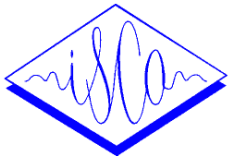

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

AN ASSESSMENT OF AUTOMATIC RECOGNITION TECHNIQUES FOR SPONTANEOUS SPEECH IN COMPARISON WITH HUMAN PERFORMANCE

Takahiro Shinozaki and Sadaoki Furui

Tokyo Institute of Technology
Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan.
{staka, furui}@furui.cs.titech.ac.jp

ABSTRACT

To investigate problems of spontaneous speech recognition using N-grams and HMMs and estimate the room for improvement in the recognition rate, an automatic speech recognizer is evaluated in comparison with performances by human listeners. The evaluation task is to recognize spontaneous speech presentations from the Corpus of Spontaneous Japanese. Both the automatic recognizer and human listeners are requested to choose the most likely word from a dictionary, given a speech signal with a three word length including $\pm$ one word context extracted from a presentation. Recognition performances are compared using the same criteria for both experiments. The results show that recognition error rate by human listeners is roughly half of that by the recognizer. By examining words that are easy for humans but difficult for the recognizer, it is found that causes of the recognition errors by the decoder include insufficiency of model accuracy and lack of robustness against vague and variable pronunciations.

1. INTRODUCTION

Recognizing spontaneous speech is a challenging task and it needs significant improvement before approaching an application level. There is a large variation of difficulties in recognizing words in spontaneous speech; while some words are easy, others require specific knowledge or longer context. Some words are impossible to recognize even for humans. Recognition methods need to be improved in different ways for each class of words according to the variation of the difficulties. However, it is unclear what are the most important problems and to what extent they are significant.

In order to investigate the possibility of improvement, we evaluate recognition performance of an automatic speech recognizer in comparison with human recognition performance.

To evaluate the difference of recognition performances between computer (decoder) and humans, several comparisons were conducted. For speech reading text, an order of magnitude higher word error rate was reported when comparing decoders with human listeners using sentences extracted from CSR'94 spoke 10 and CSR'95 Hub3 database under various SNR and microphone conditions [1]. Another experiment using sentences extracted from the Wall Street Journal database indicated roughly a five times higher error rate for a decoder [2]. For spontaneous speech, an order of magnitude higher word error rate for a decoder was reported for the Switchboard task [3].

This paper explores the possibility of improvement in spontaneous speech recognition by searching for conditions where the prescription would be relatively easy. For this purpose, the decoder and human performances are compared in restricted word contexts. Recognition results from the decoder and listeners are compared word by word and reasons for the errors by the decoder are analyzed.

2. EXPERIMENTAL SET UP

2.1. Recognition task

Recognition results are evaluated using the same recognition task performed by an automatic speech recognizer and human listeners. The task is to recognize a word in an excerpted period of utterance including a $\pm$ one word context. Both the decoder and human listeners choose the most likely words from the same vocabulary set, which consists of the most frequent 25k words occurring in 455 academic presentations in the Corpus of Spontaneous Japanese (CSJ) [4]. The presentations in the corpus have been given spontaneously and recorded using close-talking microphones.

Five hundred test utterances are randomly chosen from seven academic presentations in the CSJ given by different male speakers. Table 1 shows the contents of the test set

Table 1. Test set presentations

Presentation ID	Conference
A01M0035	Acoust. Soc. Jap
A01M0007	Acoust. Soc. Jap
A01M0074	Acoust. Soc. Jap
A02M0117	Soc. Jap. Linguistics
A03M0100	Assoc. Natural Lang. Proc.
A05M0031	Phonetics Soc. Jap.
A06M0134	Assoc. Socioling. Sciences

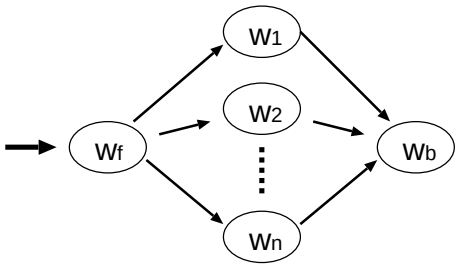

Fig. 1. Word network for the decoder.

presentations. Each test utterance is a three-word sequence excerpted from the presentations by forced alignment of an HMM sequence corresponding to the true word sequence. Since there is no spacing between words in Japanese sentences and even no clear definition of words, we used the JTAG morphological analysis program to define words. The JTAG is also used to annotate pronunciations of the words. We manually checked the resulting pronunciations so that errors did not affect the segmentation accuracy. Utterances with severe errors were eliminated after the random selection of the test set. Approximately one percent of the center words of the 500 test utterances, target words to recognize, were not included in the vocabulary. In the evaluation process, recognition results are manually checked and simple transcription variations are normalized. The major reason why we give the $\pm$ one word context, instead of the previous two words, is to avoid explicitly determining word boundaries of the center word in the wave form which is sometimes difficult due to coarticulation effects.

2.2. Recognition by decoder

For recognition by the decoder, a word network, as shown in Figure 1, is prepared for each test utterance. Finding the most likely path in the network corresponds to choosing the center word given the $\pm$ one word context. Note that in the decoding process, the optimum word boundary may be different from path to path. A language probability is assigned to each center word in the network as shown in equation (1).

$$P(w_c|w_f, w_b) \tag{1}$$

$$= \frac{P(w_f) \cdot P(w_c|w_f) \cdot P(w_b|w_f, w_c)}{\sum_w P(w_f) \cdot P(w|w_f) \cdot P(w_b|w_f, w)}. \tag{2}$$

Here w_c is a center word, and w_f and w_b are the front and back context words, respectively. The conditional probability of equation (1) is calculated using a trigram language model as shown in the equation (2).

The language model is trained using a corpus with 2.9M words consisting of 1289 academic and non-academic presentations given by both male and female speakers. Acoustic feature vectors have 25 elements consisting of 12 MFCC, their delta and the delta log energy. The CMS (cepstral mean subtraction) is applied to the sentence utterance including each three-word length test utterance. A tied state triphone model consisting of 2k states and 16 Gaussian mixtures in each state is used as a speaker independent (SI) acoustic model. The model is trained using 455 academic presentations in the CSJ given by male speakers, which has a total length of 94 hours. In addition to the SI model, speaker adaptive (SA) acoustic models are also constructed using an unsupervised adaptation method. The SI model is adapted for each speaker with the MLLR technique using the entire presentation. There is no overlap between the speakers in the training set and those in the test set.

The HTK was used for decoding. A language weight of 10 was determined by preliminary experiments. A relatively light pruning level that was also determined by preliminary experiments was used so as not to affect the recognition rate.

2.3. Recognition by human

Human listeners are given the capability of playing back the test utterances and choosing words from the vocabulary using a GUI-based system. They can listen to the same utterance as many times as they like to make a decision. However, once they make a decision, they cannot repeat the same task. Since the utterances are randomly selected from presentations, it is impossible for the human listeners to use a longer context beyond $\pm$ one word. They are informed that target words in the test utterances might not be in the vocabulary, and instructed to select the closest word even when they do not find the exact word. To facilitate finding the words from the large vocabulary, the GUI is equipped with a dictionary search using regular expressions.

Fifteen listeners, consisting of 14 male and one female, were divided into five groups, each having three listeners. They were students and staff of our laboratory. The 500 test utterances were partitioned into five blocks and each block was assigned to one of the groups. The same utterance is recognized by three different listeners in each group

to mitigate the effects of careless mistakes and individual variations due to differences of familiarity with presented topics. An upper limit of human recognition ability is estimated by determining the selected word based on a majority rule among the three listeners. The estimated upper limit is used for the comparison with the results by the decoder. If there is no overlap between the words given by the three listeners, an answer by the listener having overall the best performance among the three listeners is adopted.

The listeners practiced the task using 10 examples before performing for the test utterances. Experiments were conducted in an office using a headphone. It took about one to two hours for a listener to process the 100 test utterances.

3. EXPERIMENTAL RESULTS

3.1. Human recognition results

Figure 2 shows the recognition performance of individual listeners and that by the majority rule. Unknown words are not counted in the recognition rate. There are no insertion or deletion errors because of the experimental settings. The variation in score from listener to listener is mostly due to a difference in familiarity with academic presentations. Averaged recognition rate of the majority-rule based result is 95.3%.

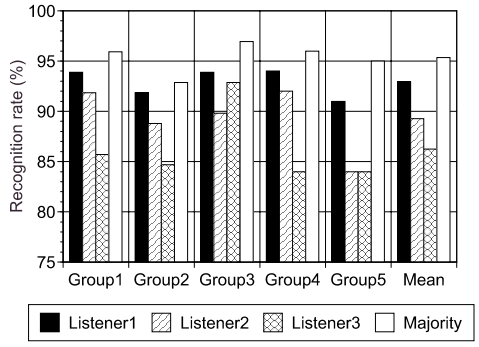

Fig. 2. Human recognition rates.

3.2. Comparison between decoder and human

Comparison of recognition rates by human listeners and a decoder is shown in Figure 3. The majority-rule based result is used as the human recognition rate. The averaged recognition rate of the decoder is 88.7% when the SI model is used, and 91.3% when the SA models are used. The human recognition rate is superior to that of the decoder under the same conditions defined for the context. The recognition error rate for human listeners is roughly half of that for the decoder. The differences of the recognition/error rate between humans and the decoder are significant at 1% level for the SI results, and at 5% level for the SA results.

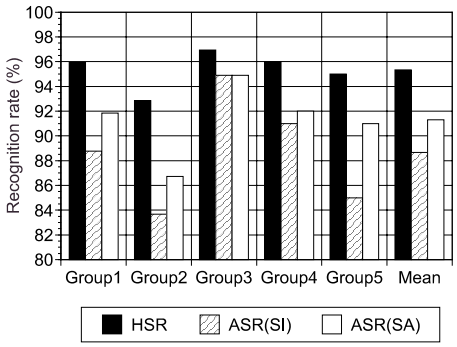

Fig. 3. Human and decoder recognition rates.

3.3. Analysis of decoding errors

Table 2 shows the classification of the experimental results using the SI acoustic model. The results based on the majority rule are used in the case of human experiments. There exist twelve words that are successfully recognized by the decoder but not by human listeners. The reasons for the errors made by the humans include vague pronunciations which make the recognition difficult and cause the same inattention errors by two or more listeners at the same time.

Table 2. Classification of recognition results

		ASR	
		Correct	False
HSR	Correct	426	45
	False	12	11

UNK: 6

There are 45 words that can be correctly recognized by humans but not by the decoder. Among these 45 words, 33 words are correctly recognized by all three listeners. If we could improve the decoder so that these words can be correctly recognized, a 6% improvement in the accuracy could be expected. In order to investigate why the decoder failed to recognize these words, we have compared acoustic and linguistic likelihood values of the true words and the outputs of the decoder. The result is shown in Figure 4. The acoustic likelihood is calculated including the $\pm$ one word context, and the language weight is incorporated into the language likelihood.

There are 13 samples in which the acoustic likelihood of the incorrect hypothesis word is lower but the language

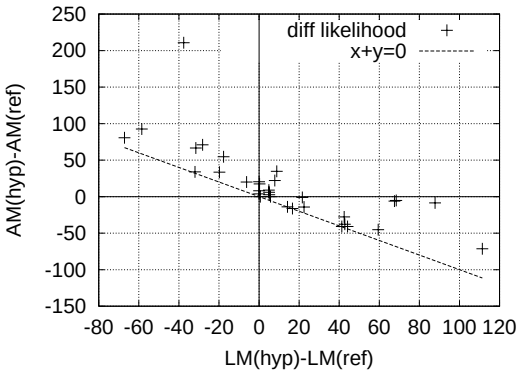

Fig. 4. Comparison of likelihood values.

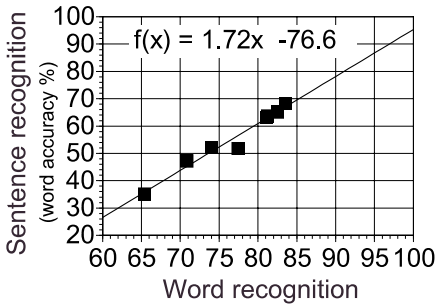

Fig. 5. Word vs. sentence recognition rates.

likelihood is higher than the true word. The inversion of the language likelihood is due to either an excessive likelihood assignment to the incorrect hypotheses or an unusual occurrence of the correct three word sequences; both are almost equally observed. The excessive likelihood seems to be caused by a backing off applied because of the sparsity of the training data. To recognize the unusual true word sequences, improvement of the acoustic model is also required.

On the other hand, there are nine samples with which the language likelihood of the misrecognized word is lower but the acoustic likelihood is higher than the true word. Among these samples, one of them is totally unvoiced and another is contaminated with a low noise. The other seven samples have no problem as far as the three-word sequences are listened. But when the center words are listened to in isolation, roughly half of the seven samples sound somewhat different from the correct word.

3.4. Relationship with continuous speech recognition

Word recognition rates of individual word recognition, given the $\pm$ one word context, conducted in the above experiments were compared to recognition rates using whole sentence continuous speech using various acoustic and language models which have different modeling accuracy. The results are shown in Figure 5. In this experiment, 3000 words and 280 sentences in the test set presentations listed in Table 1 are used. The recognition rate of this result is slightly lower than that of the task in subsection 3.2 even when the same model is used, since the results are not manually normalized.

The relationship between the recognition rates can be approximated by a straight line with an inclination coefficient 1.72, that passes through the point of $(100\%, 96\%)$. If the individual word recognition performance improves by 6% as stated in the previous subsection, a 10% improvement in the continuous speech recognition is expected.

4. CONCLUSION

Recognition performance of an automatic speech recognizer has been evaluated in comparison with human recognition performance in spontaneous presentation recognition. The recognition error rate of human listeners is roughly half of that of the decoder. There exist roughly 6% of words that are easy for humans to recognize but difficult for the decoder. Causes of the recognition errors by the decoder include the problems of model accuracy and lack of robustness against vague and variable pronunciations. If the decoder could be improved to overcome these problems and the 6% of words could be correctly recognized, approximately a 10% improvement is expected in continuous speech recognition without using contexts longer than trigrams.

5. REFERENCES

[1] N. Deshmukh, R. J. Duncan, A. Ganapathiraju, and J picone, "Benchmarking human performance for continuous speech recognition," in *Proc. ICSLP*, 1996, vol. 4, pp. 2486–2489.

[2] D. A. van Leeuwen, L. G. Van den Berg, and H. J. M. Steeneken, "Human benchmarks for speaker independent large vocabulary recognition performance," in *Proc. Eurospeech*, 1995, vol. 2, pp. 1461–1464.

[3] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, pp. 1–15, 1997.

[4] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki, and T. Ohdaira, "Toward the realization of spontaneous speech recognition," in *Proc. ICSLP*, Oct. 2000, vol. 3, pp. 518–521.