AUTOMATIC DIAGNOSIS OF RECOGNITION ERRORS IN LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION SYSTEMS

Hiroaki Nanjo   Akinobu Lee   Tatsuya Kawahara

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

Automatic diagnosis of recognition errors in large vocabulary continuous speech recognition (LVCSR) systems is addressed. It consists of two steps. The first step is to identify the module that causes recognition errors for every erroneous segment. This statistics points out which modules to be revised. The second step is to analyze the causes of the errors in detail. Specifically, the triphone and N-gram entries related to the errors are listed. The diagnostic information provides directions for improvement. This diagnosis has been applied to three LVCSR systems: read speech dictation system, lecture speech transcription system and dialogue speech recognition system. We have observed different and interesting diagnosis results. In the dictation system, the diagnosis is useful for improving our decoder Julius. In the lecture and dialogue speech recognition systems, problems in acoustic and language modeling are made clear.

1. INTRODUCTION

Diagnostic information of recognition errors is useful for improving speech recognition systems. It is helpful in debugging the system and revising algorithms as an immediate use. From a long-term viewpoint, it provides hint for future research directions and data collection. Since a large vocabulary continuous speech recognition (LVCSR) system is a complex of a sophisticated decoder coupled with an acoustic model and a language model, each of which are very large in scale and statistically trained with huge databases, it is not easy to manually diagnose the system given a number of recognition errors.

In this paper, we address a method to automatically identify the modules that cause errors and to generate diagnostic information for system improvement. The method is applied to three systems of different tasks: a dictation system based on Japanese newspaper corpus, a lecture speech transcription system and a spontaneous dialogue speech recognition system.

2. DIAGNOSIS PROCEDURE

2.1. Identification of Error-Causing Modules

The orthodox statistical speech recognition is formulated as finding the best word sequence W for an input speech X, such that the combined score $P(W|X) = P(W) \cdot P(X|W)$ is maximum. The framework is illustrated in Figure 1. Thus, a recognition system consists of three components: (1) an acoustic model to compute the acoustic score $P(X|W)$, (2) a language model to compute the language score $P(W)$, and (3) a decoder to search for the best hypothesis combining the above two scores.

In actual systems, the language score is amplified with some weight and an insertion penalty is added to every word transition. These factors are omitted in the explanation for simplicity, but taken into account in the implementation and experiments.

Recognition errors are counted when the recognized word sequence W_r differs from the correct sequence W_c. They are classified by the decision tree given in Figure 2. The procedure identifies which module is the cause of the error. The error is attributed to the decoder when the correct sequence W_c has a higher score than W_r but could not be found out. Otherwise, the error is the fault of either the acoustic model or the language model that gives a higher score to the recognized result W_r than W_c, or both of them. The errors caused by out-of-vocabulary words should be attributed to the lexicon and are not dealt with this procedure.

2.2. Implementation Issues

In applying the procedure, following issues must be taken into account to align the recognized sequence W_r and the correct sequence W_c.

1. Adjust variation of words and their compounds.

   There are variations of word segmentations, for example, ‘context-dependent’ can be counted as one compound word or separated into two words. This phenomenon is common in Japanese, as its text is written without space delimitation between words. In conventional definition of the word accuracy, the

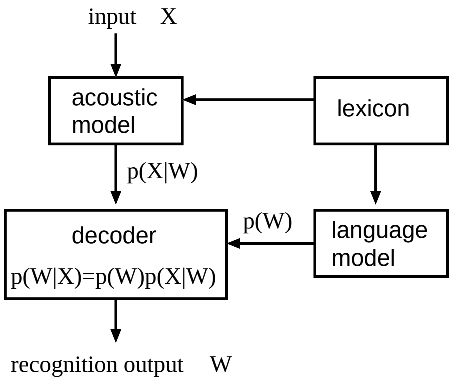

Figure 1: Framework of statistical speech recognition

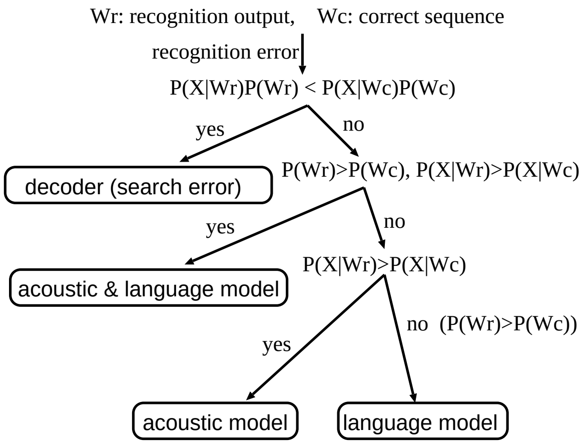
Wr: recognition output, Wc: correct sequence
recognition error
P(X|Wr)P(Wr) < P(X|Wc)P(Wc)
yes
no
decoder (search error)
P(Wr)>P(Wc), P(X|Wr)>P(X|Wc)
yes
no
acoustic & language model
P(X|Wr)>P(X|Wc)
yes
no (P(Wr)>P(Wc))
acoustic model
language model

Figure 2: Decision tree of module identification

recognized sequence is adjusted to be best matched with the correct one. In the diagnosis procedure, however, the correct sequence is adjusted to the recognized one in order to keep the word segmentations used in language model scoring at the recognition stage.

2. Handle variation of baseforms.

   Not a few lexical entries have multiple baseforms. Since different choices of baseforms cause different acoustic scores, the matching is based on baseforms, not lexical entries. Correct baseforms are estimated with Viterbi algorithm. The difference can be tolerated as in the usual word accuracy through a post-processing.

3. Handle pauses between words.

   Existence of pause models between words is not counted in the conventional word accuracy, either. But insertion of a pause model affects the acoustic score together with the choice of adjacent context-dependent phone models. In some systems, pause models correspond to end-of-phrase and end-of-sentence symbols, thus affects the language score, too. Therefore, pauses are inserted into the correct word sequence with Viterbi algorithm and adjusted to the recognized sequence if necessary.

2.3. Segmentation of Minimal Erroneous Sequences

When recognition errors occur in multiple segments of an input utterance, the identification procedure should be applied to each of them instead of the whole input for precise analysis. Here, we must note that both the acoustic score and the language score are influenced by adjacent words. Specifically, the word N-gram score can be affected by any recognition errors of preceding (N-1) words. The acoustic score is affected if either side of adjacent contexts is incorrectly recognized.

Considering these factors, we segment the whole sequence W_r and W_c into minimal erroneous portions that are independent of other errors. In other words, two segments are merged if one of them affects the other. For every segment, the specific module that caused the error is identified. The diagnosis provides guidance on which modules should be revised or re-trained to improve the whole system performance.

2.4. Causal Analysis of Errors

In order to obtain more detailed diagnostic information, the set of erroneous segments and the identified modules are investigated.

When the acoustic model is judged as the cause of errors, its phonetic contexts are investigated. Specifically when triphone models with some context-clustering techniques are used, missing contexts and logical contexts that are merged into others are reported with their statistics. The diagnosis provides suggestions on which triphones should be re-trained and hints on general contextual modeling.

When the language model is the cause of errors, N-gram entries concerned are listed with their statistics. Especially, those entries that are back-off smoothed are reported. The diagnosis reports tendency of patterns that are poorly modeled or insufficiently trained.

For search errors that are attributed to the decoder, the corresponding segments are traced and scores of competing hypotheses are reported frame-by-frame.

3. APPLICATION TO DICTATION SYSTEM

At first, the diagnosis procedure is performed on a Japanese dictation system that we are currently developing. The system is developed as a free sharable LVCSR software under the collaboration of major academic institutes of Japan and

Table 1: Diagnosis of 5K system (read speech)

decoder	acous.	lang.	both	search	total
Julius-1.1	0.8%	1.4%	1.5%	4.0%	7.6%
Julius-2.1	0.5%	1.6%	1.7%	1.7%	5.4%

Word Error Rate

Table 2: Diagnosis of 20K system (read speech)

decoder	acous.	lang.	both	search	total
Julius-2.1	0.8%	1.8%	0.9%	4.1%	7.6%
Julius-3.0	0.9%	1.2%	1.5%	2.0%	5.7%

Word Error Rate

a governmental support[1]. The specification of the modules is described in [2].

The evaluation task is 5K-word and 20K-word dictation of Japanese newspaper (JNAS) corpus. We use a gender-dependent triphone HMM of 2000 states and 16 mixture components that is trained with 20K sentences by 132 speakers for each gender. The language model is trained with the corpus of Mainichi newspaper articles of 7 years.

Since the authors (Kyoto University) take the part of development of the decoder program named Julius[3], we have used the diagnosis tool for improving the algorithms and debugging the program.

The diagnosis of 5K and 20K systems are shown in Table 1 and 2. In the 5K system using Julius 1.1, more than half of errors were attributed to the decoder. Tracing erroneous segments attributed to the decoder, we could fix several problems efficiently and reduced the error rate to 5.4%.

The 20K system using Julius 2.1 generated more errors attributed to the decoder than the 5K system. Among total error rate of 7.6%, more than half (4.1%) were caused by the decoder, and they were mainly related to one syllable words. Our decoder at that time (Julius 2.1) did not deal with inter-word triphone in the first pass, but applied it in rescoring. However, the diagnosis result motivated us to revise it to apply inter-word triphone in the first pass. The modification significantly reduced the search errors to almost half (from 4.1% to 2.0%), thus improved the overall system performance (from 7.6% to 5.7%) in the 20K system using Julius 3.0.

4. APPLICATION TO LECTURE SPEECH TRANSCRIPTION SYSTEM

Table 3: Diagnosis of lecture speech recognition system

Speaker	acous.	lang.	both	search	OOV	total
A	2.5%	3.7%	5.1%	4.2%	2.2%	17.7%
B	5.2%	3.4%	4.5%	6.4%	4.5%	24.1%
C	5.0%	5.1%	9.4%	6.0%	2.8%	28.3%

Word Error Rate

Table 4: Most frequent triphones in erroneous segments attributed to acoustic model (counts)

o-n+o (50)	sh-i+t (43)	a-sh+i (37)	i-t+a (30)
i-t+e (28)	i-m+a (28)	o-t+o (27)	w-a+sp (20)
k-o+t (19)	m-a+sh (19)	e-w+a (19)	t-o+sp (19)
a-k+u (19)	o-k+u (17)	o-d+e (17)	k-o+n (17)
a-r+i (17)	a-s+u (17)	o-k+o (16)	o-n+i (15)
i-sh+o (15)	m-a+s (15)		

Next, this method is applied to a lecture speech transcription system [4]. The test-set consists of oral presentations by 3 males, and the acoustic model is trained with read speech, exactly same as the one described in chapter 3. The language model is trained with numerous lecture transcriptions collected via World Wide Web. We got an error rate of 23.4%, and about 70% of them were caused by the acoustic or language model as shown in Table 3.

When we pick up most frequent triphones appeared in erroneous segments attributed to the acoustic model (Table 4), we find almost all of them are concerning functional words and auxiliary verbs that are not clearly articulated, for example “konoyouni” and “-node”. Also the triphones following short pauses are found, for example “-wa ,” and “-to ,”. Lecture speech is usually uttered faster than read speech and contains abrupt pauses. Thus, the acoustic model trained with read speech can not cope with such phenomena.

Table 5: Ratio of interjections and pauses in missing N-gram entries (counts)

speaker	2-gram	3-gram
A	48%(73)	71%(112)
B	34%(98)	56%(147)
C	61%(189)	78%(343)
total	51%(360)	71%(602)

In erroneous segments attributed to the language model, we list up missing N-gram entries. Over half of them are related to pauses and interjections (Table 5). In the lecture transcriptions by human, they were deleted in editing, thus not properly modeled. N-gram entries related to interjections and pauses should be estimated by the raw transcriptions of lectures or they should be handled as transparent words in decoding [5].

Table 6: Diagnosis of dialogue speech recognition system

acoustic model training database	acous.	lang.	both	search	total
dialogue (ATR)	1.5%	1.3%	2.6%	6.3%	11.8%
read (ASJ)	8.5%	2.7%	4.2%	7.3%	22.9%

Word Error Rate

Table 7: Most frequent triphones in erroneous segments attributed to acoustic model (counts)

i-m+a (18)	d-e+s (16)	e-s+u (15)	o-sh+i (11)
m-a+s (11)	N-d+e (10)	sh-i+t (10)	i-t+e (9)
i-t+a (9)	q-t+o (9)	s-u+k (8)	o-k+u (7)
a-k+u (7)	i-sh+i (7)	e-d+o (7)	g-a+i (7)
a-i+sh (7)	a-r+a (7)	u-n+o (7)	o-r+e (7)
o-n+o (7)	s-o+r (7)	t-e+r (7)	a-s+u (7)

acoustic model training database: read (ASJ/JNAS)

5. APPLICATION TO SPONTANEOUS DIALOGUE SPEECH RECOGNITION SYSTEM

Finally, we apply the diagnosis procedure to a spontaneous dialogue speech recognition system [6]. We used a large spontaneous speech database developed at ATR [7] and achieved error rate of 11.8% (Table 6). This shows the significance of matching the training database to the recognition task.

For reference, we also tested the system by replacing the acoustic model with the one trained with read speech. This doubled the error rate (Table 6). Picking up the most frequent triphones appeared in erroneous segments attributed to the acoustic model (Table 7), most of them were caused by the acoustic model in the contexts of Japanese auxiliary verbs such as “-shimasu” and “-desu” and demonstrative pronouns such as “ano” and “sore”. The former words scarcely appear in read speech, and the latter words have different acoustic patterns from those of read speech. The errors that belong to the latter case are frequently observed. For example, the word “-wo”, “-nai”, “ano” in dialogue speech matches with the word sequence “-wo o o”, “-nai i”, “ano o”. This is caused by poor modeling of the pronunciation ambiguity and the tendency of a vowel to be a long vowel in spontaneous Japanese. Although the word “ano” has a notation **/ano/** as a baseform, it is often realized as **/ano:/** in spontaneous speech.

In the system using the acoustic model trained with spontaneous speech, this sort of pronunciation variations are well modeled, thus the errors attributed to acoustic model are reduced. However, since the pronunciation of several words such as “-shimasu”, “-desu” and “-nde” varies largely, there are still remaining difficult problems in the acoustic modeling of spontaneous speech.

Next, we investigate the cause of many errors attributed to the decoder in spite of small perplexity. In erroneous segments attributed to the decoder, a lot of one-syllable interjections are found. The language model is trained with real spontaneous dialogue including interjections, thus the probability of generating interjections is quite high. This resulted in false alarms of interjections in ambiguous portion of input speech.

6. CONCLUSIONS

We have presented an automatic diagnosis of large vocabulary continuous speech recognition systems. It identifies the module that is concerned with every minimal erroneous segment, and then investigates the causes of errors in detail.

The diagnosis method has been performed on three LVCSR systems of different tasks. The analysis in the dictation task helped us improve the decoder. In spontaneous lecture and dialogue speech recognition task, it is made clear that problems mainly arise in acoustic modeling of ambiguous speech segments such as auxiliary verbs and functional words, and in language modeling of interjections and short pauses. The diagnostic information is useful for future directions of research and system improvement.

References

[1] T.Kawahara, T.Kobayashi, K.Takeda, N.Minematsu, K.Itou, M.Yamamoto, T.Utsuro, and K.Shikano. Sharable software repository for Japanese large vocabulary continuous speech recognition. In *Proc. Int’l Conf. Spoken Language Processing (ICSLP)*, pp. 3257–3260, 1998.

[2] T.Kawahara, T.Kobayashi, K.Takeda, N.Minematsu, K.Itou, M.Yamamoto, A.Yamada, T.Utsuro, and K.Shikano. Japanese dictation toolkit – plug-and-play framework for speech recognition R&D –. In *Proc. IEEE workshop on Automatic Speech Recognition and Understanding*, 1999.

[3] A.Lee, T.Kawahara, and S.Doshita. An efficient two-pass search algorithm using word trellis index. In *Proc. Int’l Conf. Spoken Language Processing (ICSLP)*, pp. 1831–1834, 1998.

[4] K.Kato, H.Nanjo, and T.Kawahara. Automatic transcription of lecture speech using topic-independent language modeling. In *Proc. Int’l Conf. Spoken Language Processing (ICSLP)*, 2000.

[5] A. Stolcke and E. Shriberg. Statistical language modeling for speech desfluencies. In *Proc. of ICASSP*, pp. 405–408, 1996.

[6] M.Mimura and T.Kawahara. Effective training of acoustic model for dialogue speech by incorporating read speech corpus. In *Proc. Int’l Conf. Spoken Language Processing (ICSLP)*, 2000.

[7] T.Takezawa, T.Morimoto, and Y.Sagisaka. Speech and language database for speech traslation research in atr. In *In Proc. Oriental COCOSDA workshop*, pp. 148–155, 1998.