ERROR ANALYSIS USING DECISION TREES IN SPONTANEOUS PRESENTATION SPEECH RECOGNITION

Takahiro Shinozaki and Sadaoki Furui

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

ABSTRACT

This paper proposes the use of decision trees for analyzing errors in spontaneous presentation speech recognition. The trees are designed to predict whether a word or a phoneme can be correctly recognized or not, using word or phoneme attributes as inputs. The trees are constructed using training “cases” by choosing questions about attributes step by step according to the gain ratio criterion. The errors in recognizing spontaneous presentations given by 10 male speakers were analyzed, and the explanation capability of attributes for the recognition errors was quantitatively evaluated. A restricted set of attributes closely related to the recognition errors was identified for both words and phonemes.

1. INTRODUCTION

To promote better understanding and to build technology for spontaneous speech recognition, the Science and Technology Agency Priority Program (Organized Research Combination System) entitled "Spontaneous Speech: Corpus and Processing Technology" started in 1999 under the supervision of Furui [1]. A large-scale spontaneous speech corpus named “Corpus of Spontaneous Japanese (CSJ)” is under construction by the project.

Previous study showed that acoustic and language models made using the CSJ were significantly superior to conventional read-speech-based models in spontaneous speech recognition [2]. However, the recognition accuracy is still rather low, and there might be many factors that affect recognition performance acoustically as well as linguistically. Analyzing these factors is crucial to improve the recognition accuracy.

This paper proposes an application of decision trees to analyze recognition errors. Words/phonemes contained in speech have many attributes, and the choice of a given word/phoneme by the speech recognition is either true (correct) or false (incorrect). To map the attributes to a true/false class, decision trees can be employed. We expect the prediction capacity of a tree to be related to the explanation capability of the set of attributes used in this tree. In addition, we investigate how these attributes cause recognition errors by analyzing the trees.

A “case” is defined as a set of attributes and a class. A decision tree is trained by using a set of cases. The performance of the tree is measured by applying the tree to a set of test cases and calculating what percentage of the classes are correctly predicted. Even if attributes having no useful information for predicting recognition errors are included, they will be harmless if the tree is constructed properly. Using trees has the advantage that any attribute can be taken into consideration.

This paper is organized as follows. In section 2, we show a speech recognition task and an experimental set up of the presentation speech. In section 3, we first review the principle of constructing decision trees, and then we show the construction and evaluation set up of trees. In sections 4 and 5, we show the current recognition performance and the experimental results of decision trees. Finally in section 6, we conclude with a general discussion and issues related to future research.

2. SPEECH RECOGNITION TASK AND EXPERIMENTAL SET UP

2.1. Recognition task

Presentation speech uttered by 10 male speakers was used as a test set for speech recognition. Table 1 shows contents of the test set.

Table 1 Recognition test set of presentations

ID	Conference name	Length [min]
A22	Acoust. Soc. Jap.	28
A23	Acoust. Soc. Jap.	30
A97	Acoust. Soc. Jap.	12
P25	Phonetics Soc. Jap.	27
J01	Soc. Jap. Linguistics	57
K05	National Lang. Res. Inst.	42
N07	Assoc. Natural Lang. Proc.	15
S05	Assoc. Socioling. Sciences	23
Y01	Spont. Speech Project Meeting	14
Y05	Spont. Speech Project Meeting	15

2.2. Experimental conditions

Speech signals are digitized with 16kHz sampling and 16bit quantization. Feature vectors have 25 elements consisting of 12 MFCC, their delta and the delta log energy. The CMS (cepstral mean subtraction) is applied to each utterance. HTK v2.2 is used for acoustic modeling. Language models are made by using the CMU SLM Tool Kit v2.05. Morphemes (which will be called "words" hereafter in this paper) are used as units for statistical language modeling. The Julius v3.1 decoder [3] is used for speech recognition.

The language model weights and the insertion penalties are chosen to maximize the recognition accuracy of the test set. Filled pauses and repairs are taken into account as words in calculating the recognition accuracy.

2.3. Language and acoustic modeling

A part of the CSJ completed by the end of December 2000, having approximately 1.5M words, is used as a training set. Speakers have no overlap with those of the test set. The training set consists of 610 presentations; 274 academic conference presentations and 336 simulated presentations. The simulated presentations talking about a wide variety of topics including the subjects' experiences in their daily lives were specially recorded for the project.

The language model used in the recognition consists of bigrams and reverse trigrams with backing-off. It is made using the whole training set. The vocabulary size is 30k. The acoustic model is made using 338 presentations uttered by male speakers (approximately 59 hours). It is a tied-state tri-phone HMM having 2k states and 16 Gaussian mixtures in each state.

3. TRAINING AND TESTING DECISION TREES

3.1 Tree construction

The decision trees are made using a data-mining tool called C4.5R8[4]. In C4.5, trees are derived by a two-path strategy. First, questions about attributes are chosen step by step under a predefined criterion. Training cases are split by the question accordingly. This partitioning continues to subdivide the set of training cases until each subset in the partition contains cases of a single class, or until no question yields any improvement. Next, to correct over-training and make the tree robust against unseen data, the tree is pruned.

In this experiment, gain-ratio is employed for the question choosing criteria, which is default in C4.5. Questions that maximize the gain-ratio are selected. Equation (1) shows the definition of the gain-ratio.

$$gain\,ratio = \frac{H(Y) - H(Y \mid X)}{H(X)}, \tag{1}$$

where X is a random variable defined for each question, whose value is its answer. $H(X)$ denotes the entropy for the distribution of X. $H(Y)$ denotes the entropy for the distribution of a class. $H(Y \mid X)$ is the conditional entropy of the distribution of a class given an answer to the question. Entropy is calculated based on the distribution of the training cases for each tree node.

3.2 Decision trees for words

Decision trees for words are constructed by defining a case as a set of attributes of a reference word and the correctness of its recognition hypothesis. The correctness is determined by matching the reference word sequence and recognition hypothesis. Since compound words are not considered in the matching process, the errors include the cases where only word segmentation boundaries are different. We analyze only substitution and deletion errors, insertion errors are not considered in this paper. Decision trees are pruned by the error-based pruning. We set the threshold to 10 based on our preliminary study.

Table 2 shows the attributes in consideration. They are either discrete or continuous. In the table, “D” or “C” indicates that the attribute is treated in C4.5 as discrete or continuous, respectively.

We use the JTAG3.03 morphological analysis program to obtain the part of speech information. For the judgment of filled pauses and repairs, annotated information in the CSJ transcription is used. The speaking rate and frame likelihood attributes are calculated by using the result of phoneme alignment on the reference sentence.

Table 2 Word attributes

Number of phonemes in the word	C
Word duration (number of frames)	C
Speaking rate (number of phonemes/number of frames)	C
Averaged acoustic frame likelihood	C
Ratio of a certain phoneme class such as vowel or nasal	C
Part of speech (noun, verb, etc.)	D
Filled pause or not	D
Repair or not	D
Quotation or not	D
Loanword or not	D
Word frequency in the training set	C
Bigram score	C
Trigram score	C
Back off class	D
Word order in the sentence from either beginning or end	C
Part of speech of the left/right context word	D
Left/Right context word is filled pause or not	D
Left/Right context word is repair or not	D
Left/Right context word is quotation or not	D
Left/Right context word is loanword or not	D

We use the first 2320 cases for each presentation in order to unify the condition in terms of the amount of data. Trees are created and tested using the cross validation method; the data set made of all selected cases is divided into 10 subsets and one of them is used for testing.

3.3 Decision trees for phonemes

Decision trees for phonemes are built in the same way using phonemes as units instead of words. Like for the word analysis, we consider only substitution and deletion errors, and neglect insertion errors. The pruning threshold is set to 10 based on our preliminary experiments.

Table 3 phoneme attributes

Kind of phoneme (a, u:, sh, etc.)	D
Left/Right phoneme kind context	D
Phoneme class (voiced, nasal, etc)	D
Left/Right phoneme class context	D
Filled pause or not	D
Repair or not	D
Left/Right context is filled pause or not	D
Left/Right context is repair or not	D
Max frame likelihood over all states	C
Minimum frame likelihood over all states	C
Average frame likelihood over all states	C
Number of states whose frame likelihood is greater than frame max minus delta	C
Frame likelihood variance over all states	C
Phoneme duration	C
Frame energy	C
Delta frame energy	C
Mono-phone frequency in the corpus	C
Tri-phone frequency in the corpus	C

Table 3 shows phoneme attributes used in the experiments. Frame-by-frame information such as likelihood and power is averaged over the period of each reference phoneme obtained by the phoneme alignment. The likelihood value for each HMM state does not include transitional probability. When counting the number of mono-phone and tri-phone occurrences, model sharing is not considered. Whether or not a phoneme is uttered in a filled pause or repair is determined according to the annotation of the CSJ.

Trees are created and tested using the cross validation method, dividing the data into 5 subsets. We use the first 8600 cases per presentation to equalize the amount of data.

4. RECOGNITION PERFORMANCE OF CSJ PRESENTATION UTTERANCES

Figure 1 presents test-set perplexity and out-of-vocabulary (OOV) rate of the task using the trigram language model. Figure 2 shows word and phoneme recognition accuracies. In the phoneme recognition, no linguistic constraint was used. The results show that the accuracies largely vary from speaker to speaker.

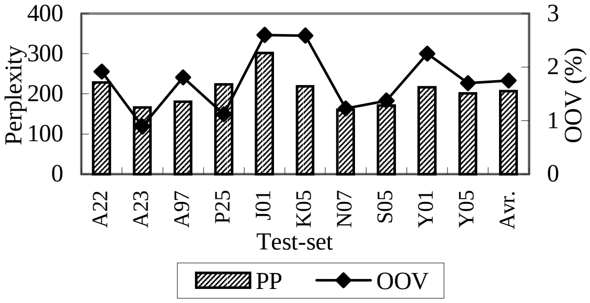

Fig. 1 Test-set perplexity and OOV rate of the task.

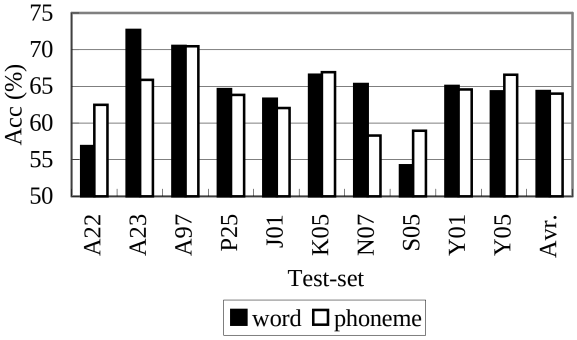

Fig. 2 Word/phoneme recognition accuracy.

5. ERROR ANALYSIS USING DECISION TREES

5.1 Decision trees for words

A set of decision trees for words was made using all the attributes listed in Table 2. Figure 3 shows prediction correctness of the trees. For comparison, word (recognition) correctness (WCorr) is also shown in the figure. TSpk denotes prediction correctness when trees are built for each speaker. TAll is also prediction correctness but when trees are built using the training data by all the 10 speakers.

The word correctness corresponds to the prediction correctness of a tree having only the root node. As can be seen, prediction correctness is higher than word correctness. This difference is believed to result from recognition errors caused by the attributes found in the tree.

Questions assigned near the root of the trees are the repair, the word occurrence frequency, the ratio of voiced phonemes, the ratio of long (double) consonants, etc.

TAll indicates better prediction correctness than TSpk. This means that the amount of data is more significant than speaker-specific variations in this analysis. That is, the difference of the sources of recognition errors among speakers is not significant in these data.

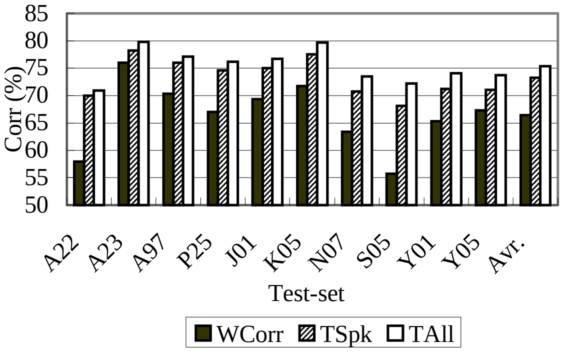

Fig. 3 Recognition and prediction correctness.

5.2 Error factor analysis of word recognition

To analyze what attributes have strong correlation with recognition errors, we selected various subsets of attributes and measured the performance of trees. As a result, it turned out that only three attributes produced almost the same performance as all the attributes in Table 2. The three attributes are the number of phonemes in a word, the speaking rate, and the frequency of word occurrences. Word recognition error tend to be higher if the word has relatively small number of phonemes, spoken fast, and observed less frequently in the language-model training corpus. But strictly speaking, the relationships are not monotonic. For example, very slow speaking rate also tends to increase errors. The other attributes are either less informative about word error or the information they provide is already included in the one given by the three major attributes.

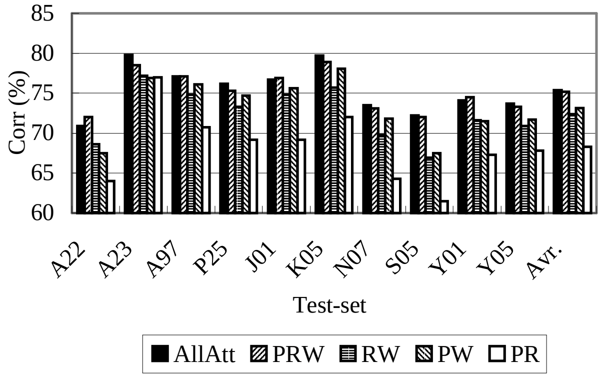

Fig. 4 Analysis of word attributes.

Figure 4 shows the prediction correctness of the trees for subsets of attributes. The trees were built using training data by 10 speakers. AllAtt indicates the correctness of trees using all attributes. P, R and W indicate the number of phonemes in the word, the speaking rate and the word frequency, respectively. It can be seen that omitting any one of them degrades the prediction correctness.

In order to analyze the sources of word recognition accuracy variation among speakers, we estimated the success rate of recognition using the decision tree with the three most significant attributes. We defined predicted success rate (*PSR*) for each utterance as follows.

$$PSR = \frac{T}{T+F} \quad , \tag{2}$$

where T indicates the number of test cases in the utterance that are predicted to be true (correctly recognized) by the tree, and F indicates that predicted to be false (incorrectly recognized). Figure 5 shows the relationship between *PSR* and the actual recognition correctness for the 10 speakers in the test set.

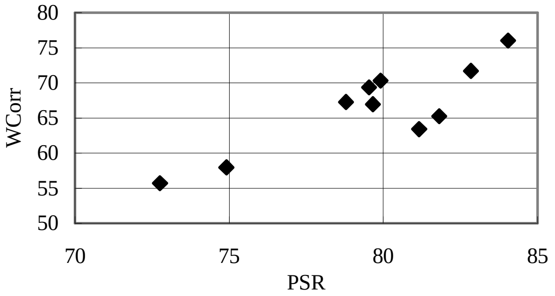

Fig. 5 The predicted success rate (PSR) and the recognition correctness.

The correlation coefficient for this result is 0.87, meaning that differences in the three attributes are highly related to the variation in the recognition accuracy.

5.3 Decision trees for phonemes

Figure 6 shows the prediction correctness of the trees for phonemes that are made using all attributes in Table 3. TSpk denotes the tree made for each speaker. TAll denotes the tree made by using all the training data from the 10 presentations. For comparison, results of phoneme correctness (PCorr) are also shown.

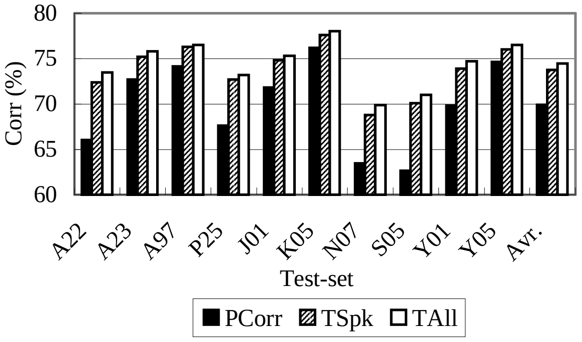

Fig. 6 Prediction correctness.

The prediction correctness of TAll is higher than that of TSpk. This suggests that the factors of recognition errors are similar among speakers.

5.4 Error factor analysis of phoneme recognition

We selected various subsets of attributes and compared the performances of the trees. We found that a subset of attributes that indicates almost the same prediction correctness as all attributes in the Table 3 consisted of the frame-max and frame variance (F), the phoneme class and phoneme class context (P), the phoneme duration (D), and the mono-phone frequency in the training data (M). Figure 7 shows the prediction correctness for several attribute sets. Among these attributes, the phoneme duration seems to contribute the most to the correct recognition of phonemes.

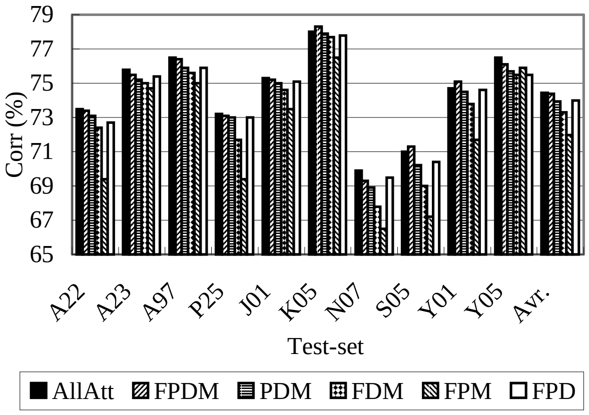

Fig. 7 Analysis of phoneme attributes.

6. CONCLUSION

In this paper, we have proposed the use of a decision tree for analyzing recognition errors. We have quantitatively analyzed to what extent the recognition error can be explained by a set of attributes. In word recognition, we have found that the number of phonemes in the word, the speaking rate and the word frequency in the training data are highly related to the recognition rate. In phoneme recognition, a set of attributes consisting of the frame-max, the frame variance, the phoneme class, the phoneme class context, the phoneme duration and the mono-phone occurrence count have been found to have the same prediction power as all the attributes used in the experiment. To increase the recognition accuracy, the following issues are important; designing words considering the number of included phonemes, modeling the effects of speaking rate, and properly increasing the training data. It might also be useful to use the decision-tree-based framework for estimating the confidence measure for recognition.

REFERENCES

[1] S. Furui, et al., “Toward the realization of spontaneous speech recognition”, Proc. ICSLP, China, Vol. 3, pp. 518-521, 2000.

[2] T. Shinozaki, C. Hori, and S. Furui, “Towards Automatic Transcription of spontaneous presentations”, Proc. EUROSPEECH, Denmark, Vol. 1, pp. 491-494, 2001.

[3] A. Lee, et al., “An Efficient two-pass search algorithm using word trellis index”, Proc. ICSLP, Australia, pp.1831-1834, 1998.

[4] J.R.Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufmann, 1996.