J. Acoust. Soc. Jpn. (*E*) **21**, 1 (2000)

ACOUSTICAL LETTER

Decoder for Japanese broadcast news transcription

Toru Imai, Kazuo Onoe, Akio Kobayashi, and Akio Ando
NHK Science and Technical Research Laboratories,
1-10-11, Kinuta, Setagaya, Tokyo, 157-8510 Japan
(*Received 30 April 1999*)

Keywords: Speech recognition, News, Decoder, Search
PACS number: 43. 72. Ne

1. Introduction

In order to generate closed-captions for TV news automatically, a Japanese broadcast news transcription system is being developed.[1] This paper describes a decoder of the transcription system and reports word accuracies and running times with various vocabulary sizes and phoneme networks of linear or tree structures.[2]

2. Decoder

The decoder consists of two passes as illustrated in Fig. 1. In the first pass, the word-dependent N-best algorithm[3] is carried out with a bigram language model, triphone HMMs, and the Viterbi beam search. During the search top-n ($n \ll N$) paths which have different previous words are saved in each HMM state. If some paths reach at the same state with the same previous word, only the best path is saved. In a final state of a word-end phoneme, the n paths and pointers to the end of their previous word-ends are saved as a word lattice, and only the best path propagates to following words. At a sentence-end the word lattice is recursively traced back to get N-best sentences.

Pruning in the beam search is based on a global beam width of log-likelihood and a narrower word-end beam width. In order to reduce computation effort, words which can follow a previous one by a back-off process are restricted to unigram's top-K words.

The first pass consists of an HMM-step which advances all the active paths in the HMM states associated with nodes of the triphones by one frame of input speech and a grammar-step which advances all the active paths in the final HMM states to following nodes (word internal and external). Two types of the phoneme network are applied in this paper: a linear structure where each node belongs to one word or a tree structure[4] where initial phoneme sequences are shared by some words.

In the second pass of the decoder, the N-best sentences from the first pass are rescored by a trigram language model to find the best sentence as a recognized transcription.

3. Phoneme network

3.1 Linear structure

Since the linear-structured phoneme network gives word identity at the first phonemes, bigrams can be applied between words immediately. However, it increases active nodes at the word beginnings. To avoid it and save computation, it is effective to dynamically control a threshold of the number of active nodes.[5]

3.2 Tree structure

Sharing initial nodes by some words, the tree-structured phoneme network can reduce the number of active nodes in the word beginnings. However, bigrams can not be applied until leaf nodes where word identity is known. To apply a language model earlier, the maximum bigram into the words sharing a node is used for pruning.[6] It requires much computation to get the previous-word-dependent maximum bigram in each node if a list of sharing words is changed. In a small vocabulary, the maximum bigrams can be computed and saved beforehand, but it requires much memories in a larger vocabulary. We propose to compute and save the values into a table beforehand, according to the size of the vocabulary, in all nodes at L and lower levels, and nodes related to unigram's top K words at level $L+1$.

The phoneme network discussed here is a static single tree, though a dynamic one copying the tree after each word is also proposed.[4,6] In the static case, n should be greater even if $N=1$ in order to reduce search errors.

4. Experiment

An experiment to transcribe NHK's Japanese broadcast news was performed in different conditions of vocabulary sizes (5 K, 10 K, 20 K, 40 K, and 65 K) and network structures. Table 1 shows the experimental condition. The number of nodes associated with triphones in the network was 160 K in the linear structure and 85 K in the tree structure with the 20 K word lexicon. The table of the maximum bigrams in the tree structure was computed beforehand so as to spend 300 MB memories, which set L to 12, 3, 2, 1.2, or 1.1 respectively for the vocabulary sizes (the non-integer means to include a part of the next level).

The result is shown in Table 2. Independent to the network structures, the word accuracy increased as the vocabulary size increasing to 20K and did not change much with 40 K or 65 K words. The tree structure gave a better word accuracy than the linear structure independent to the vocabulary size. It is because the tree network uses a greater bigram than real one for

pruning in the beginning nodes and has high possibility to save words pruned in the linear network.

The effect of the second pass to rescore sentences by trigrams was 3.1% with the 20 K word lexicon and the tree network. To limit words which can follow a previous word by a back-off bigram reduced computation effort to 92% with 0.5% higher errors (in the linear network it was reduced to 85% with 0.2% higher errors). To get the table of the maximum bigrams beforehand for pruning made the running time 24 times faster without changing the accuracy.

The tree network was superior to the linear network also in the point of the running time with a 20 K or smaller word lexicon. However, a larger lexicon than 20 K made it slower because the maximum bigrams could not be computed enough beforehand in the initial nodes. Load distribution in decoding is shown in Fig. 2.

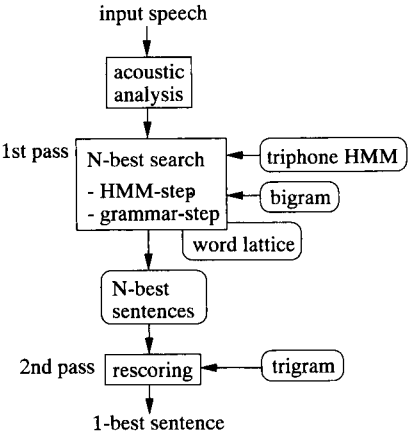

Fig. 1 Decoder.

1st pass	98%
acoustic analysis	2%
HMM-step	56%
computation of Gaussians	38%
grammar-step	39%
maximum bigram (except table)	7%
2nd pass	2%

Fig. 2 Load distribution (tree network, 20 K words).

5. Conclusion

This paper described the decoder for Japanese broadcast news transcription and its performance of word

Table 1 Experimental condition.

Acoustic analysis
16 kHz sampling Hamming window, 25 ms width, 10 ms period 39 features (12 MFCC+log-power, 1st and 2nd derivatives)
Acoustic model
gender-dependent speaker-independent state-clustered 8-mixtured triphone HMM #logical HMM=5,844 (news 20 K words) #physical HMM=1,366 (m), 1,396 (f) #tied state=2,356 (m), 2,420 (f) / 42 phonemes training data: ASJ, ATR continuous speech database, 54 males (12H), 58 females (15H)
Language model
bigram (1st pass), trigram (2nd pass) back-off smoothing, Good-Turing, cut-off 1-2 training data: NHK's Japanese news database (aired on Apr. 1, '91-Jul. 10, '96)
Decoder
#saved path in an HMM state: $n=4$ #1st pass's output sentences: $N=200$ global beam width=160 word-end beam width=50 LM score weight=14, insertion penalty=0 #possible connected word in back-off: $K=2{,}000$ #maximum active node (in linear net)=20,000
Evaluation data
NHK news[7](aired on Jul. 11, '96-Jul. 14, '96) studio anchors, clean speech 50 sentences for each gender (known) 100 sentences total (3,986 words) average length of 13.2 s
Machine
Alpha-21164 chip, 600 MHz, 768 MB memory

Table 2 Recognition result.

Vocabulary size	Test-set perplexity		OOV	Linear network		Tree network	
	bigram	trigram		accuracy	×real time	accuracy	×real time
5K	67.4	35.2	5.2%	73.9%	2.1	77.4%	1.9
10K	72.3	36.3	2.0%	78.3%	2.6	82.2%	2.3
20K	76.1	37.5	0.5%	79.7%	3.1	83.5%	2.8
40K	77.5	38.3	0.3%	79.9%	3.7	83.3%	5.9
65K	77.8	38.4	0.2%	79.9%	4.4	83.5%	9.5

accuracies and running times with various vocabulary sizes and network structures. The experiment showed that the tree network obtained higher word accuracies and faster running times than the linear one with a 20 K or smaller word lexicon, and that the vocabulary size of 20 K was sufficient for the broadcast news task.

References

1) A. Kobayashi, T. Imai, A. Ando, E. Miyasaka, H. Akamatsu, S. Nakagawa, R. Oguro, K. Ozeki, S. Furui, J. Suzuki, and K. Shirai, "A study on continuous speech recognition system for broadcast news," Proc. Autumn Meet. Acoust. Soc. Jpn. 3-1-9 (in Japanese) (1997).
2) T. Imai, K. Onoe, A. Kobayashi, and A. Ando, "A decoder for broadcast news transcription," Proc. Autumn Meet. Acoust. Soc. Jpn. 3-1-12 (in Japanese) (1998).
3) R. Schwartz and S. Austin, "A comparison of several approximate algorithms for finding multiple (N-best) sentence hypotheses," Proc. ICASSP 91, 701-704 (1991).
4) H. Ney, R. Haeb-Umbach, B.-H. Tran, and M. Oerder, "Improvements in beam search for 10000-word continuous speech recognition," Proc. ICASSP 92, 9-12 (1992).
5) H. Hattori, T. Watanabe, K. Hatazaki, K. Yoshida, T. Emori, and S. Koga, "A large vocabulary speech recognition system using a beam search technique," Proc. Spring Meet. Acoust. Soc. Jpn. 2-6-5 (in Japanese) (1997).
6) J. J. Odell, V. Valtchev, P. C. Woodland, and S. J. Young, "A one pass decoder design for large vocabulary recognition," Proc. Human Language Technology Workshop, 405-410 (1994).
7) A. Ando and E. Miyasaka, "Construction of Japanese News Speech Databases," Proc. Spring Meet. Acoust. Soc. Jpn. 2-Q-9 (in Japanese) (1997).