Speech Input | Distributed Speech Recognition | Error Correction

Technology Reports

Efficient Speech-recognition Error Correction for More Usable Speech-to-text Input

LVCSR technology has been widely used on mobile terminals in various applications such as mail message composition or speech translation. Since errors are unavoidable in speech recognition, an effective error correction mechanism is very important. We propose a method for error correction in LVCSR, in which the user identifies sections containing error word, and LVCSR is performed on the speech feature from the section containing the identified segment. We experimentally confirm that the proposed method is able to correct 30 to 70% of the error words.

Research Laboratories
Yusuke Nakashima
Zhipeng Zhang
Nobuhiko Naka

1. Introduction

The performance of Large-vocabulary Continuous Speech Recognition (LVCSR) has increased and applications such as writing e-mail and speech translation are becoming practical on mobile terminals. With these developments, mobile terminals from NTT DOCOMO have been providing various services utilizing LVCSR, such as the Japanese/English translation i-appli on FOMA 905i series and later terminals, and even on Raku-Raku PHONE series after Raku-Raku PHONE Premium, services like "Speech-input mail," which allows mail messages to be composed entering sentences with speech. However, because the input speech is not necessarily consistent with the acoustic and language models used for LVCSR, it is not possible to avoid recognition errors. Thus, for increasing the usability of speech recognition, an effective scheme for correcting recognition errors is important in addition to improving the performance of speech recognition.

The characteristics of the major approaches to correcting speech-recognition errors are shown in **Table 1**. Ogata et. al [1] have proposed a method presenting the N-best[*1] candidate words from the initial recognition if correction is necessary, and allowing the user to select the correct word. Correction is only possible if the correct result is included in the N-best, however. Other methods for entering words, such as key-input or re-speaking and re-analyzing the speech are provided, but these require additional user operation. We propose a correction method requiring a minimum number of user operations such as key input or re-speaking, and presenting the candidate words not included among the N-best of the initial recognition results.

The proposed method corrects the LVCSR errors by having the user specify an incorrect section of the initial recognition, and re-applying LVCSR to that part ("partial re-recognition") [2]. The method refers to and uses a

*1 **N-best**: A sequence of the most likely candidates.

Table 1 Representative recognition error-correction methods

Error correction method	Correction performance	Can change to candidates not included in initial recognition	Correction section	User-operation load	Processing method	Processing load
N-best	High	No	Word	Medium	Language	Light
Key input	High	Yes	Word	Heavy	Language	Light
Re-speak	Low	Yes	Multiple word	Heavy	Acoustic and language	Heavy
Re-recognition (proposed method)	Medium	Yes	Multiple word	Light	Acoustic and language	Medium

sequence of correctly recognized words from the initial recognition, preceding and following the incorrect section, as a constraint when searching for candidates during the re-recognition, which increases the corrective effect (increasing the recognition rate by correcting the errors).

The re-recognition process operates in the same way as ordinary LVCSR, so it has the benefit of being able to handle error sections spanning multiple words. By including a sequence of correctly recognized words preceding and following the error section, less processing is required than if re-recognizing the entire sentence. Also, it is able to use either the same or different models for the initial recognition and the re-recognition; so for example, a specialized model could be used for the re-recognition.

We have also studied an approach which automatically estimates the end point of an error section and proceeds with the re-recognition when the user has specified a starting point. This is intended to increase responsiveness to user operation, in addition to providing the corrective effect.

In this article, we describe the proposed recognition-error correction method based on partial re-recognition processing, and experiments to verify its effectiveness.

2. Correction of Speech-recognition Errors by Partial Re-recognition

2.1 Speech-recognition Processing

An overview of the speech-recognition process is shown in **Figure 1**. The recognition decoder takes unknown speech input X and searches for a word sequence $\hat{w}$. The search process searches for the word sequence which maximizes the product (sum on a logarithmic scale) of the probabilities from the acoustic model, $P(X|w)$, and the language model, $P(w)$ (the maximum-likelihood sequence). The acoustic model is a probabilistic model of the time-sequence signal of speech features[*2]. The language model assigns a probability of occurrence to the word array (N-gram[*3]).

$$\hat{w} = \arg\max_{w} P(X|w)P(w) \quad (1)$$

In fact, the most likely hypothesis is not always the correct one, and such cases result in recognition errors.

In actual recognition processing, the number of words and phonemes[*4] is usually large (several thousands or more), so the number of possible hypotheses is some power of that number, and unlikely hypotheses must be pruned in order to search effectively. As a result, even the most likely

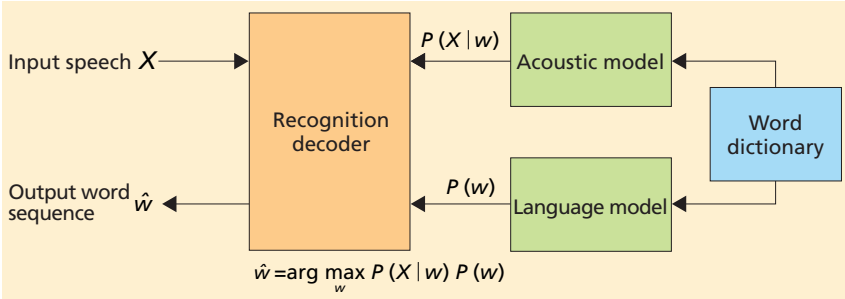


Figure 1 Overview of speech recognition process

*2 **Speech feature**: A time sequence vector of feature extracted from short-time intervals of a speech signal.
*3 **N-gram**: A chain of N words.
*4 **Phoneme**: A minimal unit of sound used to discriminate meaning in a given language.

hypothesis from all hypotheses can be pruned, and will not necessarily be output.

2.2 Re-recognition of the Correction Section

The configuration of the proposed recognition-error correction system using re-recognition is shown in **Figure 2**, and an example user interface is shown in **Figure 3**.

In order to correct errors in the initial recognition result, the user first specifies a section to be corrected by performing edit operations. Possible editing operations could be pressing the delete key or specifying the correction section using a touch panel (Fig.3(3)). Once the correction section is determined, the speech feature from a section including sequences of words preceding and following the correction section (for re-recognition section) are input to the recognition decoder (**Figure 4**). The recognition decoder uses the sequences of words preceding and following the correction section as constraints and re-recognizes the speech feature of the re-recognition section. It then presents the corrected result to the user (Fig. 3 (4)).

Restricting the initial and final words of the search hypotheses with the correct words during re-recognition is expected to have a corrective effect. Examples we confirmed are shown in **Figure 5**. With many recognition decoders, candidates with a low confidence level are pruned during the search process. Candidates with relatively high confidence level during the first recognition (Fig. 5 (a)), including "で形" and "人か" are retained, and the candidate with the correct result, "で肩" is pruned, because it has a relatively low confidence level. During

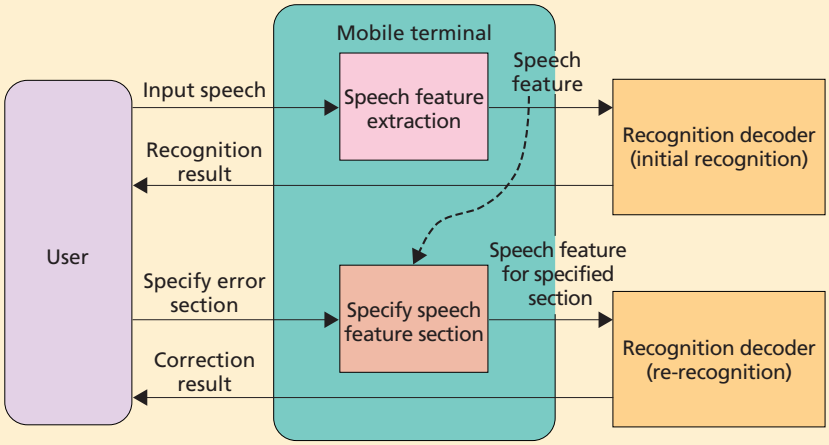


Figure 2 Example system configuration

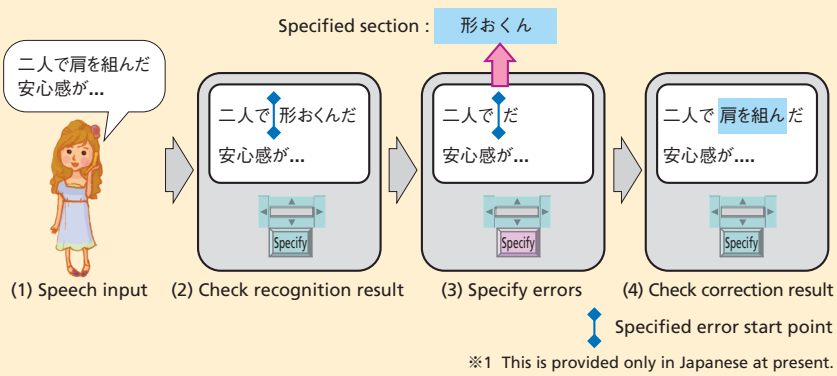


Figure 3 User interface example

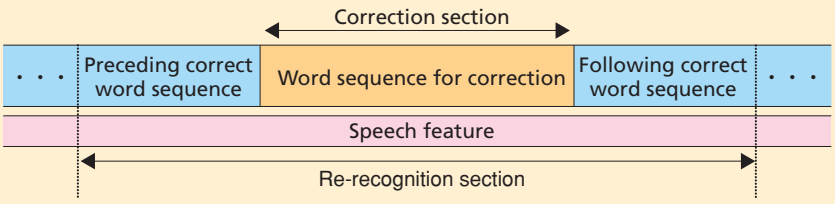
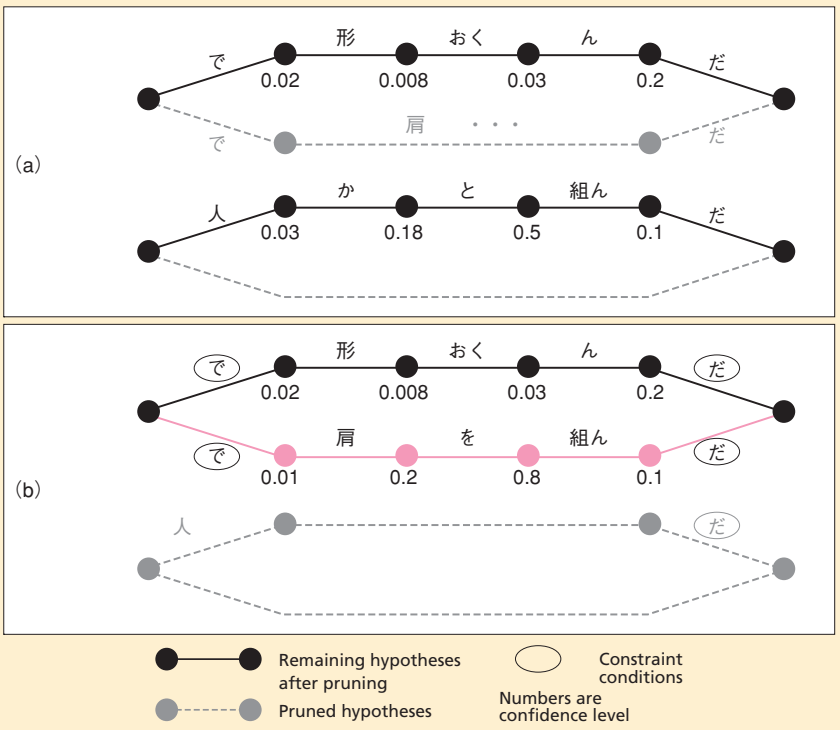


Figure 4 Section for re-recognition



※1 This is provided only in Japanese at present.

Figure 5 Example of constraining candidates using words before and after the correction section

re-recognition (Fig. 5 (b)), the search is restricted to candidates with the correct initial and final words, "で" and "だ," so the candidates connected to "人" is excluded, and "肩を組ん" is allowed to survive.

Also, by selecting the acoustic and language models used for re-recognition, we should be able to increase the corrective effect. With Distributed Speech Recognition (DSR)[*5], the high-computational-load initial recognition is performed on the server, and re-recognition, for which computational load is limited due to shorter recognition sections, can be done on the client (terminal). Then, the server can use a large-scale, general model, and the client can use a smaller-scale model tuned to the user, improving the corrective effect. Note that even if the same model is used for initial recognition and re-recognition, a corrective effect can be expected due to the constraints imposed by using words known to be correct preceding and following the error section as described above. Using the same models for both initial and re-recognition could allow a simpler, smaller-scale system to be built. For DSR in particular, retaining the speech feature and recognition results from the initial recognition on the server, would allow them to be used for re-recognition, reducing the amount of transmission between client and server, and leading to faster response.

2.3 Automatic Estimation of Re-recognition Sections

As a way of increasing responsiveness to user operation, we propose a method for estimating the end-point for a re-recognition section when the user has specified the beginning of an error section. By estimating the end point of the section, re-recognition processing can begin before the user has specified the end of the correction section. The part of the already-re-recognized result applicable to the correction section can be displayed as soon as the user specifies the end of the section, increasing responsiveness to user operation.

The words at the re-recognition end-point are used as a constraint for re-recognition, so it is desirable that they are correct from the initial recognition. We consider three possible bases for estimating the end point.

1) Confidence Level

Use the per-word confidence level resulting from the initial speech recognition process to estimate the end point for re-recognition. The first word with a confidence level above a set threshold is used as the end point for the re-recognition section and becomes the constraint (**Figure 6** (a)).

2) Number of Words

Since N-grams are usually used in LVCSR, incorrect recognition results

*5 **DSR**: A system of speech recognition in which the speech feature is extracted from the input speech on a mobile terminal, and is decoded to a recognition result on a server.

Figure 6 Example of automatic estimation of re-recognition section

usually occur in chains spanning several words. As such, the end point of the re-recognition section could be estimated in terms of a set number of words (Fig. 6 (b)).

3) Pauses

Recognition errors are less likely to occur at pauses represented by full stops (“。”) or commas (“、”) than in other locations. The nearest Short Pause (SP) after the starting point of the correction section, or the end of the sentence containing the starting point, or Long Pause (LP), could be used to estimate the end of the re-recognition section (Fig. 6 (c-1), (c-2)).

3. Experiments

In order to test the effectiveness of partial re-recognition for correcting recognition errors experimentally, we performed experiments using words known to be correct as constraints at both ends of the section being re-recognized. We also performed experiments to test the effectiveness of automatically estimating the end point by using a correct word at the start of the re-recognition section and estimating the end point automatically.

3.1 Experimental Conditions

In the experiments, we used the same acoustic model for both initial and re-recognition steps to check the effectiveness of the language model. For language models, we used a model similar to the input speech, trained from a newspaper and of 20,000 words in size, and a model with a large, general vocabulary trained from the Web, of 60,000 words in size. For the input speech, we used 100 sentences selected from a database of newspaper-article speech recordings (JNAS) (six each of male and female readers, 1,504 words) [3]. Recognition processing was done using the “Julius” [4], LVCSR engine commonly used in research and development. For speech features, 25 dimensions were used, including 12 dimensions of standard-configuration Mel-Frequency Cepstrum Coefficients (MFCC)[*6], 12 dimensions of ΔMFCC, and one dimension of Δlog power[*7]. Second-pass search with tri-gram is used to obtain the recognition result. The acoustic model uses a 16-component mixture gender-independent, 2,000 state, tri-phone[*8] Hidden Markov Model (HMM)[*9] and 129-state mono-phone[*10] [5].

3.2 Results

1) With both Ends of the Re-recognition Section Specified

We conducted experiments of re-recognition for cases with a correct word specified at both end of the section, and the results are shown in **Table 2**. The word recognition rate for the initial recognition results was 94.8% for the newspaper language model and 86.0% for the Web language model. There were 78 incorrect words with the

*6 **MFCC**: A series of speech feature coefficients modeled on human auditory perception.

*7 **Δ log power**: First order power difference.
*8 **Tri-phone**: A three-phoneme combination.
*9 **HMM**: A type of probabilistic model.
*10 **Mono-phone**: An acoustic model in which preceding and following phonemes are not considered.

Table 2 Re-recognition results

Initial recognition	Re-recognition		Word recognition rate (N=1,504)	Word correction rate
Language model	Language model	Constraint		
Newspaper			94.8%	
	Newspaper	N	96.3%	28.2% (N=78)
		Y	96.9%	41.0% (N=78)
Web			86.0%	
	Web	Y	90.2%	29.9%(N=211)
	Newspaper	Y	95.9%	71.1%(N=211)

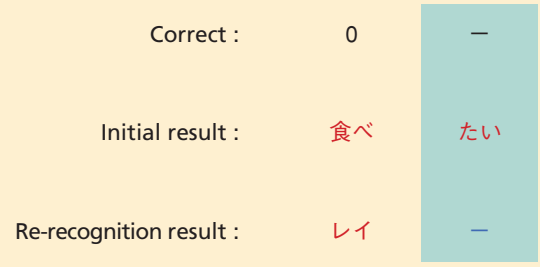
newspaper language model, and 211 incorrect words for the Web language model, to which correction by re-recognition was applied.

When the newspaper language model was used for both the initial recognition and the re-recognition (Newspaper-Newspaper), the word correction rate was 28.2% (of the 78 words) without the preceding and following correct word constraints, and 41.0% with the constraints. The re-recognition results were different from the initial results, even for cases without the constraints, because the section of the input speech is set to include the preceding and following correct words (the re-recognition section), and there are no additional constraints from outside the re-recognition section. Because of this, hypotheses different from those in the initial recognition are output as the most likely hypothesis during re-recognition, showing that even without the constraints of correct preceding and following words, re-recognition can have some corrective effect. We also confirmed that the case using constraints produces better corrective results than without the constraints. Also, as shown in Fig. 5, there are cases when a correct result is pruned in the first recognition, but is retained in the re-recognition.

Using the Web language model for the initial recognition and the newspaper model for re-recognition (Web-Newspaper), four of six words that were unknown to the initial model but known to the re-recognition model were corrected. We also confirmed that the method was able to correct recognition errors preceding or following the unknown words. For example, for the section in which the words "0ー五と" were spoken, the word for "ー," pronounced "たい," is not known, and the result from the initial recognition is "食べたい事." Upon re-recognition the result is "レイー五と," and the "ー五と" part has been corrected (**Figure 7**).

Using the newspaper language model for both initial and re-recognition (Newspaper-Newspaper) and using the preceding and following correct-words as the constraints, the combination of both results produced a 96.9% word recognition rate. We achieved word recognition rate above 95.9% for the Web-Newspaper case. Note that the Web-Newspaper recognition rate of 95.9% is higher than the initial recognition rate of 94.8% using the newspaper model.

When a language model close to the input speech is used for both the initial and re-recognition steps, recognition rates are higher than in other cases.



※1 "ー" is an unknown word in the initial recognition language model.
※2 This is provided only in Japanese at present.

Figure 7 Example results for sections containing unknown words

Even if this model is only used for the re-recognition step, we achieved results comparable or better than when it is used for the initial step. Using a model close to the input speech appears to be effective, even if only used for the initial or re-recognition steps.

On the other hand, when the Web language model was used for initial and re-recognition steps (Web-Web), the word correction rate was 29.9% (of 211 words), showing a certain corrective effect. We were able to confirm a corrective effect, even if the language model for both initial and re-recognition steps are not similar to the input speech. This confirms that error correction by re-recognition is effective, regardless of whether there is consistency between the input speech and language model used.

2) With End Point of Re-recognition Section Estimated Automatically

We also performed experiments to verify the effectiveness of using a re-recognition section starting with a correct word and with the end point estimated automatically. The word recognition rate when using the Web language model for the initial recognition was 82.4%, with 56 error sections and 264 incorrectly recognized words.

We automated estimation of error sections, performing morphological analysis[*11] using the "ChaSen" morphological analysis system and a script attached to the "Julius" package to determine the error section. The end point for re-recognition was determined using word confidence level, number of words, SP or LP (**Table 3**). The newspaper language model was used for re-recognition. Threshold values that produced the highest recognition rates were used for confidence level and number of words, resulting in recognition rates of 91.4% and 91.9% respectively. Using SP and LP, recognition rates were 93.3% and 93.6%, higher than for the cases using confidence rate or number of words. This could be because the ending word is not necessarily correct when using confidence level or number of words to determine the end of the section, and the ending word tends to be correct when SP or LP are used. Thus, even if the re-recognition could contain multiple error sections, the correction rates using SP and LP were closer to that when the end point was specified explicitly.

4. Conclusions

In this article, we have proposed a method for correcting errors in speech recognition, in which recognition is re-applied to sections containing errors specified by the user, with the preceding and following correct words as constraints. Through experiments, we confirmed that the use of constraint conditions is effective, and that re-applying recognition has a corrective effect, independently of whether there is consistency between the input speech and the language model used. In particular, when the same language model is used for the initial recognition and the re-recognition, 30% to 40% of incorrect words were corrected, indicating that improvements in performance of existing speech recognition systems may be possible without major changes.

We also proposed a method which allows re-recognition as soon as the user specifies the beginning of the section needing correction by automatically deciding the end point. We confirmed that this can produce results nearly as effectively as when the end point is specified explicitly.

We have examined ways to improve correction rates when presenting only the most likely hypotheses to the user for both the initial recognition and re-recognition, but other methods such as displaying the N-best, or that require keyboard input are candidates

Table 3 Re-recognition results when the end of sections to be re-recognized are estimated automatically

Automatic end-estimation method	Word recognition rate (N=1,504)	Word correction rate (N=264)
Confidence level	91.4%	51.1%
Number of words	91.9%	54.0%
SP	93.3%	61.9%
LP	93.6%	63.6%

*11 **Morphological analysis**: The task of dividing a sentence into its constituent morphemes, which are the smallest units of meaning.

for future study of correction methods. We will continue to study ways to further improve correction efficiency, and to introduce them into speech recognition systems.

REFERENCES

[1] J. Ogata and M. Goto: "Speech Repair: Quick Error Correction Just by Using Selection Operation for Speech Input Interfaces," in Proc. Interspeech 2005, pp. 133-136, 2005.

[2] Z. Zhang, Y. Nakashima and N. Naka: "Error Correction with High Practicality for Mobile Phone Speech Recognition," in Proc. International Workshop of Mobile HCI 2008.

[3] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Itahashi: "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research," J. Acoust. Soc. Jpn. (E), Vol. 20, No. 3, pp. 190-206, 1999.

[4] A. Lee, T. Kawahara and K. Shikano: "Julius --- an Open Source Real-Time Large Vocabulary Recognition Engine," in Proc. EUROSPEECH, pp. 1691-1694, 2001.

[5] K. Takeda, N. Minematsu, A. Ito, K. Itou, T. Utsuro, T. Kawahara, T. Kobayashi, T. Shimizu, M. Tamoto, K. Arai, M. Yamamoto, T. Takezawa, T. Matsuoka and K. Shikano: "Common Platform of Japanese Large Vocabulary Continuous Speech Recognition Research -Construction of Acoustic Model-," IPSJ SIG Technical Report, 1997-SLP-18-3, 1997.