Decoding Silent Speech in Japanese from Single Trial EEGS: Preliminary Results


1Knowledge Discovery Research Laboratories, NEC Corporation, Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa, Japan
2Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Kawazu, Iizuka, Fukuoka, Japan
3Japan Tobacco Inc., Tokyo, Japan
4SCSK Corporation, Osaka, Japan
5Hitachi Systems Ltd., Tokyo, Japan
6Medical Solutions Division, NEC Corporation, Tokyo, Japan
7Future University Hakodate, Tokyo, Japan
8Yokohama Gakuen University, Sapporo, Japan

Abstract

We propose a new scheme for speaker-dependent silent speech recognition systems (SSRSs) using both single-trial electroencephalograms (EEGs) scalp-recorded and speech signals measured during overtly and covertly speaking “janken” and “season” in Japanese. This scheme consists of two phases. The learning phase specifies a Kalman filter using spectrograms of the speech signals and independent components (ICs), whose equivalent current dipole source localization (ECDL) solutions were located mainly at the Broca’s area, of the EEGs during the actual speech. In case of the “season” task, the speech signals were transformed into vowel and consonant sequences, and these relationships were learned by hidden Markov model (HMM) with Gaussian mixture densities. The decoding phase predicts spectrograms for the silent “janken” and “season” using the Kalman filter with the EEGs during the silent speech. For the silent “season”, the predicted spectrograms were inputted to the HMM, and which “season” was silently spoken was determined by the maximal log-likelihood among each HMM. Our preliminary results as training steps are as follows: the silent “jankens” were correctly discriminated; the silent “season”-HMMs worked well, suggesting that this scheme might be applied to the discrimination between all the pairs of the hiragana.

Keywords: Silent speech recognition system; Electroencephalogram; Kalman filter; Spectrogram; Independent component analysis; Equivalent current dipole source localization; Hidden Markov model; Gaussian mixture density

Introduction

The decipherment of human thought from brain activity, without recourse to speech or action, is one of the most attractive and challenging frontiers of modern science. In particular, silent speech recognition systems (SSRSs) enable speech communication to be needed when an audible acoustic signal is unavailable [1]. In addition to “physical” SSRSs [2-5], in the “electrical” ones, articulation may be inferred from actuator muscle signals or predicted using command signals obtained directly from the brain. Especially, the latter could be speech synthesis for individuals with severe communication impairments. Electroencephalography (EEG) recorded during speech production attempts have increasingly yielded the decoding of phonemes and words [6,7] and artificial speech synthesizers [8,9]. However, the SSRSs using non-invasively recorded brain activity, such as scalp-recorded EEGs [10-14], functional magnetic resonance imaging (fMRI) [15] and functional near infrared spectroscopy (fNIRS) [16], had been still in the experimental stage, and limited almost to vowel recognition. Therefore, we propose a new scheme for a speaker-dependent SSRS using single-trial scalp-recorded EEGs for silent vowel recognition, and generalize to consonant one in Japanese. The scheme consists of two phases (learning and decoding ones). In order to exemplify this scheme, we carried out two experiments (Experiments I and II).

Materials and Methods

Subjects, tasks and electrical recordings

Ten healthy student volunteers (two females and eight males; mean age: 23.7 ± 1.42 years) participated in Experiment I, whose procedures were approved by the Ethics Committee for Human Subject Research, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology. Informed consents were obtained from all the subjects in writing for the procedures prior to the experiment. All the subjects were right-handed according to the Edinburgh inventory [17]. The subjects were requested to speak “rock”, “paper” or “scissors” (/ɡʊ/ , /pɑː/ or /tʃiː/ in English pronunciation of Japanese, respectively) into a microphone (MS-STM87SV, ELECOM CO., LTD., Japan) in the learning phase or to silently speak it in the decoding phase, according to visual cues. After the subjects gazed for 3 s a point presented at the center of a monitor 62 cm away from the subjects, a line drawing of a hand indicating “rock”, “paper” or “scissors” was presented for the next 3 s. Only the fixation point was presented for the next 3 s. Then, when the point disappeared, the subjects overtly or covertly spoke “rock”, “paper” or “scissors” corresponding to the line drawing presented just before (Figure 1). The line drawings were randomly presented ten times for each janken. Nineteen active electrodes (AP-C100-0155, DIGITEX LAB. CO., LTD., Japan) were affixed to the scalp according to the International 10-20 system. Additive six channels were included for electromyograms (EMGs) and electrooculograms (EOGs), so that

*Corresponding author: Yamazaki T, Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Kawazu, Iizuka, Fukuoka, Japan, Tel: +81-948-29-7818; E-mail: t-yamazki@bio.kyutech.ac.jp

Received August 18, 2015; Accepted August 27, 2015. Published August 31, 2015


Copyright: © 2015 Yamaguchi H, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Independent EEG sources obtained by ICA are dipolar [24]. ECDL was applied to the reconstructed EEGs, namely the projection of each of the rest ICs on the scalp surface by the deflation procedure, using “SynaCenterPro” (PC-based commercial software for multiple ECDL) (NEC corporation). This software estimates unconstrained dipoles at any timepoint [25], using the three-layered concentric sphere head model by the nonlinear optimization methods [26]. An unconstrained dipole was estimated at any timepoint with maximal peak or trough in the EEGs reconstructed by the deflation procedure for each IC.

Here, we searched for appropriate and reliable dipole solutions, by selecting localization results only with goodness of fit (GOF) of more than 90% and with the simplified confidence limits (CLs) of less than 1 mm, by restricting to the results with no drastic change in the brain sites where the unconstrained dipoles are located at least twenty successive instants including the peak or trough, and by excluding the ECDL results localized to the cerebral ventricles and the corpus callosum.

Anatomical labeling of the brain where ECDs were located, using the Japanese brain atlas for a single subject, was automatically carried out in the following: each subject’s MRI was transformed into the atlas, then the estimated ECDs were projected onto the atlas by this non-linear transformation, and finally anatomical labels on the atlas were
Kalman filter

Next, according to the hypothesis, assumed in Directions Into Velocities of Articulators (DIVA) model [8,29], that neurons in the left ventral premotor cortex present intended speech sounds in terms of formant frequency trajectories and projections from these neurons to the primary motor cortex transform the intended formant trajectories into motor commands to the speech articulators, the relationship between the extracted ICs and spectrograms of the speech signals was described by a Kalman filter. The filter was given by

\[ x_t = Ax_{t-1} + w_t, \]

where the parameters \( A, C, w \) and \( v \) were those to be estimated [30], where \( x_t \) is the two-dimensional vector consisting of the first (F1) and second (F2) formant frequencies, \( y_t \) is the one-dimensional vector representing one IC, the matrix \( A \) describes the relationship between past and future formant frequencies, \( C \) describes the expectation of the reconstructed EEGs given a set of formant frequencies and the error terms \( w_t \) and \( v_t \) are white Gaussian random variables.

In the decoding phase (Figure 5), the inputs to the Kalman filter specified in the learning phase were the ICs whose dipole solutions were located at the premotor cortex, SMA and/or Broca’s area, according to the previous neuroimaging studies related to silent speech [19,31-34], and the filter estimated spectrograms for the silent speeches using the so-called Kalman filter algorithm [35]. The 0 ms on the EEGs was defined to be average EMG onset in the learning phase for each subject (Figure 5).

HMM construction

The above SSRS for silent janken was constructed in terms of vowel recognition. Therefore, for example, spring (“haru” in English pronunciation of Japanese) and summer (“natsu” in English one) could not be discriminated by the SSRS, because the vowel transitions are the same. In order to cope with this problem, Experiment II was designated. In the learning phase for Experiment II, speech signals were transformed into vowel and consonant sequences, and these transitions were learned by hidden Markov model (HMM), in addition to spectrograms. In the decoding phase, the inputs to the HMM are spectrograms estimated from the Kalman filter specified in the learning phase. Which season was silently spoken was determined by the maximal likelihood among each HMM output in the following.

For example, the present “spring”-HMM is the left-to-right one shown in Figure 6, and is characterized by the following: (i) \( N \) is the number of states in the model; (ii) \( M \) is the number of output vowel and consonant; (iii) \( A \) (a \( ij \)), the transition matrix of the underlying Markov chain, where \( a_{ij} \) is the probability of making a transition from state \( i \) to state \( j \); (iv) \( b_j(x) \) is the probability of outputting digitalized speech signal \( x \) in state \( j \); (v) \( \pi = (\pi_i) \), \( i = 1, 2, \ldots, N \), the initial state probability vector. Figure 6 exemplifies \( N=5 \) and \( M=4 \). In the learning phase, each vowel and consonant occurrence is segmented into \( N \) states. This segmentation is achieved by finding the optimum state sequence, via the Viterbi algorithm with \( b_j(x) \) modeled by Gaussian mixture densities, in addition to the initialization of \( \pi \) and \( a_{ij} \). Parameters in the densities are estimated from spectrograms for ten actual speech trials after K-means clustering [36]. That is, for \( i=1, 2, \ldots, N \),

\[ \log \pi_i(t=0) = \max \left\{ f^*(i,t-1) + \log a_{ij} b_j(y) \right\} (t=1, 2, \ldots, T) \]
are calculated, then a maximum log-likelihood

\[ L = \max_{i \in F} f^\prime (i, T) \, l \]

is obtained [36], where \( F \) is a set of final states. Thus, the HMM parameters were initialized by the Viterbi algorithm and then re-estimated by the Baum-Welch algorithm. These procedures were carried out by HTK (a portable toolkit for building and manipulating HMMs in C: http://htk.eng.cam.ac.uk/). In the decoding phase, a silently spoken season was assumed to be maximal among each season-HMM likelihood value for the predicted spectrogram from the Kalman filter.

**Results**

**Grand averages**

Figure 7 shows the grand averages for the actual and silent janken tasks. This figure reveals similar BP-like component both for the actual and silent speeches, while motor potential (MP)-like ones [37] for only the actual one. Therefore, for both the tasks, we should pay attention to Bereitschaftspotential (BP)-like components [38].

**Correlation coefficients**

We statistically examined the hypothesis in the DIVA model for our SSRS. From the correlation coefficients, with the minimal P-values, for all the tasks by all the subjects, between each formant frequency and the IC, it followed that significant correlations were found for both F1 (\( r=0.62-0.88, p \leq 6.00 \times 10^{-15} \) for "rock"; \( r=0.64-0.94, p \leq 7.64 \times 10^{-11} \) for "paper"; \( r=0.54-0.95, p \leq 3.27 \times 10^{-11} \) for "scissors") and F2 (\( r=0.65-0.92, p \leq 7.55 \times 10^{-15} \) for "rock"; \( r=0.60-0.98, p \leq 2.83 \times 10^{-11} \) for "paper"; \( r=0.68-0.92, p \leq 7.88 \times 10^{-15} \) for "scissors") except for one subject (DK) (Table 1), thus confirming the hypothesis. The Kalman filter parameters were calculated by using one pair of the ICs and the spectrogram, whose correlation coefficient had the minimal P-value.

---

**Figure 5:** Decoding phase in the present SSRS.

**Figure 6:** Representation of the present left-to-right HMM with six states.
Estimated spectrograms for the silent janken

As the training performance, in case of one subject, diagonal parts of Figure 8 show the predicted spectrograms in the F1-F2 plane for the silent “rock”, “paper” and “scissors” with ellipsoidal distributions of five Japanese vowels [39]. In case of “rock” (/gu:/) and “paper” (/pa:)/, when the formant frequency trajectories reach the /u/ and /a/ regions, respectively, the predictions were considered to be correct, while, in case of “scissors” (/tʃɔki/), the trajectory was regarded as right if it passed through the region /o/ then the distribution /i/. In terms of Japanese pronunciation, a main difference between “scissors” and the other is that the former has two different vowels, and the latter one. To incorporate this difference in the Kalman filter algorithm, the initial values of the covariance matrix [35] were set to be variances of F1 and F2 and their covariance. Figure 9 plots all the spectrograms for each janken in the F1-F2 plane, including the covariance matrices. Note that the covariance for “scissors” was much larger than those for “rock” and “paper”. The diagonal parts of Figure 8 show the outputs from the Kalman filter algorithm with the initial values (V) depicted in Figure 9, and all indicates the correct predictions. The same tendencies as in Figure 8 were obtained for all the rest subjects. The rest of Figure 8 exemplifies the misapplication of our predictors. For example, the “rock” predictor correctly estimated only for the silent “rock” EEGs.

Generalization to silent consonant recognition

Table 2 shows a confusion matrix for the silent “season” tasks in terms of HMM log-likelihood values. For example, at the first row (Silent “spring” EEG), for the estimated silent “spring” spectrograms, the log-likelihood of the “spring”-HMM was higher than those of the other-HMMs. The other rows demonstrate the same tendency. Therefore, if higher log-likelihood values are accepted, it could be demonstrated that these HMMs work well. As a preliminary result, the accuracy was 86% (“spring”), 29% (“summer”), 43% (“autumn”) and 100% (“winter”) for one subject.

Discussion and Future Outlook

In order to decode silent speeches from single trial EEGs, we used Kalman filters for the vowel recognition, and HMMs for continuous speech one including consonants. The performance of the present Kalman filters might be improved in the following.

Three-dimensional kalman filter

By constructing three-dimensional Kalman filter, that is, involving F3, we obtained more discriminative results for the silent “rock” and “paper” tasks (Figure 10 (A) and (B), respectively).
Figure 8: A confusion matrix of the three silent janken predictors for one subject.

Table 2: A confusion matrix for the silent “season”-HMMs.

<table>
<thead>
<tr>
<th></th>
<th>Silent “Spring” EEG</th>
<th>Silent “Summer” EEG</th>
<th>Silent “Autumn” EEG</th>
<th>Silent “Winter” EEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Spring”-HMM</td>
<td>-2844</td>
<td>-5749</td>
<td>-4663</td>
<td>-6441</td>
</tr>
<tr>
<td>“Summer”-HMM</td>
<td>-2878</td>
<td>-4246</td>
<td>-4246</td>
<td>-6064</td>
</tr>
<tr>
<td>“Autumn”-HMM</td>
<td>-3377</td>
<td>-5428</td>
<td>-4204</td>
<td>-6055</td>
</tr>
<tr>
<td>“Winter”-HMM</td>
<td>-3310</td>
<td>-5940</td>
<td>-5609</td>
<td>-5510</td>
</tr>
</tbody>
</table>

Figure 9: Spectrograms in the F1-F2 plane, obtained from the speech signals during the actual janken tasks: (A) “rock”; (B) “paper”; (C) “scissors”. Each task includes 30 trials.
Kalman filter using two ICs

Intrinsically, vowels and consonants are known to be processed by distinct neural mechanisms [40]. For example, vowels and consonants increased activation in right middle temporal and frontal areas, respectively [41]. Tentatively, we constructed a Kalman filter with one IC whose dipole solution was located at the temporal area, in addition to the frontal-area-dopole IC, in the learning phase, and then the silent “haru” spectrogram was estimated in the learning phase. Figure 11 (A), (B) and (C) show spectrograms obtained by the Kalman filter with only one IC, that with the above two ICs and that with two ICs whose dipoles were localized to the other areas, respectively. Figure 11 (B) revealed the best performance.

Practical problem

In practice, it is unknown during which silent task EEGs were recorded. So, using all the estimated spectrograms that were obtained by all the KFs with such EEGs, all the HMMS outputted log-likelihoods. Table 3 shows a confusion matrix for one trial by one subject. This table indicates that the present method worked well.

Even if EEGs would be able to be recorded when a subject attempted to voice, speech signals might not be measured. In this case, we could obtain the speech signal by a person physically fitted to the subject

Future research

Since Japanese has syllable-timed rhythm [42], the present method could be easily generalized to all the pairs of hiragana. However, because the present results are limited to the training performance,
proper testing of our method must be done in future. Moreover, there are several parameters affecting the present algorithm, such as EEG intervals which ECDL is applied to, EMG onset, speech signal features, and so on. Thus, it is hoped to generalize the present framework to continuous silent speech recognition using cepstra.

Acknowledgments

This research was partly supported by a Grants-in-Aid for Scientific Research on Scientific Research (C) (23500151) and (C) (15K00276) - The Japan Society for the Promotion of Science.

References