Detection and Recognition of Correction Utterance in Spontaneously Spoken Dialog

Norihide Kitaoka, Naoko Kakutani and Seiichi Nakagawa

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

Abstract

Recently, the performance of speech recognition was drastically improved, and the products with the interface based on speech recognition have been realized. However, when we communicate with computers through a speech interface, misrecognition is inevitable, and it is difficult to recover from it because of the immaturity of the interface. Users try to recover from misrecognition by a repetition of the same content. So, the detection of user's repetition is helpful for a system to detect its misunderstanding, and to recover from the misrecognition.

In this paper, we assume the utterance which includes repetitions a correction and propose a method to detect correction utterances in spontaneously spoken dialog using a word spotting based on DTW (dynamic time warping) and N-best hypotheses overlapping measure. As a result, we achieved recall rate of 92.7% and precision of 89.1%. Moreover, we tried to improve recognition accuracy using the detection. Using the choice of vocabulary and grammar setup based on the detection, we achieved improvement in recognition performance from 42.7% to 50.0% for correction utterance and from 70.5% to 77.9% for non-correction utterance.

1. Introduction

Recently, the performance of speech recognition was drastically improved, and the products with the interface based on speech recognition have been realized. However, when we communicate with computers through a speech interface, misrecognition is inevitable, and it is difficult to recover from it because of the immaturity of the interface. Users often repeat the sentence or a part of the sentence to recover it but the same misrecognition often occurs repeatedly. This will be much load to the users. Therefore, the method with which the system detects its misunderstanding is necessary to recover from it. As written above, users try to recover from misrecognition by a repetition of the same content. So, the detection of user's repetition is helpful for a system to detect its misunderstanding, and to recover from the misrecognition.

Some attempts have been made to detect correction utterances. Many prosodic differences of user's utterances were found after a system's misrecognition [1]. Levow [3, 4], tried to detect correction utterance using prosodic effects on user's utterance caused by system's misrecognition and achieved error rate of 25%. When misrecognition occurs, users prefer to respeak even when they have other input methods like typing, handwriting, etc [2], so it is important to obtain higher recognition accuracy for correction utterances. Measures have been proposed [5, 6] to detect repeated speech of an isolated word for correction of misrecognition, using overlapping candidates,
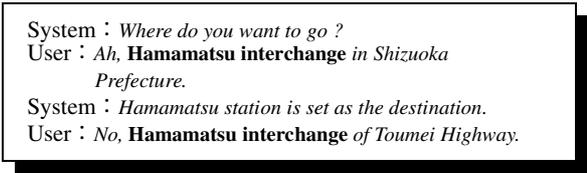
System：*Where do you want to go ?*
User：*Ah,* **Hamamatsu interchange** *in Shizuoka Prefecture.*
System：*Hamamatsu station is set as the destination.*
User：*No,* **Hamamatsu interchange** *of Toumei Highway.*

Figure 1: An example of repetition in a dialog

likelihood difference and distance between the time sequence vectors. Good detection performance was achieved by combination of overlapping candidates and distance measures.

In [7], we proposed a method to detect the correction utterance based on DTW (Dynamic Time Warping) and N-best hypotheses overlapping measure on a location name input task for a car navigation system, and achieved recall rate of 92.5% and precision of 86.0%. Moreover, we tried to improve recognition accuracy using the detection. Using the choice of vocabulary setup based on the detection, we achieved significant improvement in recognition performance for both correction and non-correction speech.

In this paper, we extended this method to apply more spontaneously spoken dialog as described in Fig. 1, and tried to improve the recognition accuracy using the detection.

As this example, the second user utterance does not have to be a part of the first utterance. So we have to find part-to-part identification between the utterances.

2. Detection Methods

2.1. Detection based on Dynamic Time Warping

We proposed the method to detect partial repetitions based on DTW in an isolated location name input task [7]. We applied a word spotting technique to detect it on the assumption that users repeat only a part of the first utterance. On the other hand, we can find that repetition parts as "Hamamatsu interchange" in the example of Fig. 1 can appear at any position in correction utterance. That is, different from the task in [7], the second utterance does not have to be a part of the first utterance. Thus, the method should be modified to be applied to the new task.

Fig. 3 shows an example of an optimal path of DTW using DTW path as Fig. 2 between an original utterance $(a_1,a_2,a_3,...,a_I)$ and a correction one $(b_1,b_2,b_3,...,b_J)$. We used 10 dimensions of LPC mel cepstral coefficients as feature parameters a_i and b_j. The segments of the path corresponding to the repetition segments (*"iie"* and *"hamamatsu inta-desu"*) of utterances are almost linearly matched. Thus, we pick such segments without matching continuously along I-axis or
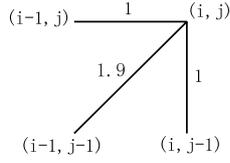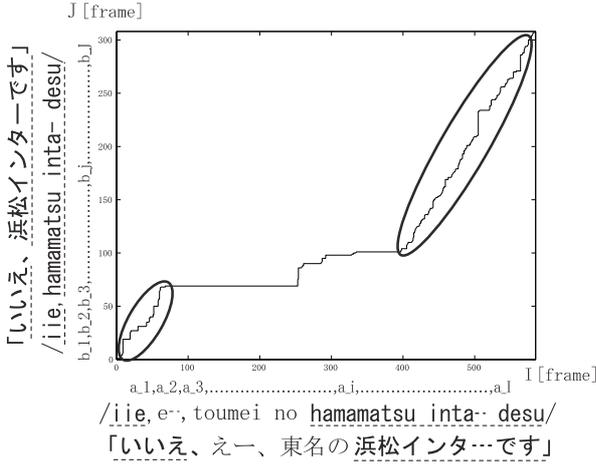
Figure 2: *DTW path and that weights*

Figure 3: *Detection of repetition based on DTW*

Figure 4: *Hypotheses for original utterance and correction utterance*

J-axis. We set the weight of a diagonal path in Fig. 2 to 1.9 to give priority to a diagonal path to match stable sounds as vowels linearly. Then we reject the segments with small lengths, large average DTW scores or large local DTW scores. We also adopted a threshold of the power to delete non-speech segments. The rests are the pairs of original and correction utterances. Finally, if one or more pairs are found in the utterance pair, the second utterance is judged as a correction.

2.2. Detection based on N-best hypotheses overlapping measure

When the utterances including same words were recognized, there tend to be many same words in the sentence hypotheses of each utterance even if the hypotheses are not correct. So we defined the hypotheses overlapping measure L as below and detect the correction utterance based on this measure:

$$L = \frac{\sum_{i=1}^{I} 2 \times \sqrt{F_{W_i,A} \times F_{W_i,B}}}{C_A + C_B} \qquad (1)$$

where $F_{W_i,A}$, $F_{W_i,B}$ mean the frequencies of word W_i in N-best sentence hypotheses for original utterance A and correction utterance B, respectively, C_A and C_B describe the numbers of all words appearing in hypotheses for utterance A and B, respectively, and I is the number of words appearing in hypotheses. The utterance B with large L is detected as a correction utterance. We used at most 200-best (average 150-best).

Fig. 4 shows an example of hypotheses for original utterance and correction utterance. There are many same words ("shizuoka ken" , "iwata eki" and "kikugawa eki") in the sentence hypotheses of both utterances.

2.3. Combination of DTW score and N-best hypotheses overlapping measure

We propose to combine above the two methods based on "repetition probability", which means the probability that the current utterance includes a repetition of the previous one. Repetition probability is defined as below. $C(W)$ equals 1 when the current utterance includes repetition of the previous one, and $C(W)$ equals 0 without repetition. We can write the repetition probability of the pair of current and previous utterances (which we call W) under the condition that the feature(s) $\mathbf{x}$ equals $\mathbf{x}_W$ as:

$$P(C(W) = 1 | \mathbf{x} = \mathbf{x}_W). \qquad (2)$$

Assuming that this probability follows the function $f(\mathbf{x})$ and $\mathbf{x}$ consists of averaged DTW score x_1 and N-best hypotheses overlapping measure x_2 and constant 1:

$$f(\mathbf{x}) \quad = \quad \frac{1}{1 + \exp(\mathbf{a}^T \mathbf{x})} \qquad (3)$$

$$= \quad \frac{1}{1 + \exp(ax_1 + bx_2 + c)}, \qquad (4)$$

where $\mathbf{x} = (x_1, x_2, 1)^T$ and $\mathbf{a} = (a, b, c)^T$. We can estimate the parameter $\mathbf{a}$ to minimize the sum of the square errors between the repetition/non-repetition (1/0) and the values of $f(\mathbf{x})$ over all training data:

$$\mathbf{a} = \arg\min_{\mathbf{a}} \sum_W \{C(W) - f(\mathbf{x}_W)\}^2. \qquad (5)$$

3. Experiments

3.1. Experimental setup

3.1.1. Training and Test Set

Training and test set were recorded by using respectively different spoken dialog systems. Table 1 shows tasks, vocabulary of the systems, number of speakers, number of recorded correction utterances and non-correction utterances. We used a dialog system for a setup of a destination for a car navigation system, whose speech recognition was driven by a context-free grammer with 15,000 word vocabulary including location names and

Table 1: *Training and Test set*

	Training Set	Test Set
Task	destination setting	reception
System vocabulary	15,000	864
Num. of speaker	10 males	10 males
Num. of correction	192	124
Num. of non-correction	320	136

facilities' names in Japan. The system used to record a test set is a prototype reception system. This system asks the name of the person to interview, appointment time and visitor's name step by step, and the recognition system driven by a context-free grammar with 864 word vocabulary including digits for phone number recognition was used for calculating N-best hypotheses overlapping measure. Training set was used to estimate each threshold for DTW and N-best hypotheses overlapping measure and the parameters a, b, c of repetition probability function.

All utterances were sampled at 12kHz and converted to 10 dimensional LPC mel-cepstral coefficients every 8 ms frames.

3.1.2. Evaluation measure

The performance was evaluated by recall rate, precision and F-measure. The recall rate expressed the rate of the number of detected correction utterance in real ones and the precision means the rate of the number of real correction utterance in detected ones as corrections. F-measure is the harmonic mean of the recall rate and the precision.

3.2. Result

Fig. 5 shows recall-precision curves on the test set for the methods based on DTW, N-best hypotheses overlapping measure and combination of the two methods when varying thresholds. The nearer to the top-right corner the line gets, the higher the performance is. The combination method outperformed the DTW-based and the N-best hypotheses overlapping measure-based methods. The discriminative performance on the test set with the thresholds and the combination parameters obtained by maximizing F-measure on the training set (open test) are shown in Table 2. As a result, we achieved recall rate of 92.7% and precision of 89.1% by combining two methods. Table 3 shows discriminative performance on the training set with the same thresholds and the combination parameters as Table 2 (that is, closed test). These results show that the performance on open test was almost same as that on closed test. Moreover, Table 4 shows the discriminative performance when the test set was used to estimate the thresholds and the combination parameters. Comparing Table 4 with Table 3, the recall and precision with N-best overlapping measure, especially the recall, differed on each test because of the difference of the vocabulary size, but all the performance were almost the same. This indicates that the thresholds and the parameters can be estimated robustly.

4. Effects on recognition performance

4.1. Application to dialog systems

We assumed to construct a system initiative dialog system. The system can confirm the recognition result at each step, but such a strategy makes users dull. Without the confirmation, users
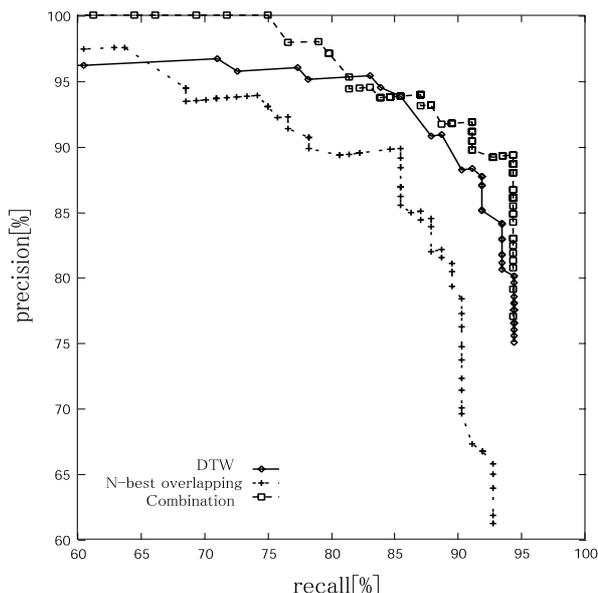

Figure 5: *Recall-precision curves on the test set*

Table 2: *Discriminative performance (Open test: thresholds and combination parameters are estimated by maximizing F-measure on the training set)*

Method	threshold	Recall	Precision	F-measure
DTW	mean=8	91.9%	87.7%	89.8
Overlapping	0.013	92.7%	62.5%	74.9
Combination	0.42	92.7%	89.1%	90.9

may repeat the utterance (that is, a correction utterance) for the previous prompt even when the system asks a new question. So the system should recognize the utterance not only with the grammar for the state of dialog but also with that for repetition which occurs after a misrecognition. This results in the performance degradation. Using the detection, the system can select either of two grammars based on the detection and, as a result, the recognition performance can be improved.

4.2. Introduction of "gray zone"

A discrimination error of correction/non-correction utterance leads to misrecognition. So, we also introduced a "gray zone" to the discrimination, which can reduce this critical effect of discrimination errors. Fig. 6 shows how to set a gray zone. When the repetition probability is between threshold γ_1 and γ_2, the recognition grammar is set assuming both cases of repetition and non-repetition. We tested two sets of thresholds: (a) γ_1=0.42, γ_2=0.6 and (b) γ_1=0.3, γ_2=0.7.

4.3. Experiments

We evaluated the recognition performances on the test set described in Section 3.1.1. The reception dialog system had 5 questions. Without the detection, system used grammars for both repetition and non-repetition to allow users to allow users to repeat the utterance on misrecognition without the detection

Table 3: *Discriminative performance (Closed test: tested on training set with the thresholds and combination parameters estimated with the same data (same as Table 2))*

Method	threshold	Recall	Precision	F-measure
DTW	mean=8	95.3%	87.6%	91.3
Overlapping	0.013	51.0%	93.3%	66.0
Combination	0.42	94.8%	89.2%	91.9

Table 4: *Discriminative performance (Tested on test set with the thresholds and combination parameters are estimated by maximizing F-measure on the test set)*

Method	threshold	Recall	Precision	F-measure
DTW	mean=8.5	92.2%	88.1%	90.1
Overlapping	0.34	85.5%	89.8%	87.6
Combination	0.38	94.4%	89.3%	91.8

and the average perplexity of the grammars was 97. Using the detection, system used a grammar for either repetition or non-repetition and the average perplexity was reduced to 54.

We compared three conditions: without detection, with ideal detection (oracle condition) and with real detection. Recognition performance without detection is the baseline, where "recognition" means recognition rates of words or phrases for embedding into slots on the reception task, for example, digit string recognition rate for a telephone number. Ideal condition means that the recall rate and the precision are both 100% which lead the upper limit of recognition performance. Under the real condition, the recall rate and precision are 91.9% and 87.7%, respectively and we also adopted gray zones.

We can find the recognition performances for correction and non-correction utterances in the test set by these methods in Table 5. Because the correction utterances tends to consist of words that are difficult to recognize, the recognition rate of correction utterances was relatively worse than that of non-correction utterances. Recognition performances were improved for both correction and non-correction utterances by the grammar constraint based on the detection of correction. Compared to the oracle, the performance improvement with real detection (without gray zone) was a little low. This is because a discrimination error always causes misrecognition. The risk of the discrimination error could be reduced by a "gray zone", and the recognition rate approached that under the oracle condition.

5. Conclusion

In this paper, we proposed correction utterance detection methods for robust dialog strategy and showed the performance improvement of speech recognition with the detection. We proposed a method to detect the repetition in spontaneously spoken dialog using a method based on DTW and N-best hypotheses overlapping measure. As a result, we achieved recall rate of 91.9% and precision of 87.7% using a method based on DTW, and recall rate of 92.7% and precision of 62.5% using N-best hypotheses overlapping measure. We also achieved recall rate of 92.7% and precision of 89.1% by combining two methods.

We tried to improve recognition accuracy using the detection. Using the choice of vocabulary and grammar setup based on the detection, we achieved improvement in recognition per-
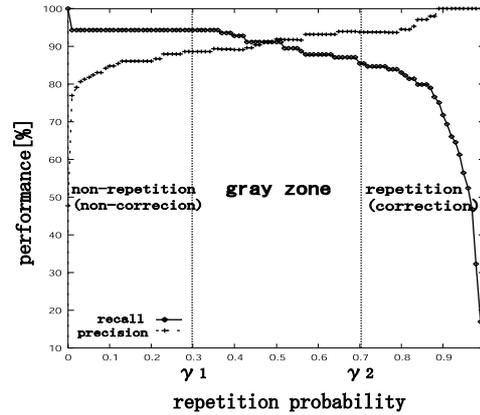

Figure 6: *Detection of repaired speech with "gray zone"*

Table 5: *Effects of correction utterance detection on recognition performance*

Method	Correction	Non-Correction
W/o detection	42.7%	70.5%
Ideal detection	51.6%	79.4%
Real detection(w/o gray zone)	49.2%	77.9%
Real detection(gray zone (a))	49.2%	78.7%
Real detection(gray zone (b))	50.0%	77.9%

formance from 42.7% to 50.0% for correction utterances and from 70.5% to 77.9% for non-correction. In addition, the correction utterance detection with "gray zone" was also effective.

6. References

[1] Swerts, M., Litman, D., Hirschberg, J., "Correction in spoken dialogue system", Proc. of ICSLP2000, Vol.2, pp.615-618, 2000.

[2] Suhm B., Myers, B., Waibel, A., "Multimodal error correction for speech user interfaces", ACM Transactions on Computer-Human Interaction, Vol.8, no.1, pp.60-98, 2001.

[3] Levow, G.-A., "Characterizing and recognizing spoken corrections in human-computer dialogue", Proc. of the COLING-ACL '98, pp.736-742, 1998.

[4] Levow, G.-A., "Adaptation in spoken corrections: Implications for models of conversational speech", Speech-Communication, Vol.36, pp.147-163, 2002.

[5] Inoue, N., Imai, N., Hashimoto, K., Yoneyama, M., "Repeated speech detection method for correcting speech recognition errors", Trans. IEICE, Vol.J84-D-II, No.9, pp.1950-1959, 2001 (in Japanese).

[6] Imai, Y., Inoue, N., Hashimoto, K., Yoneyama, M., "Detection method of repeated speech for unknown words processing", Technical Report of IEICE, SP99-26, pp.1-6, 1999 (in Japanese).

[7] Kakutani, N., Kitaoka, N., Nakagawa, S., "Detection and recognition of repaired speech on misrecognized utterances for speech input of car navigation system", Proc. of ICSLP, Vol.2, pp.833-836, 2002.