MESSAGE-DRIVEN SPEECH RECOGNITION AND TOPIC-WORD EXTRACTION

K. Ohtsuki[2], S. Furui[1], A. Iwasaki[1], N. Sakurai[1]

[1]Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
[2]NTT Human Interface Laboratories
1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0837 Japan

ABSTRACT

This paper proposes a new formulation for speech recognition/understanding systems, in which the a posteriori probability of a speaker's message that the speaker intend to address given an observed acoustic sequence is maximized. This is an extension of the current criterion that maximizes a probability of a word sequence. Among the various possible representations, we employ co-occurrence score of words measured by mutual information as the conditional probability of a word sequence occurring in a given message. The word sequence hypotheses obtained by bigram and trigram language models are rescored using the co-occurrence score. Experimental results show that the word accuracy is improved by this method. Topic-words, which represent the content of a speech signal are then extracted from speech recognition results based on the significance score of each word. When five topic-words are extracted for each broadcast-news article, 82.8% of them are correct in average. This paper also proposes a verbalization-dependent language model, which is useful for Japanese dictation systems.

1. INTRODUCTION

State-of-the-art automatic speech recognition systems employ the criterion of maximizing $P(W|X)$, where W is a word sequence and X is an acoustic observation sequence. This criterion is reasonable for dictating read speech. However, the ultimate goal of automatic speech recognition is to extract the underlying messages of the speaker from the speech signals. Hence we need to model the process of speech generation and recognition as shown in Fig. 1 [1], where M is the message (content) that a speaker intended to convey.

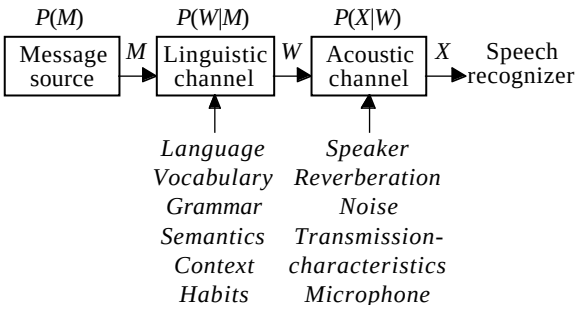

Figure 1: A communication - theoretic view of speech generation and recognition.

According to this model, the speech recognition process is the problem of estimating M to maximize $P(M|X)$. This paper proposes a new formulation to solve this problem. It covers the various approaches that have been attempted and also suggests new approaches. We consider that message M is represented by a co-occurrence of words, and based on this consideration we propose a new formulation for speech recognition.

We apply this formulation to broadcast-news speech transcriptions and show how it improves word accuracy. We then explicitly represent message M as a combination of topic-words, and automatically extract topic-words from the word sequence obtained as the result of the speech recognition process. This paper also describes language models taking care of multiple verbalizations of words, which is one of the difficult problems for Japanese speech dictation systems.

2. MESSAGE-DRIVEN SPEECH RECOGNITION

According to Fig. 1, the speech recognition process is represented as the maximization of the following a posteriori probability,

$$\arg\max_M P(M|X) = \arg\max_M \sum_W P(M|W)P(W|X). \quad (1)$$

Using Bayes' Rule, Eq. (1) can be expressed as

$$\arg\max_M P(M|X) = \arg\max_M \sum_W P(X|W)P(W|M)P(M). \quad (2)$$

For simplicity, we can approximate the equation as

$$\arg\max_M P(M|X) \approx \arg\max_{M,W} P(X|W)P(W|M)P(M). \quad (3)$$

$P(X|W)$ is calculated using hidden Markov models in the same way as in usual recognition processes. We consider $P(M)$ has a uniform probability for all M. Then we only need to consider further the term $P(W|M)$. We assume that $P(W|M)$ can be expressed as follows.

$$P(W|M) \approx (1-\lambda)P(W) + \lambda P(W|M), \quad (4)$$

where λ, $0 \le \lambda \le 1$, is a weighting factor. $P(W)$, the first term of the right hand side, represents a part of $P(W|M)$ that is independent of M and can be given by a general statistical language model. $P(W|M)$, the second term of the right hand side, represents the part of $P(W|M)$ that depends on M. The latter term can be represented in various ways, whether the dependency of M is represented explicitly or implicitly. The explicit formulation usually needs to represent M by a finite number of topic classes. Approaches that change language models according to the estimated topic class M (e. g. [2])

correspond to this formulation. Approaches using probabilistic state transition networks [3] or HMM [4] for forming semantic language models are also classified into this category. A cache model [5] is one of the approaches in which M is implicitly represented.

In this paper, we consider that M is represented by a co-occurrence of words based on the distributional hypothesis by Harris [6]. Similar methods include a method that uses a thesaurus for measuring semantic similarity between words and the one that clusters words based on some similarity measures. Since these approaches formulate $P(W|M)$ without explicitly representing M, they can use information about the speaker's message M without being affected by the quantization problem of topic classes.

3. BASELINE JAPANESE BROADCAST-NEWS DICTATION SYSTEM

3.1 Acoustic Models

The feature vector consists of 16 cepstral coefficients, normalized logarithmic power, and their delta features (derivatives). The total number of parameters in each vector is 34. Cepstral coefficients were normalized by the cepstral mean subtraction (CMS) method.

The acoustic models we used were shared-state triphone HMMs designed using tree-based clustering. The total number of states was 2,106, and the number of Gaussian mixture components per state was 4. They were trained using phonetically-balanced sentences and dialogues read by 53 speakers. They were completely different from the broadcast-news task. All of the speakers were male. The total number of training utterances was 13,270 and the total length of the training data was approximately 20 hours.

3.2 Language Models

Broadcast-news manuscripts prepared during the period from August 1992 to May 1996, comprising roughly 500k sentences and 22M words, were used for constructing language models. To calculate word n-gram language models, we segmented the broadcast-news manuscripts into words by using a morphological analyzer since Japanese sentences are written without spaces between words. A word-frequency list was derived for the news manuscripts, and the 20k most frequently used words were selected as vocabulary words. This 20k vocabulary covers about 98% of the words in the broadcast-news manuscripts. We calculated bigrams, trigrams, and 4-grams, and estimated unseen models using Katz's back-off smoothing method.

To prevent data sparseness in training 4-grams, we employed class 4-grams, i.e.,

$$P\left(w_k \middle| w_{k-3}, w_{k-2}, w_{k-1}\right) \approx P\left(w_k \middle| C_k\right) P\left(C_k \middle| C_{k-3}, C_{k-2}, C_{k-1}\right), \quad (5)$$

where C_k is a word class. We first classified the vocabulary words into 138 classes on the basis of their grammatical characteristics given by the morphological analyzer. The classes of verbs, adjectives and nouns, consisting of many vocabulary words, were divided into smaller classes using the thesaurus, and the total number of classes was 1,000. A preliminary experiment showed that the thesaurus-based classes achieved lower perplexity than the classes based on mutual information under a condition of 1,000 classes. The 3,500 most frequent words in the training text were excluded from the classification and were handled as independent classes.

In written Japanese, a particular word may have a number of different verbalizations associated with it [7]. This is also true to much less frequent extent in isolated examples in English, (e. g. “minute”, “read”). With this feature of Japanese in mind, we considered three methods of building language models from written Japanese.

In the first and second methods, the language model is constructed without regard to the different verbalizations. Essentially, this represents a language model of the written form of words. During recognition, for a given word, a phoneme sequence for each possible verbalization is given by a dictionary. In the first method, all verbalizations are given equal probability. In the second method, a sequence of words in the training text is converted into the actual verbalization, which it represents using a morphological analysis program. This conversion is difficult and error prone. The probability of each verbalization for each word averaged over every kind of context was calculated and used during recognition. As we previously reported [7], the second method gives better recognition performance than the first method, but its disadvantage is that it does not consider the dependency of verbalization on the context.

In the third method, the training texts are also converted to the actual verbalization and the words with the same written form but with different verbalizations are regarded as different words. A language model is constructed and used in a manner similar to the way it would be used in English.

3.3 Evaluation Experiments

News speech data, which were broadcast on TV in July 1996, were divided into two parts, a clean set and a noisy set, and these were separately evaluated. The clean set consisted of 50 utterances with no background noise, and the noisy set consisted of 50 utterances with background noise. The noisy set included spontaneous speech such as reports by correspondents. Each set included the speech of six or seven speakers. All utterances were manually segmented into sentences. The out-of-vocabulary (OOV) rates were 0.8% and 2.9% for the clean and noisy sets, respectively.

Table 1 shows the experimental results and test-set perplexities for the first and the third method of language model construction. The results in the parentheses correspond to the results for the third method. Although there was almost no change in test-set perplexities, the word accuracies of the clean set and the noisy set for the third method were 2.0% and 3.6% higher respectively, than those for the first method when trigrams were used. This result is slightly better than the one obtained by the second method reported in [7]. Even though the third method increased the vocabulary size, the coverage of the most frequent 20k words decreased only by 0.1% from 97.7% to 97.6%.

The table also shows that there is a large difference between the results of the clean set and the noisy set. This was mainly due to the fact that the latter had a much larger test-set perplexities and was contaminated by background noise.

The N-best hypotheses obtained using bigram language models were rescored by using (word-) trigrams and (class-) 4-grams. A combination of trigram and 4-gram models achieved a 0.5% absolute value improvement in word accuracy over the case of only using one or the other.

Table 1: Recognition results of the baseline system using bigrams and trigrams; (): verbalization-dependent language model.

Evaluation set	Bigram		Trigram	
	Test-set perplexity	Word accuracy [%]	Test-set perplexity	Word accuracy [%]
Clean	124.4 (124.2)	77.9 (79.2)	64.4 (64.9)	80.7 (82.7)
Noisy	187.9 (187.3)	58.0 (60.3)	113.5 (114.5)	60.3 (63.9)

4. MESSAGE-DRIVEN SPEECH RECOGNITION EXPERIMENTS

4.1 Word Co-Occurrence Score

As mentioned in Section 2, $P(W|M)$, the second term on the right hand side of Eq. (4), is represented by word co-occurrences in this paper. Since most of the words that express messages or topics are nouns, we extracted only nouns from the N-best hypotheses of the word sequences which were obtained using the trigram language model. Message-driven speech recognition results were obtained by rescoring the hypotheses by adding word co-occurrence scores for every pair of nouns. The co-occurrence score was calculated based on the mutual information as follows;

$$CoScore(w_i, w_j) = \log \frac{p(w_i, w_j)}{\left(p(w_i)p(w_j)\right)^{1/2}}, \qquad (6)$$

where $p(w_i, w_j)$ is the probability of observing words w_i and w_j in the same news article, and $p(w_i)$ and $p(w_j)$ are the probabilities of observing word w_i and w_j in all the articles, respectively. In order to compensate the probabilities of the words with very low frequency, a square root term was employed in the denominator of the equation. The co-occurrence scores were calculated using the same database as that was used for language modeling.

4.2 Experimental Results

Table 2 shows the word accuracies obtained by rescoring with word co-occurrence scores. The results before rescoring are also shown in Table 2 for comparison. The weighting factor for co-occurrence score, λ, was appropriately set on the basis of preliminary experiments. As shown in Table 2, word accuracy for the clean set was improved by incorporating $P(W|M)$.

Table 2: Comparison of word accuracies [%] with or without $P(W|M)$.

Evaluation set	Language model	
	$P(W)$	$(1-\lambda)P(W)+\lambda P(W|M)$
Clean	82.7	83.5
Noisy	63.9	63.8

5. TOPIC-WORD EXTRACTION

5.1 Word Significance Measure

As we mentioned in Section 4, most of the topic-words that expressed messages or topics of broadcast-news were nouns. We therefore investigated a method for extracting topic-words from nouns in the speech recognition results on the basis of a significance measure for each word.

Many of the measures that have been used in information retrieval from text databases were tried in a preliminary experiment, and the following measure [8] was chosen.

$$SgScore(w_i) = g_i \cdot \log \frac{G_A}{G_i}, (i = 1, 2, ..., N), \qquad (7)$$

where N is the vocabulary size (nouns only), g_i is the frequency of word w_i in a news article, G_i is the frequency of word w_i in all news articles, and G_A is the summation of all G_i's:

$$G_A = \sum_i G_i \qquad (8)$$

Equation (7) gives the amount of information born by the word w_i in the particular news article. The Nikkei newspaper articles over a five year period were used for calculating G_i and G_A values.

5.2 Evaluation Experiments

Performance of topic-word extraction was evaluated using the following scores for each broadcast-news articles.

$$Recall = \frac{C}{T} \cdot 100 \qquad (9)$$

$$Precision = \frac{C}{H} \cdot 100, \qquad (10)$$

where C is the number of correctly extracted topic-words, T is the total number of correct topic-words, and H is the total number of extracted topic-words.

Twenty-nine broadcast-news articles comprising 142 utterances by 15 male speakers (8 anchors and 7 others) were used for evaluation. Each news article had 2 - 14 utterances (5 utterances on the average). True topic-words were given by three subjects; 4 - 10 phrases were given for each news article by each subject. A true topic-word set was constructed for each article from topic-words given by at least one subject (35.7 words on average). Supplementary experiments were also conducted by giving correct texts instead of transcription results as input. Eighty-nine percent of the true topic-word set were nouns.

Figure 2 shows the results averaged over the 29 news articles. It is observed that, if transcription results are used

as input, precision as well as recall is reduced by roughly 10% in comparison with that obtained by using the correct texts as input. The values of precision when choosing 5, 10 or 15 topic-words for each news article are shown in Table 3. When five topic-words were chosen from speech recognition results, 82.8% of them were correct on average.

In light of the fact that, on average 74% of the topic-words given by two subjects overlap, the topic-word extraction performance obtained for speech recognition results seems good enough for practical use. A supplementary experiment with G_i and G_A values calculated using five years worth (from July 1992 to May 1996) of broadcast-news manuscripts instead of newspaper articles achieved almost the same performance.

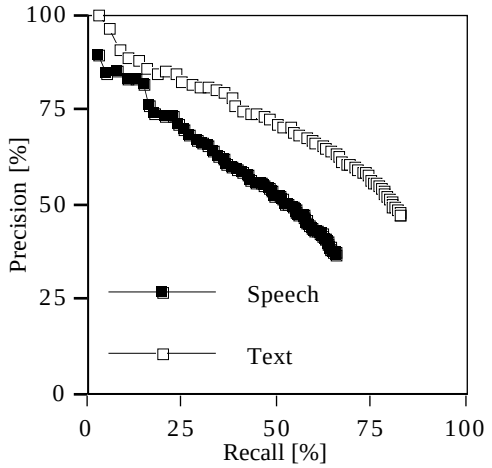

Figure 2: Topic-word extraction from broadcast-news speech or news text.

Table 3: The precision [%] obtained when 5, 10 or 15 topic-words were extracted.

Input data	Number of extracted topic-words		
	5	10	15
Speech	82.8	73.4	66.2
Text	88.3	82.1	79.5

6. CONCLUSION

This paper proposed a new formulation of speech recognition based on maximizing the a posteriori probability of message M that the speaker intended to convey given an acoustic observation sequence X, $P(M|X)$. This formulation is an extension of the current speech recognition criterion which maximizes $P(W|X)$, where W is a word sequence. $P(W|M)$ can be represented in various ways; for example, it can represented as a topic-dependent language model or a cache model. In this paper, M was represented by word co-occurrences so that we were able to handle M as a continuous quantity without explicitly representing topic classes.

We constructed a baseline speech recognition system for Japanese broadcast-news dictation. As one of the Japanese-specific problems, we investigated the issue of multiple verbalizations associated with each word, and showed that word accuracy could be improved by a new language model in which words with the same written form but with different verbalization were treated as different words. We carried out speech recognition experiments based on the above-mentioned framework by introducing mutual information-based word co-occurrence scores; better word accuracy was achieved than was possible with the current framework.

We then extracted topic-words representing topics of a given news article from broadcast-news speech recognition results. The topic-words were extracted from a set of nouns in the hypothesized word sequence based on the word significance measure. When five topic-words were chosen from each hypothesis, 82.8% of them were correct on average.

The proposed new formulation of speech recognition framework suggests extensions in various ways to new formulations of speech understanding systems. The system should improve with the elaboration of a more sophisticated $P(W|M)$, a better morphological analyzer which gives more precise verbalizations for words, and a more appropriate measure for extracting topic-words.

ACKNOWLEDGMENTS

A part of this research was conducted during S. Furui's stay at Bell Laboratories in US. The authors wish to express their appreciation of Dr. B.-H. Juang at Bell Labs for several fruitful discussions. The authors also would like to thank Mr. T. Matsuoka and Mr. T. Hori for their contribution in constructing the baseline speech recognition system. The authors would like to thank NHK (Japan Broadcasting Corporation) for providing us with the broadcast-news database. The authors are also grateful to Nihon Keizai Shinbun Incorporated for allowing us to use the newspaper text database (Nikkei CD-ROM 90-94) in our research.

REFERENCES

[1] B.-H. Juang, "Automatic speech recognition: Problems, progress & prospects," IEEE Workshop on Neural Networks for Signal Processing, 1996.

[2] S. F. Chen, et al., "Topic adaptation for language modeling using unnormalized exponential models," Proc. ICASSP'98, pp. II-681-684, 1998.

[3] S. Miller, et al., "Statistical language processing using hidden understanding models," Proc. DARPA Human Language Technology Workshop, pp. 278-282, 1994.

[4] R. Pieraccini, et al., "A speech understanding system based on statistical representation of semantics," Proc. ICASSP'92, pp. I-193-196, 1992.

[5] R. Kuhn and R. De Mori, "A cache-based natural language model for speech recognition," IEEE Trans. PAMI-12, 6, pp. 570-583, 1990.

[6] Z. S. Harris, "Co-occurrence and transformation in linguistic structure," Language, 33, pp. 283-340, 1957.

[7] S. Furui, et al., "Japanese broadcast news transcription and topic detection," Proc. Broadcast News Transcription and Understanding Workshop, pp. 144-149, 1998.

[8] T. Noreault, et al., "A performance evaluation of similarity measure; Document term weighting schemes and representations in a Boolean environment," in R. N. Oddy ed. Information Retrieval Research, London, Butterworths, pp. 57-76, 1997.