ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Speech Interface for Name Input based on Combination of Recognition Methods Using Syllable-based N-gram and Word Dictionary

Hironori OSHIKAWA, Norihide KITAOKA, Seiichi NAKAGAWA

Toyohashi University of Technology
Department of Information and Computer Sciences
1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi, 441-8580, JAPAN
{oshikawa,kitaoka,nakagawa}@slp.ics.tut.ac.jp

10.21437/Interspeech.2004-112

Abstract

We propose an interface for a name input based on speech recognition using syllable-based N-gram and a word dictionary. Name utterance is hard to recognize accurately because of the large vocabulary size, so the system uses continuous syllable recognition with syllable-based N-gram and isolated word recognition with a dictionary containing frequent words. User first utters a name and then chooses the correct word/syllables by pen touch from word/syllable candidates which were obtained from speech recognition. System displays word candidates, syllable sequence candidates and a syllable lattice on a touch panel and user can select a desired word from the candidates. We evaluated this interface. User could find the correct answer from word candidates or syllable sequence candidates at a rate of 82-86%, and could input correct name at a rate of 94-96% using syllable selection from the syllable lattice. Some subjects used this interface and felt that it was efficient and useful.

1. Introduction

Recently, we can use many information retrieval services on the WWW with easy access interfaces based on graphical user interfaces(GUI). However, for some kind of the information retrieval tasks and under certain operating conditions, such interface may not be so efficient. For example, users with a small laptop PC, a Personal Digital Assistant (PDA)[5, 6] or a wearable computer cannot access such services efficiently or satisfactory with such interfaces. Speech input is one of promising alternatives to the keyboard and the mouse as an input modality. However, since the existing forms on the WWW are not originally intended to be filled in by speech, further investigation will be needed to apply a speech to the WWW services as an alternative of the keyboard or mouse operation. So far, some researchers have proposed speech interface systems for the WWW services [1, 2, 3, 4, 5, 6].

We can find a lot of forms which expect inputs of arbitrary character sequences. So it is important to give speech-based input modality to such forms. We discuss the interface for person's name input as an example of arbitrary character sequence input because there are many opportunities to input a name. In the speech recognition of isolated names, it is not realistic to have all names in a dictionary because of the large variety. For example, there are over 40000 Japanese first names and new names are also continuously generated. Figure 1 shows the vocabulary size and the coverage of all the names.

We developed a multimodal interface based on continuous syllable recognition: the N-best syllable sequence candidates and syllable candidates (syllable lattice). These candidates were displayed on a touch panel, and then user could select a sequence or syllables by pen touch. This interface, however, only used the syllable recognition, so the recognition rate was not good enough. In addition, Chinese character(Kanji) input was often required for Japanese text input form. Many Kanji sequences share the same pronunciations, that is to say, there are many homonyms. We propose a speech recognition technique which combines a word recognition of the frequent words with a continuous syllable recognition using a syllable-based N-gram to overcome these problems.

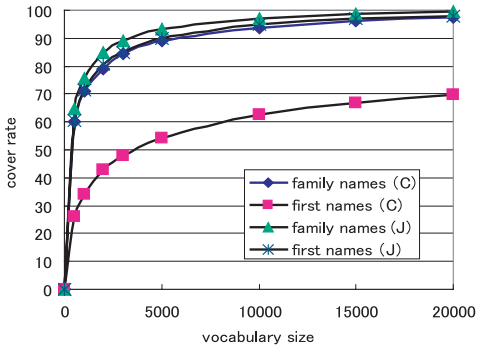

Figure 1: *Relations between the vocabulary size and the coverage of Japanese first and family names. J:Japanese (Hiragana) character, C: Chinese (Kanji) character*

2. Speech interface for name input

So far, speech interface for WWW browser operation and choosing a candidate from a menu has been investigated. However, speech input methods for an arbitrary character sequence has not been realized yet.

So, we discuss speech interface on the name input task for the WWW browser. As name utterances were hard to recognize accurately and the recognition took high computational costs because of the large vocabulary size. For European languages, an alphabet-unit sequence has been used as the name input manner [7]. We adopted continuous syllable recognition with syllable-based N-gram. All the Japanese words can be expressed as a Japanese character[Hiragana] sequence and each character is uniquely associated with a certain pronunciation of a syllable. However, syllable recognition results were not so accurate and we could not use the results without any modification. So, our system displayed N-best syllable sequence

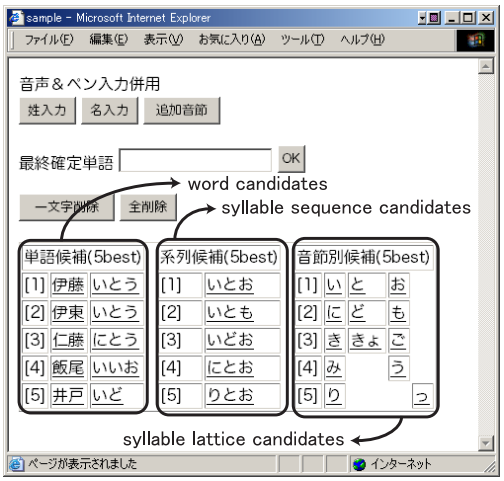

Figure 2: *Interface for Name Input*

candidates and a syllable lattice on the touch panel, and user could choose a sequence or syllables from the candidates by pen touch to construct the desired name [8]. When inputting a name, however, not only a Japanese syllable character sequence but a Chinese character sequence may be needed. Japanese syllable characters are phonograms, whereas Chinese characters are ideograms. Chinese character sequence set has many homonyms. Only using the syllable recognition, users cannot select one of such homonyms.

To overcome these problems, the system also adopted isolated word recognition where each word was associated to a Chinese character sequence in combination with continuous syllable recognition. That is to say, only frequent names were registered to the word dictionary.

Based on this hybrid recognition method, we designed an interface for name input (Figure 2). Using this interface, the user first utters a name and then chooses a correct word/syllables by pen touch from word/syllable candidates displayed on the touch panel. The system displays 5-best word candidates, 5-best syllable sequence candidates and syllable lattice with 5-best candidates for each syllable segment obtained by the recognition.

Using a word dictionary, the constraints for syllable concatenation get stronger and thus the recognition accuracy gets better(see Section 3.). As each word candidate is associated to a certain Chinese character sequence, user can choose not only a word written in Japanese (Hiragana) characters but also a word written in Chinese (Kanji) characters. The 5-best syllable sequence candidates are top five hypotheses generated by continuous syllable recognition. There are some exception of correspondence between Japanese character sequence and pronunciations. For example, when Japanese character sequence describes "o u", the sequence is pronounced a long vowel "oh". Syllable sequence candidates were modified to absorb such differences. Word candidates and syllable sequence candidates can be selected only with one click. The syllable lattice is created as follows: the best hypothesis generated by continuous syllable recognition is picked and the syllables contained in the hypothesis are taken. The other syllables (up to 5) ended at the same time frame as each syllable contained in the best hypothesis in HMM trellis are also picked.

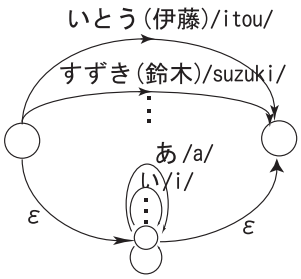

Figure 3: *Language model*

3. Recognition of the name utterances

3.1. Continuous syllable recognition

There are too many names for Japanese persons, and the number of them gets larger. Thus, it is not realistic to register all the names to the dictionary from the viewpoints of computational cost and the coverage. So we adopted a continuous syllable recognition to the name recognition task. In this recognition, the statistical language knowledge of Japanese names can be introduced to the continuous syllable recognition [9]. Because the statistics of the Japanese names has a certain characteristics in syllable arrangements, we can expect the effect to apply a statistical language model (N-gram) of the syllable unit: where S_n, S, and ϵ are a syllable, a syllable sequence, and a null transition, respectively, and λ is a penalty to reduce insertion errors.

$$\begin{aligned} P(S) &= P(S_1, S_2, \ldots, S_N) \\ &= P(S_1|\epsilon) \cdot P(S_2|S_1) \cdots P(S_N|S_{N-1}) \\ &\Rightarrow P(S_1|\epsilon) \cdot \lambda_1 \cdot P(S_2|S_1) \cdot \lambda_1 \cdots P(S_N|S_{N-1}) \cdot \lambda_1 \end{aligned}$$

3.2. Combining isolated word recognition with continuous syllable recognition

We discussed isolated word speech recognition with a word dictionary in combination with the continuous syllable recognition. In this combination, we treated words for isolated word recognition and syllables for continuous syllable recognition equally. Therefore, isolated word recognition and continuous syllable recognition can work simultaneously on one recognizer. Figure 3 shows the concept of the language model. M most frequent names are contained with first and family name dictionaries. Connection probability to a word W from the start condition, $P(w|\epsilon)$, is obtained from the word frequency. Continuous syllable recognition has disadvantage, because penalties are applied between syllables. So, we also applied another penalty to each concatenation of syllables in isolated word recognition. This method could absorb the gap of likelihoods by the difference in recognition units between isolated word recognition and continuous syllable recognition. This combination method may degrade individual recognition performance because beam search may prune more hypotheses which are generated by both isolated word recognition and continuous syllable recognition. However, the results of both methods can be obtained simultaneously by only one recognition process.

4. Experiments

We conducted name recognition/input experiments by combination of continuous speech recognition using syllable N-gram and isolated word recognition to examine the performance of the speech interface for name input.

Table 1: *Coverage for the test data (coverage for all the Japanese names is also shown in parentheses)*

vocabulary size	Family name		First name	
	C (Kanji)	J (Hiragana)	C (Kanji)	J (Hiragana)
1000	70.0(70.9)	70.0(75.7)	45.0(34.1)	70.0(71.4)
5000	85.0(89.0)	88.0(93.3)	67.0(54.2)	89.0(89.8)
10000	89.0(93.9)	92.0(97.1)	75.0(62.4)	96.0(94.8)
15000	91.0(96.1)	92.0(98.6)	76.0(66.8)	98.0(97.0)

4.1. Recognition performance evaluation

4.1.1. Experimental conditions

100 first names and 100 family names were uttered by 5 male speakers and totally 500 first name utterances and 500 family name utterances were used as test sets. These utterances were analyzed as follows: sampling frequency of 16 kHz; Hamming window size of 25 ms; frame shift of 10 ms, and feature parameters of 16 dimensional MFCC, $\Delta/\Delta\Delta$ MFCC, and $\Delta/\Delta\Delta$ energy.

Left-to-right context-independent segmental unit input HMMs were used as syllable acoustic models. Each model had 5 states, 4 of them had 4 Gaussian distributions with full-covariance matrices. The number of syllables was 114. The initial models were trained with syllable-segmented data from A-J sets (50 sentences each) of the ATR speech database consisting of utterances spoken by 6 male speakers. For syllable categories with a small amount of data in the database, 216 word data sets were additionally used. Afterward, the HMMs were retrained with MAP estimation using an Acoustic Society of Japan (ASJ) database consisting of utterances by 30 male speakers (4,518 sentences) and a Japan Newspaper Article Sentences (JNAS) database with utterances by 124 male speakers (12,703 sentences).

The language models were made from the different data for the first and family name recognition. The bigram language model for family name recognition was made from 178,000 Japanese family names, which contains 49,000 unique names. The first name bigram model was made from 500,000 Japanese first names with 110,000 unique names.

Table 1 shows the coverages of test data by N most frequent words for first and family names. In this table, C means that Chinese characters were considered and J means only Japanese characters were considered. Because of the homonyms, Cs were lower than Js.

4.1.2. Evaluation by the recognition rate

Table 2 shows syllable sequence recognition rates, syllable correct rates, syllable accuracy of the 1-best continuous syllable recognition results. Word recognition rates denote the rates considering Chinese characters (WD(C)), and only considering Japanese characters (WD(J)), respectively. As mentioned above, continuous syllable recognition and isolated word recognition were carried out at the same time, then recognition rates slightly decrease from the recognition results obtained by the recognition carried out independently. In addition, Japanese first name with a certain Japanese character sequence tends to be associated with many names written by various Chinese character sequences, resulting that the recognition rates WD(C) in Table 2 were much lower than WD(J).

The words registered in the dictionary (that is, in-vocabulary words) were recognized with 80-85% recognition rates.

Table 2: *Recognition rates of 1-best candidates[%]*

Family name	syllable recognition			word recognition	
	SEQ	COR	ACC	WD(C)	WD(J)
0	50.4	82.8	79.9	—	—
1000	48.8	82.0	77.9	57.6	58.6
5000	48.6	81.8	77.6	65.8	68.6
10000	48.8	81.9	77.7	66.2	70.2
15000	49.0	81.7	77.6	65.8	69.8
First name	syllable			word	
	SEQ	COR	ACC	WD(C)	WD(J)
0	60.6	84.0	80.7	—	—
1000	59.2	84.1	80.7	22.0	58.4
5000	59.2	83.8	80.3	21.8	70.8
10000	59 2	83.7	80.1	24.0	74.8
15000	59.0	83.6	80.0	24.0	75.2

4.1.3. Evaluation of name input ability through the interface

We examined ability to input first/family names through the interface described in Figure 2 based on the recognition results in Table 2 with 10000 word lexicons. Table 3 shows the results. In the table,"WD(C)" denotes that users could choose correct ones from 5-best Chinese character candidates, "WD(J)" from 5-best Japanese character candidates, and "SEQ" from 5-best syllable sequence candidates. "CMB(J)" denotes that users could choose from "WD(J)" or "SEQ". "SYL" means the rates that the users could construct the correct Japanese character sequences by selecting characters from the syllable lattice. The rate that the users could input correct names through at least one of these methods are shown as "SYS" (that is, the total system performance).

When using a 10000 word dictionary for family name recognition, we obtained about 76%, 78% and 87% for WD(C), WD(J) and SEQ, respectively, and these results led the CMB(J) input rate of 82%, which means the rate user could input each name with only one time of pen touch. We obtained about 94% input rate using speech interface for family names. This result indicates that the combinational use of continuous syllable recognition and isolated word recognition for the family name input is effective for family name input task. In the case of first names, we obtained only about 45% WD(C), but WD(J) and CMB(J) were about 82% and 87% respectively, and SYS was about 96%. These results proved that this combination interface was also effective for the first name input task.

4.2. On-line experiments

4.2.1. Name input experiment

We implemented the proposed interface on a laptop PC with touch pen and 5 male subjects (different from the speakers of Section 4.1) used the interface. Each subject tried to input the same 100 family and 100 first names as Section 4.1. Two word dictionaries containing 10000 frequent words were used.

The users first uttered a family name or a first name, and then constructed the desired name from the candidates of speech recognition by pen-touch. If the user could not input the desired name on the first attempt, the user could retry it. We used a real-time speech recognition system and the response time was less than 1 second. Tables 4 and 5 show the results of the ability of the interface on this on-line test. Some results were a little inferior to the off-line test described in Table 3 because of voice activity detection errors, but almost all the results were as well as those of off-line test.

Table 3: *The rate that users could input the desired name using the interface[%]*

Family name	WD (C)	WD (J)	SEQ	CMB (J)	SYL	SYS
0	—	—	77.2	—	90.2	92.8
1000	62.8	63.0	71.2	81.8	90.4	94.0
5000	74.2	75.6	70.6	83.0	90.4	94.2
10000	76.0	78.0	70.8	83.4	89.6	94.0
15000	76.8	77.8	71.0	83.4	89.4	94.2

First name	WD (C)	WD (J)	SEQ	CMB (J)	SYL	SYS
0	—	—	80.0	—	94.2	95.6
1000	33.6	61.8	78.6	86.2	93.6	95.8
5000	43.0	76.8	76.8	86.8	94.0	96.0
10000	44.8	81.8	77.4	86.8	93.4	96.2
15000	44.6	82.6	76.6	87.9	93.2	96.2

Frequencies of the use of items displayed on the touch panel are shown in Table 6. Upper columns have the total frequency of the 1st and 2nd attempts and the cumulative totals of the items are also shown in parentheses the 1st and 2nd attempts are shown separately in lower columns. Comparing Tables 4, 5 and 6, we could find that users prefer retrying speaking to select syllables from syllable lattice. Selection from the syllable lattice was the family resort for names which was too hard to recognize.

4.2.2. Subjective evaluation

We carried out a questionnaire survey about the total usability, adequateness of displayed items, and response speed of the interface. Subjects replied to each question with a value from 1(bad) to 5(good).

The results are shown in Table 7. We obtained the usability scores of 4.0 and 3.8 for family name and first name inputs, respectively, indicating that the interface gave a good impression on users. In addition, we obtained good scores for adequateness of displayed items and response speed.

We also asked the subject which items they often used. Almost all of them said they used word candidates and sequence candidates frequently, but they hardly used syllable lattice especially in the first attempt for every trial. This result was consistent with the finding in Section 4.2.1.

5. Conclusion

We proposed an interface for a name input based on continuous syllable recognition with syllable-based N-gram and isolated word recognition with a dictionary containing frequent words. The system displayed word candidates, syllable sequence candidates and a syllable lattice on a touch panel and user can select a desired word from the candidates. User could find the correct answer from word candidates or syllable sequence candidates at a rate of 82-86%, and could input correct name at a rate of 94-96% using syllable selection from the syllable lattice. Five subjects used this interface and felt that it was efficient and useful.

Table 4: *Results of on-line test (Family name)[%]*

Subjects	WD (C)	WD (J)	SEQ	CMB (J)	SYL	SYS
A	75	77	49	79	68	84
B	66	69	43	73	74	84
C	75	78	54	82	71	84
D	79	82	63	88	72	90
E	77	81	62	86	79	92
Ave.	74.4	77.4	54.2	81.4	72.8	86.8

Table 5: *Results of on-line test (First name)[%]*

Subjects	WD (C)	WD (J)	SEQ	CMB (J)	SYL	SYS
A	41	71	56	77	70	85
B	40	77	77	88	78	92
C	45	85	76	90	82	94
D	50	94	75	96	80	97
E	44	87	74	91	74	94
Ave.	44.0	82.8	71.6	88.4	76.8	92.4

Table 6: *Frequencies of the use of displayed items[%]*

	WD(C)	WD(J)	SEQ	SYL	SYS
Family name	78.8 74.6/4.2	3.2(82.0) 2.6/0.6	4.4(86.4) 3.8/0.6	6.4 2.6/3.8	92.8 83.6/9.2
First name	46.4 43.4/3.0	39.8(86.2) 37.8/2.0	6.2(92.4) 4.6/1.6	2.8 1.4/1.4	95.2 87.2/8.0

Table 7: *Subjective evaluation result*

	usability		display result		response speed	
	family	first	family	first	family	first
Ave.	4.0	3.8	4.2	3.6	4.2	4.0

6. References

[1] A.Kai and T.Nakano and S.Nakagawa, "A speech interface system for information retrieval tasks on the WWW", Proc. of SPECOM'99, pp.141-144, 1999.

[2] R.C.Rose, I.Arizmendi, S.Parthasarathy, "An efficient framework for robust mobile speech recognition services", Proc. of ICASSP2003, pp.316-319, 2003.

[3] D.Jouvet et al., "About improving recognition of spontaneously uttered French city-names", Proc. of ICASSP2003, pp.544-547, 2003.

[4] M.Johnston et al., "Multimodal Language processing for mobile information access", Proc. of ICSLP2002, pp.2237-2240, 2002.

[5] Y.Y.Wang, "Robust language understanding in MiPad", Proc. of EUROSPEECH2001, pp.1555-1558, 2001.

[6] L.Comerford et al., "The IBM personal speech assistant", Proc. of ICASSP2001, pp.1-4, 2001.

[7] D.Jouver, J.Monne, "Recognition of spelled names over the telephone and rejection of data out of the spelling lexicon", Proc. of EUROSPEECH99, pp.286-89, 1999.

[8] H.Oshikawa, N.Kitaoka, S.Nakagawa, "Speech interface fo rarbitrary character string input on WWW browser", Proc. of Spring Meeting of ASJ(in Japanese), pp.217-218, 2002.3.

[9] K.Tanigaki, H.Yamamoto, Y.Sagisaka, "A hierarchical language model incorporating class-dependent word models for OOV words recognition", Proc. of ICSLP2000, Vol.3, pp.123-126, 2000