which is a weak electrical potential caused by eyes, has been used.

The other is based on key gazing interface, in which a key is selected based on detecting the patient's gazing point on the keyboard. The success of this approach depends on the precision of the gazing point detection. To cope with low precision, incremental precision method has been proposed in which a key is selected by first specifying a region containing several keys. Then, the specified region is automatically magnified and the target key is specified. While it does not need a waiting period unlike the sweeping cursor interface, it still requires a certain pause period for the gazing point detection. To detect the gazing point, reflection of infra-red rays by eyes is used. However, there is a health concern about exposing infra-red rays to eyes. Another detection method uses a image camera and estimates eye directions by image processing. The input speed measured for a system with this interface is around 3.1 second per character for Japanese Hiragana alphabet [3].

In this research, we propose a speech synthesis interface that is based on eye motion recognition using EOG input focusing on realizing efficient means of speech conversation having high interactivity. Compared to the conventional interfaces, the proposed interface does not use

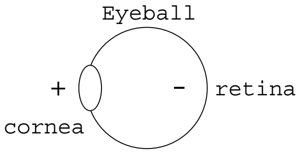

Figure 1: Corneo-retinal potential.

a software keyboard and no display is needed. Therefore, patients do not need to face a large display installed in their room nor wear a heavy head mounted display. There is also a potential advantage for the input speed since there is no intrinsic waiting period.

The organization of this paper is as follows. In Section 2, the basics of EOG are briefly reviewed. In Section 3, the proposed system is explained. Section 4 describes a database used to train and evaluate our system. Experimental conditions are described in Section 5 and the results are shown in Section 6. Summary and future works are given in Section 7.

2. Electrooculogram

There is electrical potential in the eyeball between the cornea and the retina as shown in Figure 1, which is called corneo-retinal potential (CRP). The cornea side has positive charge and the retina side has negative charge. CRP is observed as EOG through electrodes attached on the skin around eyes. EOG changes according to movements of the eye. Therefore, it can be used to estimate eye movements. The magnitude of EOG is around 290 µV/rad to 1100 µV/rad. EOG based eye motion detection works even when eyes are closed. It is reported that there is no big difference between EOG of able-bodied persons and ALS patients [4]. EOG detector is non-invasive and generally simpler than that for electroencephalogram (EEG) as EOG has larger signal magnitude.

3. Proposed system

Figure 2 shows an overview of the proposed eye-input speech interface. It consists of an EOG input module, a recognition module, and an output module. The EOG input module detects EOG signal using biopotential electrodes and digitizes it. The recognition module recognizes eye motions using a hidden Markov model (HMM) based decoder and maps eye motions to pronunciation symbols. The output module takes the sequence of pronunciation symbols and synthesizes speech sound. The synthesized speech is then output via a loudspeaker. In our implementation, the speech synthesizer is also based on HMM. The details of the EOG detection and recognition are described in the following subsections.

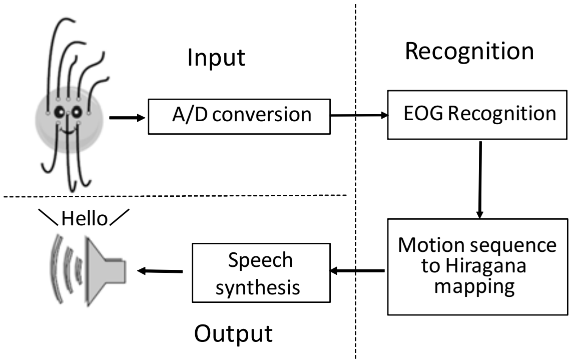

Figure 2: System overview.

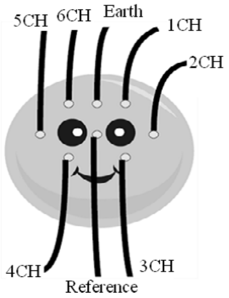

Figure 3: Arrangement of electrodes.

3.1. EOG detection

To detect EOG signal, nonpolarizable biopotential electrodes are used. In this research, eight electrodes are arranged around eyes as shown in Figure 3. Among them, one is ground electrode, another is reference electrode, and the others are measurement electrodes. EOG signals are obtained through the six measurement electrodes relative to the reference electrode. Two of the measurement electrodes (1ch and 6ch) are attached above eyes, another two (3ch and 4ch) are attached below eyes, and the others (2ch and 5ch) are attached at left and right sides of eyes.

Figure 4 shows an example of EOG signal that was observed when eyes were moved in the order of up, down, left, and right. As can be seen, EOG reflects the eyes motions. However, it is rather noisy and not trivial to accurately estimate eye motions based on the signal.

3.2. EOG recognition

An isolated eye sign recognition system using HMM has been proposed by Bulling et al. [5]. The EOG recognizer used in our system can be regarded as an extension of that system, and can recognize a sequence of eye motions by applying continuous speech recognition techniques. The recognizer is based on T^3 WFST decoder which supports live decoding where partial output is obtained when a prefix of recognition hypotheses is determined without waiting for the end of the input [6]. This is important for interactive systems.

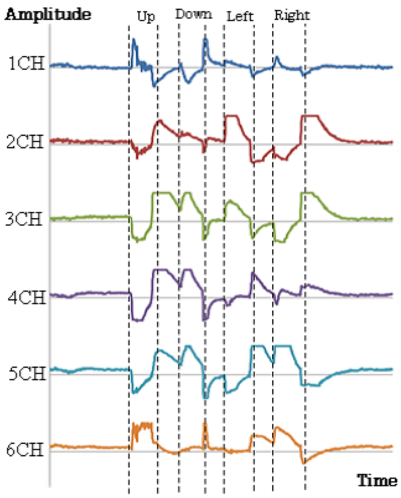

Figure 4: Example of EOG signal.

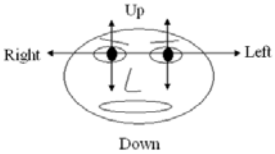

Figure 5: Definition of eye motions.

In our system, five eye motions are used as units for recognition, which correspond to phones in a speech recognition system. Figure 5 shows the motions, which are up, down, left, right, and center. In order to express arbitrary Japanese pronunciations, Hiragana symbols expressing syllables are treated as words in the recognition system. More precisely, 48 basic symbols are each expressed by a sequence of three motions of the five categories, and other derivative symbols are each expressed by an escape character and a corresponding original symbol. To make the recognition easier, a constraint is introduced that the center eye motion must exist between motion sequences expressing two symbols. For the recognition, a six dimensional feature vector is simply formed by gathering the six channels of EOG signal, and a sequence of the vectors is input to the decoder.

4. Database

Since there is no existing database that can be used for our purpose, we have recorded data by ourselves. The data were recorded from three able-bodied persons who were male and were in their twenties. They were asked to move their eyes following indications given by voice. The indication speed was about 0.7 second per motion. This was relatively slow and the speed could be increased with the practice of subjects.

We defined a set of eye motions that consisted of eight types of sequences such as five repetitions of “up” and “down”. In total, the set consisted of 64 motions. The first subject, who is referred to as subject A, had two sessions of recording performed on different days. The amount of data recorded in the first session was 10 sets and that of the second session was 40 sets. The second subject, referred to as subject B, had three sessions on different days, in which 15, 10, and 25 sets of data were recorded, respectively. The last subject, subject C, had a session in which 15 sets of data were recorded.

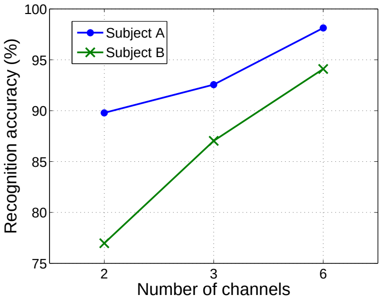

Figure 6: Number of channels and recognition accuracy.

For a safety reason (electrodes attached to the skin) as well as a technical reason (to avoid electrical noise), the recording system was operated using batteries without connecting to outlets. As for the electrodes, silver-silver chloride electrodes were used, which were fixed on the skin with conductive paste using a tape. The sampling frequency for the A/D conversion was 100 Hz. In this paper, we focus on recognition of the five classes of eye motions.

5. Experimental setup

Each of the five eye motions was modeled by a four state 16 mixture HMM. Recognition experiments were performed both in user dependent and independent conditions. The effect of session variability was also evaluated. Experiments were performed using HTK [7].

6. Experimental results

Figure 6 shows the number of channels and recognition accuracy. The result was obtained by five-fold cross-validation using 50 sets of data per subject and using two channels (1ch and 2ch), three channels (1ch, 2ch, and 3ch), and all of the six channels. While two channels theoretically suffice to recognize two-dimensional motions, it can be seen that higher recognition performance is obtained by using more channels compensating for the noisy signal. When six channels were used, the averaged accuracy was 96.1%. The rest of the experiments were performed using the six channels.

Table 1 shows session independent recognition accuracy using 25 and 40 sets of data as training set. As the test set, 10 sets of data were used, which was recorded in

Table 1: User-dependent EOG recognition results with session-independent condition.

Training data	Subject A	Subject B	Average
25 sets	92.8	89.1	91.0
40 sets	93.8	90.9	92.4

Table 2: User-independent EOG recognition results.

Train	Test	W/o adapt	With adapt
Subject B+C	Subject A	79.2	80.8
Subject A+C	Subject B	74.5	76.7
Average		76.9	78.8

a session different from the training set. This is a realistic condition for the proposed system considering session variability. That is, variations can arise in EOG signal recorded in different sessions because of physiological conditions of the subjects, and because it is impossible to arrange the electrodes exactly on the same place on the skin. The averaged recognition accuracy was 91.0% when 25 sets of training data were used. By increasing the amount of training data to 40 sets, the accuracy was improved to 92.4%.

Finally, we have evaluated EOG recognition with a user independent condition. To recognize subject A's 10 sets of data, a model was trained using 50 sets of data from subject B and 15 sets of data from subject C. Similarly, to recognize subject B's 10 sets of data, 50 sets of subject A's data and 15 sets of subject C's data were used. Table 2 shows the result. As can be seen, the performance decreases substantially compared to that of the user dependent systems. The averaged accuracy was 76.9%. To improve the performance, maximum likelihood linear regression (MLLR) [8] based supervised user adaptation was tested using 5 sets of data. The results are shown in the right hand side column of the same table. With the adaptation, 78.8% of averaged accuracy was obtained. This result indicates MLLR is effective not only for speech recognition but also for EOG recognition. However, the accuracy is still low and further improvement is required to make the user independent system practical. Although, it is not necessarily required for this application to make the system user independent since the cooperation of users is expected.

7. Summary and future work

We have proposed a speech interface that is based on EOG input to provide a means for speech based conversation for ALS patients. The system works by recognizing eye motions based on EOG using HMM based decoder that supports live decoding. The combinations of eye motions are mapped to pronunciation symbols, which are then converted to speech sound. We performed pioneering continuous EOG recognition experiments for several conditions. When six channels were used in session dependent cross-validation condition, 96.1% accuracy was obtained. In session independent experiments, accuracies for user dependent and independent conditions were 92.4% and 76.9%, respectivelly. MLLR adaptation was useful to improve the accuracy in the user independent condition. Future work includes integrating a language model to the recognizer. It will be effective to improve the coding to map motions to pronunciations so that more accurate and faster input is realized. For example, error correcting codes can be adopted. From a user's point of view, the code needs to be easily remembered, which has been pointed out by a subject who has evaluated our preliminary system incorporating a speech synthesizer. Our future work includes objective and subjective evaluation of the total system. It is also interesting to evaluate how the users can adapt to the system to improve the accuracy and the input speed.

8. Acknowledgements

This research has been supported by Kurata Grants from the Kurata Memorial Hitachi Science and Technology Foundation. Part of this research has also been supported by KAKENHI (23650068) and by KAKENHI (21300060).

9. References

[1] T. Kihira, S. Yoshida, M. Hironishi, H. Miwa, and T. K. K Okamato, "Changes in the incidence of amyotrophic lateral sclerosis in wakayama," *Amyotroph Lateral Scler Other Motor Neuron Disord*, vol. 6, no. 3, pp. 155–163, 2005.

[2] S. Söderholm, M. Meinander, and H. Alaranta, "Augmentative and alternative communication methods in locked-in syndrome," *J Rehabil Med*, vol. 33, no. 5, pp. 235–239, 2001.

[3] S. Handa and Y. Ebisawa, "Head-mounted display with eye-gaze detection function for the severely disabled," *The Journal of The Institute of Image Information and Television Engineers*, vol. 63, no. 5, pp. 685–691, 2009.

[4] T. Ohya, "Research of the character input for communication tool by voluntary blinks utilizing EOG," The Japan ALS Association, Tech. Rep., 2007, (in Japanese).

[5] A. Bulling, D. Roggen, and G. Tröster, "Wearable EOG goggles: Seamless sensing and context-awareness in everyday environments," in *Journal of Ambient Intelligence and Smart Environments*, vol. 1, no. 2, 2009, pp. 157–171.

[6] P. R. Dixon, D. A. Caseiro, T. Oonishi, and S. Furui, "The titech large vocabulary wfst speech recognition system," in *Proc. IEEE ASRU*, 2007, pp. 443–448.

[7] S. Young *et al.*, *The HTK Book*, Cambridge University Engineering Department, 2005.

[8] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. Eurospeech*, 1995, pp. 1155–1158.