A HYBRID WORDSPOTTING METHOD FOR SPONTANEOUS SPEECH UNDERSTANDING USING WORD-BASED PATTERN MATCHING AND PHONEME-BASED HMM

Hiroshi Kanazawa†, Mitsuyoshi Tachimori† and Yoichi Takebayashi††

†Toshiba Corporation, Kansai Research Laboratory.
††Toshiba Corporation, Research and Development Center.
8-6-26 Motoyama-minami-cho, Higashinada-ku, Kobe 658, JAPAN

ABSTRACT

We have proposed a new wordspotting method, combining word-based pattern matching and phoneme-based Hidden Markov Model(HMM). Word-based pattern matching based on the time-frequency representation of a whole word pattern is robust against pattern variations and background noise, while the phoneme-based HMM, which represents phonemic features within a word pattern, is flexible for expanding the vocabulary. Because of the difference in scope, these two have their own characteristics in terms of robustness and accuracy. To take advantage of the features of these two, we have integrated these different types of wordspotting results under a unified criterion. A syntactic and semantic parser is also utilized to prune the wordspotting results for spontaneous speech understanding. Experimental results indicate the effectiveness of the proposed method.

1. INTRODUCTION

As wordspotting is known to be effective for understanding ill-formed spontaneous speech, much research is being done on this topic. Among many wordspotting methods proposed, HMM, which is a promising model in recognizing large vocabulary continuous speech, has given relatively good results using the garbage model [1,2]. In particular, phoneme HMM is widely used because it is suitable for representing speech events and can easily expand the vocabulary, but it is not robust enough to handle pattern variations in utterance style and background noise. With this in mind, we previously proposed Noise Immunity wordspotting method [3], and built a prototype of speaker-independent real-time speech dialogue system TOSBURG II [4]. The Noise Immunity method employs Multiple Similarity using whole word pattern matching [3,5], which is robust against pattern variations and background noise, but makes it difficult to expand the vocabulary.

Several hybrid methods were proposed to increase recognition accuracy. Most of them adopt a multistage framework because different measures and parameters cannot be integrated easily. To name a few, HMM is integrated with a neural network or Learning vector quantization[6~9]. To take this into account, we have proposed a new hybrid wordspotting method in which word-based pattern matching and phoneme-based HMM are integrated. The key idea here is to carry out wordspotting in parallel by different recognition unit, to integrate these results under a unified criterion and to utilize a parser for understanding spontaneous speech. This paper first describes a comparison between word-based pattern matching and phoneme-based HMM. Next, our approach and an integrating algorithm are introduced. Finally, experimental evaluations are presented.

2. A HYBRID WORDSPOTTING METHOD

2.1 Word-based pattern matching vs. phoneme-based HMM

With a time-frequency spectrum as in Fig.1(a), a whole word can be globally described, while a phonemic pattern, consisting of single frame frequency spectrums, insufficiently represents dynamic features of a whole word, as shown in Fig.1(b); therefore, the former is more robust in terms of pattern distortion due to noise or utterance style [5]. However, generating whole word reference vectors requires a large amount of word speech training data, which makes it difficult to expand the vocabulary. Because of different recognition units of word pattern, these two approaches give different wordspotting results. Therefore, wordspotting perfor-

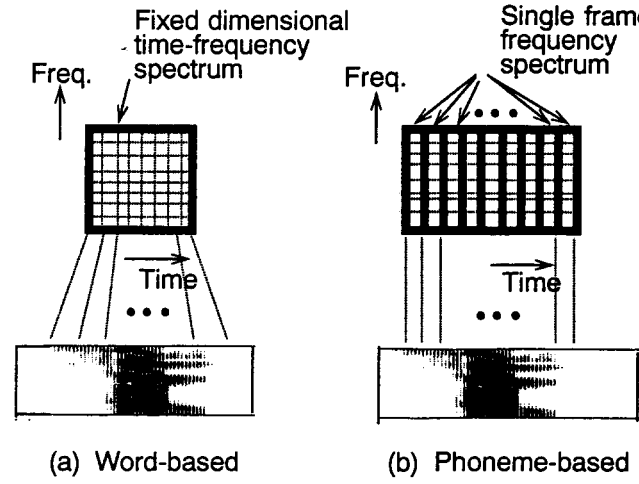

Fig. 1 Representations of a word pattern

0-7803-2431-5/95 $4.00 © 1995 IEEE

mance is expected to improve by integrating the two types of wordspotting results.

2.2 Approach

We propose a wordspotting method that effectively makes use of the features of the two different methods, namely, word-based wordspotting and phoneme-based wordspotting. Fig. 2 shows a block diagram of our method. To make the system robust against pattern variations and at the same time flexible for expanding the vocabulary, we have integrated the results of the two methods under the following strategies, so as to improve wordspotting performance:

(1) When calculating an HMM likelihood appropriate for word-based wordspotting results, two different types of results are compared under a unified criterion.
(2) For spontaneous speech understanding, the results of wordspotting are pruned through semantic utterance representations extracted by a syntactic and semantic parser.
(3) Word-based wordspotting results with robust recognition performance are given priority by weighting values.
(4) Word HMMs are constructed automatically using phoneme HMMs and a string of words to carry out phoneme-based wordspotting.

2.3 A hybrid algorithm

In word-based wordspotting, fixed-dimensional word pattern vectors are first extracted time-continuously by uniformly re-sampling the time series spectrum between the assumed start and end points. Then, the Multiple Similarity values are time-continuously computed from the pattern matching of these vectors with word reference vectors. This process is repeated for all word classes to extract spotted words after checking the thresholds [4].

In phoneme-based wordspotting, the Multiple Similarity values are calculated frame by frame using phoneme reference vectors. In this process, the candidates of a starting frame and the corresponding likelihood values are computed by using the continuous Viterbi decoding algorithm for each input frame [10]. The likelihood values are obtained by matching the time series of Multiple Similarities and word HMMs, designed by concatenating phoneme HMMs. Keywords are extracted by comparing the likelihood values with the thresholds.

The integrating method is explained below. First, word-based wordspotting and phoneme-based wordspotting are carried out at the same time. Both wordspotting results are inputted into a parser to extract the semantic utterance representations individually. Our parser analyzes only discrete keywords extracted from spontaneous speech and outputs semantic utterance representations [11]. The parser utilizes an LR parsing table obtained from augmented context-free grammar consisting of a set of production rules and semantic processing functions. The word candidates (W1,W2) of the N-best semantic utterance representations are inputted into the integrating section. Here, W1 is those obtained by word-based wordspotting and W2, by phoneme-based wordspotting. By selecting the N-best representations, the system can reduce the number of false alarms. Next,

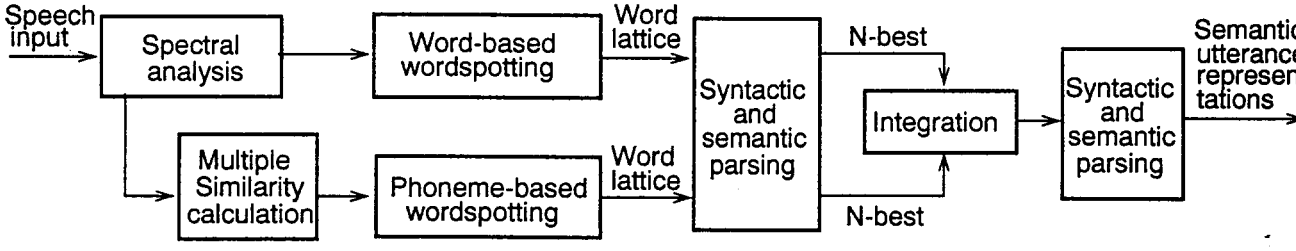

Fig. 2 A block diagram of the hybrid wordspotting method

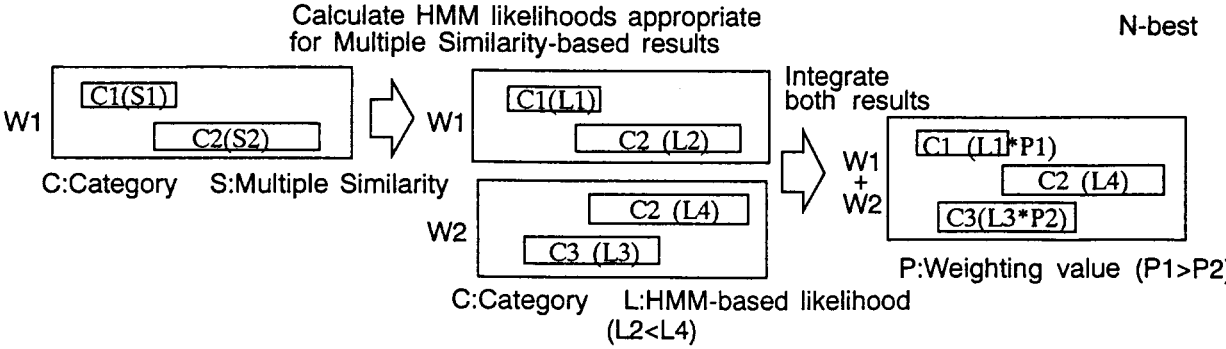

Fig. 3 Integration of word candidates W1 and W2

the word candidates in W1 and W2 are compared under a unified criterion. For this purpose, the likelihood of word candidates in W1 is calculated using the appropriate word HMM present in the same period, as in Fig. 3, so that the Multiple Similarities can all be replaced with HMM-based likelihoods.

To prioritize the word-based wordspotting results, the likelihoods of word candidates found in W1 but not in W2 are given a weighting value P1, and those of word candidates found in W2 but not in W1 are given a weighting value P2, which is smaller than P1. When word candidates in W1 and W2 belong to the same category and the overlapping duration is beyond the threshold, a word candidate with a higher likelihood is selected and will not be weighted. Likewise, additional words, for which word-based wordspotting is not carried out, are not weighted. Thus, after checking the threshold, merged word candidates are re-analyzed in the parser to produce the final semantic utterance representations.

2.4 Design of word reference vectors, phoneme HMMs and word HMMs

For designing word reference vectors, we employed Noise Immunity Learning [3]. In this process, initial word reference vectors are generated using clean speech data consisting of isolated words. As the learning process progresses, SNR of learning data is gradually decreased by contaminating the speech signal with noise in order to obtain noise immunity. In addition to background noise, unintentional utterances are used for synthesizing learning speech data [5]. Word feature vectors for learning are automatically extracted based on the Multiple Similarity values by wordspotting. Word reference vectors for the MS method are generated using the spotted word pattern vectors via K-L Expansion based on covariance matrix modification, as below:

$$K^{(l)} = K_0^{(l)} + \alpha \Sigma X^{(l)} X^{(l)t}$$

where $K^{(l)}$ is a modified covariance matrix, $K0^{(l)}$ the original covariance matrix, $X^{(l)}$ a word feature vector, α a modification coefficient, and t a transpose operator.

Word HMMs which consist of phoneme HMMs, are produced as follows. First, after dividing labeled training speech data into two, phoneme data are extracted from one half of the data to be used for generating the phoneme reference vectors for calculating Multiple Similarity. The time series of Multiple Similarities is calculated for the other half using these vectors. With this as input, parameters of a continuous mixture phoneme HMM are trained using the forward-backward algorithm. This means that pattern variations are absorbed by both the Multiple Similarities and continuous mixture HMMs. Next, the word HMM is generated so as to allow phoneme variations arising from coarticulations. To represent them in the HMM model, a phonemic transition network is formed after applying phonemic transformational rules to a string of words, as in Fig. 4; then, the phonemes in the network are replaced by each corresponding phoneme HMM, thereby deriving the word HMM. Thus, for an additional word, a word HMM is constructed automatically once a string of words is registered.

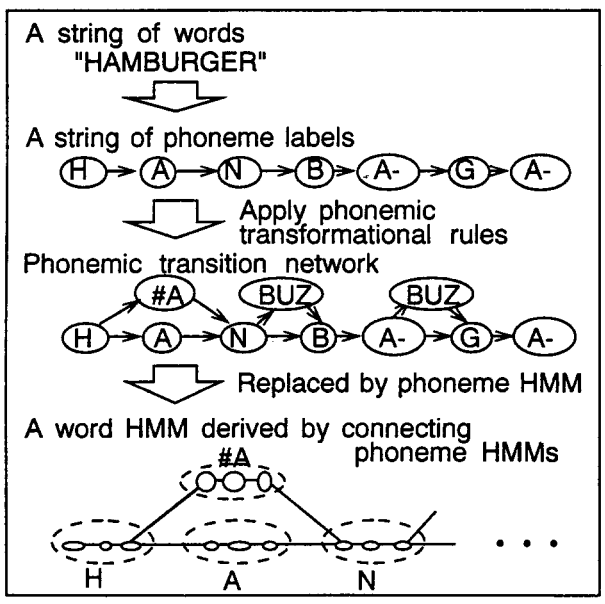

Fig. 4 An example of generating a word HMM

3. EXPERIMENTS

3.1 Database

The experimental conditions are shown in Table 1. For training, 4406 words (uttered by 90 male speakers) were used for making the initial word reference vectors, and 2900 sentences (uttered by 29 male speakers) are used for Noise Immunity learning. 8827 word data, which consist of phoneme balance 492 words uttered by 18 males, was divided into a 4908 word-set by 10 males

Table. 1 Experimental conditions

Training data	Word unit	49 words x 90 males 100 sentences x 29 males
	Phoneme unit	492 words x 18 males
Testing data		350 sentences x 5 males (including 5490 keywords)
Vocabulary		49 words (fast food ordering task)
Word pattern		16 ch x 12 frames
Phoneme label		27 phonemes
Phoneme HMM		4 states, 3 mixture

for training phoneme reference vectors and a 3919 word-set by eight males for training phoneme HMMs. The phoneme boundaries were labeled by hand. The feature vectors for training phoneme HMMs consist of 27-dimensional Multiple Similarities between phoneme reference vectors and spectral patterns. For evaluation, 1750 sentences (uttered by five male speakers) are used.

3.2 Wordspotting experiments

Using the proposed method, we have conducted the following three types of wordspotting experiments:

(1) Using only word pattern matching.
(2) Using the proposed hybrid method.
(3) Using only phoneme HMMs.

In experiment (2), the word candidates of the five best semantic utterance representations in the word-based wordspotting results, are inputted into the phoneme-based wordspotting section. P1 is set to 0.9, and P2 to 0.7. Thresholds of Multiple Similarity and likelihood are defined for each word. These results are given in Table 2. The word detection rates in experiments (1), (2) and (3) are 94.8%, 96.2%, 88.3%, respectively; the sentence understanding rates in experiments (1), (2) and (3) are 64.2%, 68.8%, 55.6%, respectively. On the whole, experiment (2) scored highest, followed by experiments (1) and (3).

The best results in experiment (2) are attributable to the combined use of two different methods and the adoption of a unified criterion. By adopting a parser in the hybrid method, false alarms are reduced to two-thirds of those in experiment (1), and the sentence understanding rate is increased by 4.6%. These results argue for the effectiveness of the proposed wordspotting method.

4. CONCLUSION

We have described a hybrid wordspotting method, combining word-based pattern matching and phoneme-based HMM. Through several experiments, we have confirmed that this method is effective for improving wordspotting performance. Also, our parser contributes to reducing false alarms and to increasing the sentence recognition accuracy. However, since the phoneme-based wordspotting performance is still insufficient, phoneme labels and duration control should be accounted for. We are planing to implement the proposed hybrid wordspotting method to our spontaneous speech dialogue system.

Table. 2 Experimental results

Experiment	(1)	(2)	(3)
Word detection rate (%)	94.8	96.2	88.3
Sentence understanding rate (%)	64.2	68.8	55.6
FA/H/W	25.6	17.5	27.9

ACKNOWLEDGMENTS

We would like to thank H. Tsuboi and H. Hashimoto for their technical suggestions, and M. Shimazu for her help in the preparation of this paper.

REFERENCES

[1] A. L. Higgins and R. E. Wohlford: "Keyword Recognition Using Template Concatenation", ICASSP'85, pp. 1233-1236 (1985)

[2] J. G. Wilpon, L. R. Labiner , C. H. Lee and E. R. Goldman: "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", IEEE Trans. on ASSP, Vol. 38, No. 11, pp. 1870-1878 (Nov. 1990)

[3] Y. Takebayashi, H. Tsuboi and H. Kanazawa: "A Robust Speech Recognition System Using Word-Spotting with Noise Immunity Learning", ICASSP'91, pp. 905-908 (1991)

[4] Y. Takebayashi, Y. Nagata and H. Kanazawa: "Noisy Spontaneous Speech Understanding Using Noise Immunity Keyword Spotting with Adaptive Speech Response Cancellation", ICASSP'93, pp. II-115-II-118 (1993)

[5] Y. Takebayashi, H. Tsuboi and H. Kanazawa: "Keyword-Spotting in Noisy Continuous Speech Using Word Pattern Vector Subabstraction and Noise Immunity Learning", ICASSP'92, pp. II-85-II-88 (1992)

[6] S. Katagiri and C. H. Lee: "A New Hybrid Algorithm for Speech Recognition Based on HMM Segmentation and Learning Vector Quantization", IEEE Trans. on Speech and Audio Process., Vol. 1, No. 4, pp. 421-430 (Oct. 1993)

[7] K. Y. Su and C. H. Lee: "Speech Recognition Using Weighted HMM and Subspace Projection Approaches", IEEE Trans. on Speech and Audio Process., Vol. 2, No. 1, pp. 69-79 (Jan. 1994)

[8] E. A. Martin, R. P. Lippmann and D. B. Paul: "Two-stage Discriminant Analysis for Improved Isolated-word Recognition", ICASSP'87, pp. 709-712 (1987)

[9] L. R. Labiner and J. G. Wilpon: "A Two-pass Pattern-recognition Approach to Isolated Word Recognition", Bell Syst. Tech. J., Vol. 50, No. 5, pp. 739-766 (May, 1981)

[10] R. C. Rose and D. B. Paul: "A Hidden Markov Model Based Keyword Recognition System", ICASSP'90, pp. 129-132 (1990)

[11] H. Tsuboi and Y. Takebayashi: "A Real-Time Task-Oriented Speech Understanding System Using Keyword-Spotting", ICASSP'92, pp. I-197-I-200 (1992)