論文 / 著書情報
Article / Book Information

Title	Time-lag Adaptation for Semi-synchronous Speech and Pen Input
Authors	Yasushi Watanabe, Koichi Shinoda, SADAOKI FURUI
Citation	Proc. INTERSPEECH2008, Vol. , No. , pp. 2675-2678,
Pub. date	2008, 9
Copyright	(c) 2008 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

interspeech 2008 incorporating SST 08

Time-lag Adaptation for Semi-synchronous Speech and Pen Input

Yasushi Watanabe, Koichi Shinoda, and Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology
yasusi@ks.cs.titech.ac.jp, {shinoda, furui}@cs.titech.ac.jp

Abstract

In a previous study, we developed an interface using semi-synchronous speech and pen input. In this interface, a user speaks while writing, and the pen input complements the speech, enabling a higher recognition performance than with speech alone. When a user inputs speech and pen, there is a time lag between the two modes, and the lag differs among users. We propose a method for adapting to the different time lags of individual users. This method was evaluated in a Japanese continuous speech recognition task with three different pen-input interfaces including a QWERTY keyboard interface. The time-lag adaptation improved recognition accuracies by up to 0.5 point.
Index Terms: user interface, speech recognition, handwritten character recognition, multimodal recognition

1. Introduction

Mobile devices such as PDAs and cellular phones are very popular, and an interface in which users can easily input long sentences is desirable for e-mailing, scheduling, and other purposes. At present, a ten-key pad, speech, or handwriting can be used for these purposes. However, users often have difficulty entering long sentences with a ten-key pad. Moreover, while a speech interface has a recognition accuracy of more than 90% in quiet conditions and an input speed that is faster than keying, its performance is seriously degraded in a noisy mobile environment. On the other hand, the recognition accuracy of handwritten characters is generally high and not affected by acoustical ambient noise. However, inputting characters is much slower than speaking. Thus, speech and pen inputs have complementary advantages and disadvantages. Combining these two modes should be a better way to deal with noise than speech alone, and it should also enable faster input speed than pen alone.

There have been several studies related to multimodal input using speech and pen. For example, Bann *et al.* [1] proposed a speech interface that uses finger tapping at word boundaries. Zhou *et al.* [2] combined speech and pen input for entering Chinese characters. Hui *et al.* [3] used speech and pen gestures for navigational inquiries. While all of these approaches were proven effective, they have limitations either in speech input information or in its application. Different from these works, we previously developed a multimodal interface that uses pen input to improve recognition performance for general *large vocabulary continuous speech recognition* (LVCSR) [4]. In this interface, a user speaks while writing, and the pen input complements the speech, enabling a higher recognition performance than with speech alone. As time lags between the two modes occur in practice, a multimodal recognition algorithm that can handle the asynchronicity of the two modes by using a segment-based unification scheme was used. This method was evaluated under noisy conditions with four different pen-input interfaces (character, stroke, pen-touch, and point-to-character), each of which is assumed to be given for a phrase unit in speech. We demonstrated that this method improved recognition accuracy in comparison with that by speech alone in all pen-input conditions.

Users of these interfaces will have different time lags, and to integrate the two modes effectively, this difference needs to be taken into consideration. In our previous study, the time lags were optimized using test data for each subject. To improve recognition performance in practice, adaptation to the time lags of individual users is necessary. In this paper, we propose a method that can adapt to the different time lags of individual users . We also propose a new interface using keyboard input for mobile devices in addition to the four interfaces of our previous study [4].

Section 2 describes three pen-input interfaces used in this study. Section 3 summarizes our previously proposed multimodal recognition algorithm [4]. Section 4 describes the time-lag adaptation method. Section 5 describes our experiments and their results, and Section 6 concludes this paper.

2. Semi-synchronous speech and pen input

In our previous study [4], we proposed four interfaces (character, stroke, pen-touch, and point-to-character), each of which is assumed to be given for a phrase unit in speech. Since character input and point-to-character had higher recognition performances than the other two, we decided to use them in this study. We also propose a new interface using keyboard input for mobile devices in addition to the interfaces of our previous study. Here, we introduce these three interfaces.

Our multimodal interfaces use semi-synchronous speech and pen input for entering Japanese sentences consisting of many phrases. Inputting the same information by speech and pen is practically impossible since the speed of pen input is very slow compared with speech. Therefore, we need to place constraints on the pen input that corresponds to each speech phrase. First, we assume that a pen input will be given at the start of a speech phrase (Figure 1). Second, we assume that a user will try to synchronize the start of the pen input with the start of the corresponding speech phrase. Here, a phrase means a Japanese *bunsetsu*, which contains one or more words. A user would not need to give pen inputs for all phrases or insert pauses between phrases. We use the following three interfaces on the basis of these assumptions.

1. **Character input** A user writes the initial character of each phrase in *hiragana*, which are phonetic Japanese characters. The interface has several input boxes, and a user writes one character in one box from left to right (Figure 2).
2. **Point-to-character** Each *hiragana* is either a vowel (V) or a consonant followed by a vowel (CV). They are classified

Accepted after peer review of full paper

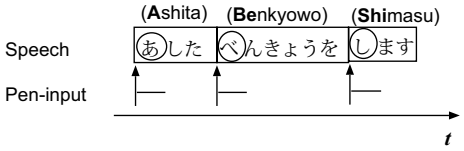


Figure 1: Relationship between speech and pen input. Arrows indicate the start times of pen input for the circled characters. Each box of speech indicates one phrase.

Figure 2: Example of character inputs for the sentence in Figure 1. Here, the first character for each phrase is input.

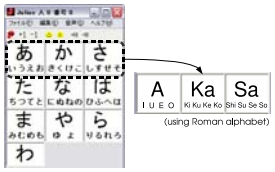


Figure 3: Point-to-character interface. Each region corresponds to 1–5 characters that share the same consonant.

into ten groups: one corresponds to characters only for vowels, and the rest correspond to the characters sharing the same consonant. There are five vowels and nine consonants. the interface provides a character table with this classification (Figure 3), of which each element indicates a character group. A user taps the group to which the first character of the speech phrase belongs.

3. Keyboard input A user inputs the initial character of each phrase expressed in Roman letters with a QWERTY keyboard layout (Figure 4). For instance, if a user want to input "Fuji", he/she keys in "F" while uttering "Fuji" (Figure 5). A mobile PC, Sony VAIO type U uses this as an input device. The PC has a 4.5 inch touch panel and weighs 500 g. It is so small that the user cannot touch type. A software keyboard is used for the input because the actual keyboard of the PC is not easy to use. As "Q" and "X" cannot be initial characters of phrases in Japanese, so their inputs are ignored. A user can use first fingers or a stylus pen for input. In this study, we treat inputs with fingers as "pen input".

3. Multimodal recognition algorithm

Here, we describe the multimodal recognition algorithm proposed in the previous study [4]. The multimodal recognition algorithm we developed for the interface uses a two-pass search process. This algorithm is based on the conventional LVCSR algorithm, for which the linguistic unit is a word.

A user will try to synchronize the starting time of his/her pen input with that of the corresponding speech phrase, but asynchronicity between the two modes always occurs in practice; some users start the pen input before the corresponding speech phrase, others start it after starting too speak. To merge pen-input recognition hypotheses with their corresponding speech-recognition hypotheses, this asynchronicity needs to be taken into consideration. We assume that each user has his or her own tendency for time lags between the two modes.

Let the pen-input starting time be t and the corresponding phrase starting time be t'. The time lag Δt is defined as $t - t'$. The mean μ of Δt is estimated using the user's data. In

Figure 4: Keyboard input interface

Figure 5: Example of keyboard input

the recognition phase, the pen-input starting time is modified to $t - \mu$ (Figure 6). The mean differs among users and can be negative.

In the first pass, each speech recognition hypothesis is merged with the pen-input recognition hypothesis weighted by factor α at the time the pen-input begins. To do this, the speech recognition process is suspended from the time the pen-input begins to the time the pen-input recognition finishes. The algorithm is shown below.

Step 0: Set time $t = 0$.

Step 1: If the start of a pen-input is detected, go to Step 4.

Step 2: If speech input ends, terminate.

Step 3: Set $t = t + 1$, and go to Step 1.

Step 4: Perform the following steps for each hypothesis h of speech recognition at time t.

1. Extract the initial character C of the hypothesis.
2. Obtain the pen-input recognition likelihood of C, $L(C)$.
3. Calculate the likelihood for h using
$$L(h) = L_s(h) + \alpha L(C),$$
where $L_s(h)$ is the speech recognition likelihood of h, and α is a weighting factor.

Step 5: Set $t = t + 1$, and go to Step 1.

The hypotheses having a word whose initial character has a high probability in the pen-input recognition are likely to remain within the beamwidth, so the candidates are effectively narrowed down.

While we consider the mean of the time lag between the two modes in the first pass, its variance is also important for our multimodal recognition. Its variance is considered in the second pass. (See our previous report [4] for details.)

While this method was proved to be effective in our evaluation under noisy conditions [4], the mean of the time lags for each speaker was optimized using the test data for him/her. Time-lag adaptation is needed to improve recognition performance in practice.

4. Time-lag adaptation

The time-lag adaptation method deals with individual differences in multimodal recognition. The word corresponding to the pen input is determined automatically, and the value of μ is adapted. The supervised adaptation method is described below.

First, let the pen inputs be $c_n (n = 1, \cdots, C)$ and word inputs be $w_m (m = 1, \cdots, M)$. Let the start time of c_n be t_n

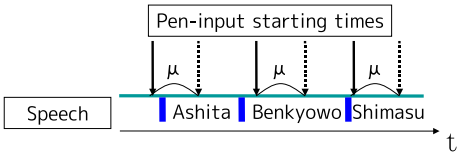


Figure 6: Example of modifying pen-input starting times. Thick lines represent starting times of speech phrases. Dashed arrows represent pen-input starting times in practice and full arrows represent modified pen-input starting times.

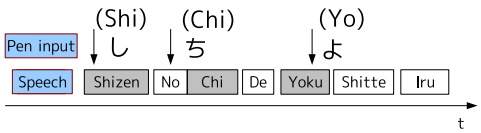


Figure 7: Example of association of speech and pen inputs. Each pen input is associated with a word indicated by a gray box.

and the start time of w_m be t_m. t_m is determined using forced alignment with the word transcripts. The word associated with each pen input c_n is determined as follows. It should be noted that auxiliary verbs and postpositional particles that cannot be the initial word of phrases in Japanese are ignored.

Step 1 Obtain $W = \{w_m | \ |t - t_m| < I\}$ from the input data.

Step 2 Obtain the subset W' of W. The initial word of each word of W' is one of characters corresponding to the pen input c_n.

Step 3 Associate w_m for which $|t_n - t_m|$ is the smallest of W' with the pen input c_m. If no word can be associated in this way, go to step 4.

Step 4 Associate w_m at which $|t_n - t_m|$ is the smallest with the pen input c_m.

Here, I is a threshold regarding the time lag. Step 1 considers words within I before and after the pen input. In character-input, characters corresponding to a pen-input are defined as characters in the top N in handwritten recognition. N is determined so that the accumulated recognition rate is over 97%. Figure 7 shows an example of an association between speech and pen input. Finally, the mean of the time lags Δt for the adaptation data of each user is calculated using the association between speech and pen input, and its value is used as μ.

In unsupervised adaptation, utterances are first recognized, and the start time of each word is determined using the 1-best recognition result with timing information about the word boundaries. Then, pen inputs are associated with words by using the above method, and the mean of the time lags Δt for each user's data is used as μ.

5. Experiments

5.1. Database

We collected simultaneous input data of speech and pen input from 20 Japanese subjects (19 males and 1 female) in a recording room. Each subject input 50 sentences for each of the three interfaces described in Section 2. The sentences were randomly chosen from the ASJ-PB database [5], which has a phonetically balanced sentence set and the ASJ-JNAS database [6], which has a sentence set from Japanese newspaper articles. The subjects used each interface for about ten minutes beforehand to get accustomed to it. They mainly used their fingers to input with the keyboard interface. When the subject failed to input a sentence correctly, he/she would re-input it.

5.2. Experimental conditions

The 50 sentences each subject input for each interface were divided into two groups: the first 15 sentences were used as adaptation data and the rest were used as test data. The adaptation data were used for acoustic model adaptation and for time-lag adaptation.

We used triphone HMMs with 16 mixture components per state included in the IPA Japanese Dictation Toolkit (2000 version) [7] as the acoustic model, and adapted the HMMs using the data from each subject. The lexicon was created using *Mainichi* newspaper articles from 1995 to 2001. We removed symbols without pronunciations and extracted 60,000 high-frequency words.

For the handwritten character recognition, we used continuous HMMs trained on 43,800 characters, which were in *hiragana*, written by 10 writers, and obtained from the online handwritten character database [8]. Our handwritten character recognition method was based on stroke-based left-to-right HMMs [9], for which a recognition unit was a stroke and the number of stroke units was 25. Pen-down strokes had three states, and pen-up strokes had one state without self-loop probability. The number of mixture components for each state was one. We created a handwritten character dictionary that consisted of 71 *hiragana*, each of which was represented as a concatenation of strokes. Characters for which the writing order significantly differed among writers had multiple entries in the dictionary. As a result, the number of handwritten characters in the dictionary was actually 82.

We implemented our algorithm with the Julius speech decoder [10] of our previous study [4]. We obtained the pen-input recognition likelihood in advance by using Julius and unified the two modes offline by using our algorithm.

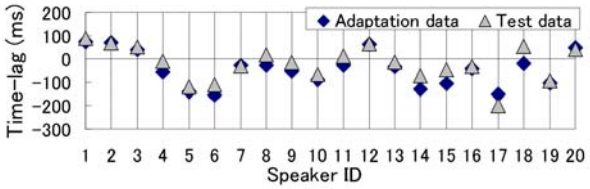
The original HMMs for speech were adapted using supervised maximum likelihood linear regression (MLLR) [11]. Regarding time-lag adaptation, we experimented with the case in which the mean μ_s was estimated by supervised adaptation using adaptation data or by unsupervised adaptation using test data. Regarding I and N described in Section 4, I was 300 ms, and N was 10 because, in the database of our previous study, most time lags were less than 300 ms and the 10-best recognition rates for the handwritten character recognition were 97%.

We also tested the case in which the mean μ was not adapted, where the μ_0 for each subject was estimated using the adaptation data of all the subjects other than him/her.

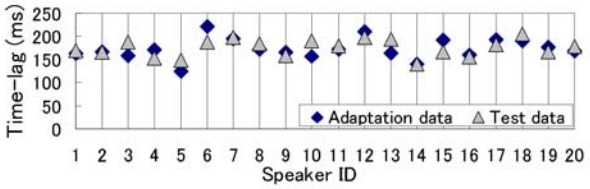
5.3. Results

Figure 8 shows the time lags between speech and pen input for each subject. This result shows that each subject had his or her own tendency to some extent regarding time lags.

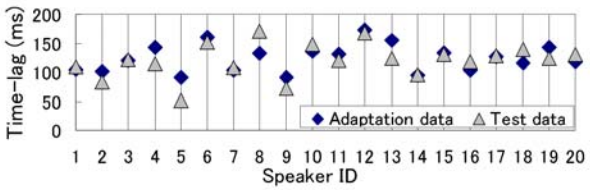
Figure 9 shows the recognition accuracy of speech recognition and the proposed multimodal recognition with speech and pen input. Compared with the case in which μ_0 was used, the time-lag adaptation improved word accuracies by up to 0.5 point. Regarding time-lag adaptation for point-to-character and key inputs, unsupervised adaptation had higher recognition accuracy than supervised adaptation. The reason is as follows. The adaptation data were 15 sentences that were input early, and the test data were 35 sentences that were input late. As the subject's tendency regarding time lag changed gradually while inputting, unsupervised adaptation using the test data gave a better



(a) Handwritten character



(b) Point-to-character



(c) Keyboard

Figure 8: Mean time lag. The mean values of the adaptation data and test data are shown.

estimate than supervised adaptation.

By applying acoustic model adaptation, we improved the word accuracies of speech recognition by 2.5–2.6 points. By combining acoustic model adaptation and multimodal recognition, word accuracies improved by up to 4.5 points. Among the three interfaces, keyboard input had the highest accuracy and point-to-character input had the second.

6. Conclusions

We developed an adaptation method for semi-synchronous speech and pen input in multimodal recognition that deals with the characteristically individual time lags of users. This method was evaluated with three different pen-input interfaces, including a new interface using a QWERTY keyboard. Time-lag adaptation improved recognition accuracies by up to 0.5 point. Our future work will include adapting the weighting factor α in multimodal recognition and evaluating a multimodal interface on a mobile phone.

7. Acknowledgements

This research was partially supported by JSPS Grants-in-Aid for Scientific Research (B) 15300054. Our thanks go to the Nakagawa Laboratory of the Tokyo University of Agriculture and Technology for providing the online handwritten character database.

8. References

[1] H. Bann, C. Miyajima, K. Itou, K. Takeda, and F. Itakura, “Speech recognition using synchronization between speech and figure tapping,” *Proc. INTERSPEECH 2004 – ICSLP*, Jeju, Korea, 2004.

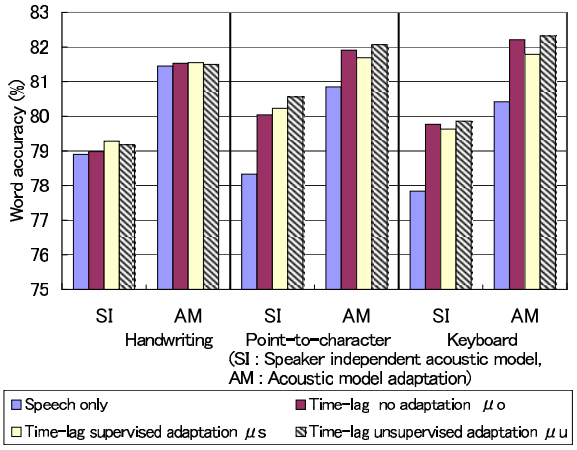


Figure 9: Comparison of recognition accuracies. The figure shows the case in which a speaker-independent acoustic model was used and the case in which adapted acoustic models were used.

[2] X. Zhou, Y. Tian, J. Zhou, F. K. Soong, and B. Dai, “Improved Chinese character input by merging speech and handwriting recognition hypotheses,” *Proc. ICASSP 2006*, Toulouse, France, vol. 1, pp. 609–612, 2006.

[3] P. Y. Hui and H. M. Meng, “Joint Interpretation of Input Speech and Pen Gestures for Multimodal Human-Computer Interaction,” *Proc. INTERSPEECH 2006 – ICSLP*, Pittsburgh, Pennsylvania, pp. 1197–1200, 2006.

[4] Y. Watanabe, K. Iwata, R. Nakagawa, K. Shinoda, and S. Furui, “Semi-Synchronous Speech and Pen Input”, *Proc. ICASSP 2007*, Hawaii, U.S.A., SPE-L5.1 (IV-409), 2007.

[5] S. Hayamizu, S. Itahashi, T. Kobayashi, and T. Takezawa, “Design and creation of speech and text corpora of dialogue,” *IEICE Trans. Information and Systems*, vol. E76-D, no. 1, pp. 17–22, 1993.

[6] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, “The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus,” *Proc* ICSLP, pp. 3261–3264, 1998.

[7] http://winnie.kuis.kyoto-u.ac.jp/dictation/

[8] M. Nakagawa, T. Higashiyama, Y. Yamanaka, S. Sawada, L.V. Tu, and K. Akiyama, “Collection and Utilization of On-line Handwritten Character Patterns Sampled in a Sequence of Sentences without Any Writing Instructions,” Technical Report of IEICE, 95, 278, pp. 43–48, 1995 (in Japanese).

[9] M. Nakai, N. Akira, H. Shimodaira, and S. Sagayama, “Substroke Approach to HMM-based On-line Kanji Handwriting Recognition,” *Proc. ICDAR 2001*, pp. 491–495, 2001.

[10] A. Lee, T. Kawahara, and K. Shikano, “Julius – an open source real-time large vocabulary recognition engine,” *Proc. EUROSPEECH 2001*, pp. 1691–1694, 2001.

[11] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, Vol. 9, No. 2, pp. 171–185, 1995.