Speech recognition using HMM-state confusion characteristics

Yumi Wakita, Harald Singer, Yoshinori Sagisaka

ATR Interpreting Telecommunications Research Laboratories
Tel: +81 774 95 1332, Fax: +81 774 95 1308, E-mail: yumi@itl.atr.co.jp

Abstract

In our previous work, we proposed a re-entry modeling of missing phonemes which are lost during search process. In the re-entry modeling, the recognition results are postprocessed and originally recognized phoneme sequences are converted to new phoneme sequences using HMM-state confusion characteristics spanning several phonemes. We confirmed that HMM-state confusions are effective for the re-entry modeling. In this paper, we propose a re-entry modeling during recognition using a multiple pronunciation dictionary where pronunciations are added using HMM-state confusion characteristics. The pronunciations are added considering part-of-speech (POS) dependency of confusion characteristics. As a result of continuous recognition experiments, we confirmed that the following two points are effective to improve word recognition rates:

(1) confusions are expressed by HMM-state sequences, (2) pronunciations are added considering part-of-speech dependency of confusion characteristics.

1. Introduction

In continuous speech recognition, though fine context dependent HMM models have been studied intensively [2, 3], it looks very difficult to expect perfect recognition for each phoneme-sized interval. Even if the conditions of acoustic modeling, e.g. number of states and mixtures are increased and huge databases are used for training, it is difficult to recover all confused phonemes. It is necessary to make a model which can recover phoneme recognition confusions so that the latter processing can handle these confusions using other knowledge such as linguistic information.

By analyzing the misrecognitions which cannot be corrected even if acoustic models are adapted, we observed that the same phoneme sequences frequently lead to the same misrecognized phoneme sequences. Furthermore, the misrecognitions depend on the previous and following context of misrecognized sequences. Therefore, if the tendency of misrecognized sequences is known a priori, it would be possible to predict the correct phonemes from the resulting, misrecognized phoneme sequences.

For recovering the recognition confusions, several confusion correction methods have been proposed using phoneme or word confusion tables[4, 5]. Though these approaches have been proven to be effective for phoneme-to-phoneme or phoneme sequences confusion, they cannot cope with the confusion in consideration of the previous and following context of misrecognized sequences.

In our previous work, we proposed a re-entry modeling of missing phonemes which are lost during the search process[1]. In the re-entry, the recognition results are postprocessed and originally recognized phoneme sequences are converted to new phoneme sequences using HMM-state confusion characteristics spanning several phonemes. A result subsequence is matched to multiple misrecognition sequences and many result candidates are generated. Reliability scores are defined as appearance frequencies of the misrecognized state sequences and the new phoneme candidates are re-entered by using left-to-right beam search to select the correct sequence which shows the best score. In this process, we confirmed the effectiveness of the HMM-state confusion characteristics spanning several phonemes for improvement of phoneme recognition rates. However the re-entry modeling should be improved as follows:

First, the re-entry uses only the final recognition results via a post-processing approach and can not use the words which already have been lost in beam space during the recognition process. It seems that the re-entry during recognition should be more effective than the re-entry by post-processing approach, because the re-entry during recognition can use the words which do not remain in the final recognition results. Next, the results of analyzing the HMM-state confusion characteristics show that some confusion characteristics depend on part-of-speech (POS). It seems that use of the confusion characteristics in consideration of the dependence on POS improve the search performance more than the already proposed re-entry modeling.

In this paper, we thus propose a word re-entry modeling during recognition process. To achieve this, a multiple pronunciation dictionary in which new pronunciations are added using HMM-state confusion characteristics are used for speech recognition. Furthermore the new pronunciations are added in consideration of the POS dependency of the confusion characteristics.

In section 2, we describe the method of multiple pronunciation dictionary generation and in section 3, we show the evaluation results of the proposed multiple pronunciation dictionary by speech recognition experiments.

Table 1: Conditions for confusion characteristics extraction

domain	travel arrangement dialogues
acoustic model	speaker adaptated HMnet, 401 states, 10 mixtures
phoneme recognition	one-pass-DP, N-best search (phoneme typewriter)[6]
training data for extraction of confusions	109 sentences/speaker, 4854 phonemes/speaker, 3 speakers

Table 2: Examples of confusions which occur only in a specific POS
correct : appearance of correct state sequences in training data
confusion : appearance of confusions in recognition results

confusions * (/correct/ → /error/)	appearance frequency by each part-of-speech part-of-speech (appearance frequency)
/z/ → /d/	*correct* : noun(23),proper noun(21),verb(24),interjection(24),adverb(2)
	confusion : verb(6),interjection(3)
/ga/ → /e/	correct : proper noun(14),interjection(20)
	confusion : interjection(8)
/o/ → /e/	correct : noun(7), proper noun(9), interjection(19),postpositional particle(11)
	confusion : proper noun(4)

∗ confusions are expressed as HMM-state sequences and only for illustration purposes expressed here as sequences of the corresponding phonemes.

2. Multiple Pronunciation Dictionary Generation using HMM-state Confusion Characteristics

2.1. Analysis on part-of-speech dependency of confusion characteristics

We counted the appearance frequency of each confusion by each part-of-speech using the database in Table 1. As a result, we confirmed that some confusion characteristics occur dependent on POS. Table 2 shows the examples of confusions which occur only in a specific POS. For example, the correct state sequences corresponding to phoneme /z/ occur in different POS but the confusion state sequences /z/ → /d/ occur only in "verb" or "interjection". Many correct state sequences corresponding to /o/ occur in "postpositional particle" and "interjection", but the confusion state sequences /o/ → /e/ occur only in "proper noun". These results show that some confusion characteristics depend on POS.

To make this dependence clearer、 we calculated correlation values between the following two distributions: one is *appearance frequency of correct sequences in training data by each POS*, and another is *appearance frequency of confusions in recognition results by each POS*. To get reliable correlation values, we select only the confusion characteristics which appear over 25 times in the recognition results. A high correlation value means that a POS including many correct sequences includes many confusions too, and therefore the confusion characteristics are independent of the POS. Inversely, a low correlation value means that a confusion occurs only in the specific POS, and the confusion characteristics depend on POS. Table 3 shows some correlation values. The upper five confusions have high correlation values and they seem to be independent of POS. The bottom three confusions show the low values and they seem to be dependent of POS. The correlation value of the confusion /z/ → /d/ in Table 2 is 0.46. This value is close to the values of the bottom three confusions in Table 2.

Table 3: Correlation values of appearance distributions between correct sequences and confusions

confusion characteristics (/correct/ → /error/)	correlation value
/eo/ → /o/	0.999
/go/ → /ko/	0.998
/s/ → /sh/	0.992
/m/ → /g/	0.972
/ch/ → /z/	0.873
...	
/m/ → /n/	0.545
/do/ → /ro/	0.541
/d/ → /t/	0.411

These results show that some confusion characteristics depend on POS. It seems that using the confusion characteristics in consideration of the part-of-speech dependency should be effective to improve the re-entry modeling.

2.2. Extraction of HMM state confusion characteristics

We extract misrecognized HMM-state sequences to express confusion characteristics. This has the following merits:

- The HMM-state confusion characteristics can express misrecognitions spanning several phonemes which are not solved by the usual HMM re-training and adaptation methods which use only phoneme-phoneme or state-state relation for updating of HMM parameters.

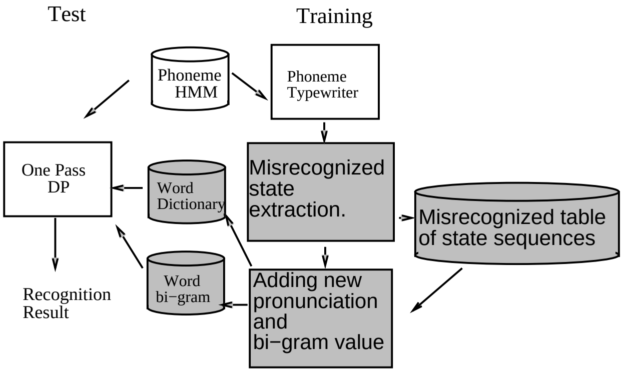
Test
Training
Phoneme HMM
Phoneme Typewriter
One Pass DP
Word Dictionary
Misrecognized state extraction.
Misrecognized table of state sequences
Word bi-gram
Adding new pronunciation and bi-gram value
Recognition Result

Figure 1: Speech recognition using HMM-state confusion characteristics

- The HMM-state sequences is useful to predict a correct phoneme sequence in consideration of the previous and following context of misrecognized sequences.

To extract confusion characteristics over HMM states, Viterbi alignment is performed on training data resulting in a *correct state sequence*. Then, an unconstrained recognition is performed on the same training data giving a (possibly) misrecognized state sequence. Then the two sequences are compared and a misrecognized state sub sequence is extracted not only for single phoneme confusions, but for variable length phoneme sequence confusions. After the misrecognized state sub sequences are extracted using all training data, an appearance frequency for each misrecognized sequence is calculated. In the case that a confusion occurs only in a specific POS, the appearance frequency is calculated for each POS. The triplet of misrecognized state sequence, correct state sequence, and a reliability score is kept in a *misrecognized table of state sequences.*

2.3. Adding pronunciations to a dictionary using HMM-state confusion

To re-enter the new words during recognition, we made a multiple pronunciation dictionary in which added pronunciations are defined by using HMM-state confusion characteristics. The process of adding new pronunciations is as follows: At first the words including the same state sequences as the correct state sequences of extracted confusion characteristics are selected from the word dictionary. Next, new pronunciations are added by using the misrecognized state sequences corresponding to the above correct state sequences. To consider the POS dependency of confusions, in the case when confusions occur only in the specific POS, the pronunciations are added only to the words which belong to the specific POS.

3. Recognition Experiment

3.1. Experimental conditions

We evaluate the proposed re-entry modeling using the continuous recognition system shown in Figure 3.1. The experimental conditions are shown in Table 4. We found that only 10% of all confusions are common across all speakers. This shows that the confusion characteristics depend heavily on the speaker. In consideration of this speaker dependency, the confusion characteristics are extracted after speaker adaptation. The multiple pronunciation dictionary is then constructed separately for each speaker. For high reliable re-entry, the confusions which appear only once or twice in the training data are not used. The number of additional pronunciations is restricted to three per word to avoid an excessive increase of the dictionary size.

To confirm the effectiveness of the proposed multiple pronunciation dictionary, the following two effects were evaluated:
(A) Effect of HMM-state confusion characteristics
(B) Effect of using the POS dependency of confusion characteristics

3.2. Effect of the proposed multiple pronunciation dictionary on speech recognition

3.2.1. Effect of HMM-state confusion characteristics

To confirm the effectiveness of using HMM-state confusions, we compared the word misrecognition rate [1] using the multiple pronunciation dictionary in which pronunciation have been added via HMM-state confusions (STATE-withoutPOS) vs. via phoneme confusions (PHONE-withoutPOS). Each word misrecognition rate is shown in Table 5. The word misrecognition rate using HMM-state confusions is lowest (11.45% (average rate)) for all speakers. The average of the misrecognition reduction rate by using HMM-state confusions is 9.5% ((12.79-11.45)/12.79). In the case of using phoneme confusions, the word misrecognition rate for speaker A and speaker C increase compared with the single pronunciation dictionary (without re-entry). The results show that it is effective for reliably improving word recognition rate to use confusion characteristics in consideration of the previous and following phoneme context of misrecognized sequences using HMM-state. On the other hand, when we do not consider phoneme context, the word misrecognition rate sometimes increases because of adding unnecessary pronunciations.

3.2.2. Effect of using the relation between confusions and POS

To confirm the effectiveness of using the POS dependency of the confusion characteristics, we compare the word misrecognition rates using a multiple pronunciation dictionary by considering the POS dependency of confusions (STATE-POS) with by not considering the POS (STATE-withoutPOS). As we have seen in section 2, only some confusions depend on POS. In this evaluation, we consider the confusion characteristics which occur only within two kind of POS as confusions which depend strongly on POS, and other confusions are used not considering POS. For this condition, 73% of confusion characteristics extracted from the database shown in Table 1 are dependent of POS.

Word misrecognition rates are shown in Table 5. In the case of considering POS, the word misrecognition rates are lower than when not considering POS

[1] based on the number of deletions substitutions and insertions

Table 4: Experimental conditions

language modeling	word bi-gram
training data for bi-gram	3476 sentences, 44147 words
continuous speech recognition	one-pass DP, N-best search, 1000 beam width
vocabulary size for recognition	2343 words
training data for extracting confusions	109 sentences/speaker, 4854 phoneme/speaker
test data	80 sentences/speaker, 907 words/speaker, 3 speakers (open to the training data for extracting confusion)

Table 5: Decrease of word misrecognition rates by using HMM-state confusions in consideration of the POS dependency

	speaker A	speaker B	speaker C	average
without re-entry	10.38%	5.79%	22.21%	12.79%
PHONE-withoutPOS	11.40%	5.33%	23.40%	13.38%
STATE-withoutPOS	9.18%	5.20%	19.96%	11.45%
STATE-POS	8.85%	3.81%	17.31%	9.99%

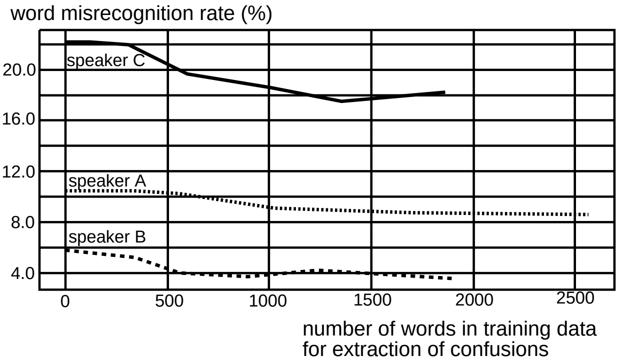

Figure 2: Relation between the decrease of word misrecognition rates and number of words in training data

(from 11.45% to 9.99% (average)). The average of the misrecognition reduction rate using HMM-state confusion considering POS dependency is 21.9% ((12.79-9.99)/12.79). These results show that it is effective to consider the POS dependency of the confusion characteristics for decreasing word misrecognition rates.

3.3. Effect of training data size for extracting confusions

In the experiments of section 3.2, we used the confusion characteristics extracted for each speaker using a speaker adaptated HMnet, because almost all confusion characteristics are dependent on the speaker. The following conditions are important to use speaker dependent confusions: (1) efficient confusion characteristics should be extractable from as few training as possible, (2) improvement of recognition rate should be stable when increasing the training data.

To investigate that the proposed re-entry modeling satisfies the above two conditions, we examined the relation between decrease in word misrecognition rate and number of words in training data separately for each speaker. The results are shown in Figure 3.3. The improvement rate of word misrecognition is different for each speaker, but there is a general trend that the word misrecognition rates decrease stably for all speakers. For 2 of our 3 speakers we noticed a significant decrease in misrecognition when using about 500 words for confusion extraction. For all 3 speakers the performance saturated at about 1000 words.

We can conclude that for applications with enough speaker specific data, e.g. a dictation task, our method is efficient and sufficiently stable. To deal with applications with less speaker specific data, we plan to make speaker dependency of confusions more clear and use speaker independent confusion characteristics or speaker-class confusions positively.

4. Conclusion

We have proposed a re-entry modeling during recognition using a multiple pronunciation dictionary where pronunciations are added using HMM-state confusion characteristics spanning several phonemes. The pronunciations are added considering POS dependency of confusion characteristics. As a result of continuous recognition experiments, we confirmed that the following two points are effective to improve words recognition rates: (1) confusions are expressed via HMM-state sequences, (2) adding pronunciations considering POS dependency of confusion characteristics.

5. REFERENCES

[1] Y. Wakita, H. Singer, Y.Sagisaka. Phoneme Candidate Re-entry modeling using Recognition Error Characteristics over Multiple HMM States. In *ESCA Workshop on Spoken Dialogue System*, pages 73–76, 1995.

[2] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul. Context-dependent modeling for acoustic-phonetic recognition of continuous speech. In *Proc. ICASSP*, pages 1205–1208, 1985.

[3] J. Takami and S. Sagayama. A successive state splitting algorithm for efficient allophone modeling. In *Proc. ICASSP*, volume 1, pages 573–576, San Francisco, 1992.

[4] A. Ito and S. Makino. Word pre-selection using a redundant hash addressing method for continuous speech recognition. In *Proc. ICSLP*, pages 309–312, 1992.

[5] E.K. Ringer, J.F. Allen. Error correction via a post-processor for continuous speech recognition. In *Proc. ICASSP96*, pages 427–430, 1996.

[6] H.Singer and J.Takami. Speech recognition without grammar or vocabulary constraints. In *Proc. ICSLP94*, pages 1994.