SPEAKER INDEPENDENT SPEECH RECOGNITION USING DEMISYLLABLE-BASED HIDDEN MARKOV MODELS

L. C. Sauter

Division Informatique,
Laboratoires de Marcoussis
91460 MARCOUSSIS, FRANCE

ABSTRACT

Previously reported work [1-2] used dynamic programming for demi-syllable based speaker independent recognition of isolated words. This paper describes the use of Hidden Markov Models (HMM) for representing the demisyllables, with and without vector quantization of the parametric representation of the speech signal. In a first experiment, whole word models are obtained by concatenating segment models in a manner equivalent to the one described in [1]. The main advantage of the HMM approach is a reduction of both processing time and storage requirements.

INTRODUCTION

Speaker independence in isolated word recognition can be achieved by different methods such as speaker adaptation, the use of multiple references for dynamic template matching or, more recently, Hidden Markov Models (HMM). In the two latter approaches, the training set contains utterances of each of the words in the vocabulary by enough speakers to take into account inter speaker variations [3]. This type of training makes creating a new vocabulary a long and costly process. This is not important for standard vocabularies such as numbers, but is a problem for task specific vocabularies or whenever the vocabulary must change in time. For these reasons, an approach to speaker independence is being investigated where each word in the vocabulary is represented by a sequence of segments. Each segment must be represented by several recordings taking into account inter speaker variations. In previously reported work, such an approach has been described using dynamic template matching [1-2]. Using Hidden Markov Models should allow to reduce the storage requirements as well as the processing time for such a recognizer. The present paper describes some experiments using speaker independent models for words obtained by concatenating speaker independent models for segments. In order to keep the discussion self-contained, we first recall the results of our experiments using dynamic time warping. We then give some background material on Hidden Markov Models. We will then describe our current experiments and give results. Finally we will attempt to outline future work based on these results.

PREVIOUS RESULTS ON SEGMENT BASED SPEECH RECOGNITION

Segments for Speech Recognition

One of the major problems in speech recognition is how to deal with coarticulation. The way the context influences different phonemes is not well known and is very difficult to handle automatically. An efficient way of dealing with phonological variability is to use segments containing several phonemes, such as diphones, syllables or demisyllables [4-5].

Whole word reference patterns can be reconstructed by concatenating segment patterns. Note that this method requires one dynamic time warp for each entry in the lexicon. We have examined how this method can be used for speaker independent recognition.

Partially supported by the CEC (ESPRIT project 64)

A different approach is based on algorithms designed for connected word recognition using whole word templates, yielding a "connected segments" algorithm. The search procedure is guided by the lexicon, in the same way as syntax is used for connected speech.

Test Corpus

The different recognizers were tested using a corpus of 35 words pronounced by 20 different speakers (10 male and 10 female). Each word was split into a sequence of segments, yielding 87 different segments. The total corpus contains 1740 segments.

We have also picked a test vocabulary of 18 words that can be obtained by concatenating segments belonging to the corpus (Table I). The vocabulary is difficult : it contains several minimal pairs such as vitre-mitre-titre. Also, the words were chosen so that whenever possible the context for each segment was not the same as in the word from which it had been extracted. The test vocabulary was recorded by 10 speakers (5 male and 5 female), none of which belonged to the original 20 speakers. The end-point detection was done automatically.

titre	mitre	vitre
mité	cité	paté
traire	traitre	termite
santé	sente	quantité
fête	fez	cassette
douter	divin	passoire

Table I : Test Vocabulary

Construction of Whole-Word References using Segments

In this approach, whole word templates were constructed by concatenating demisyllables, using in the same template segments pronounced by only one speaker. The boundaries between segments were smoothed. 5 reference templates for each word in the vocabulary were computed from the available templates using the Dynamic Clusters Algorithm and dynamic averaging. It is also possible to construct whole word templates by concatenating reference segments obtained by clustering and averaging.

The number of errors was 16 (less than 9%). Directly using words that have been obtained by concatenating reference segments yields poor performances.

Connected Segments Type Recognition

The best results obtained by this approach were 15 errors for the 180 word test corpus (the recognition rate is approximately 92%).

HIDDEN MARKOV MODELS

In this paragraph we shall give some background information on Hidden Markov Models (see also [6-8]). A Markov process is a stochastic process in which the transition probabilities depend only on the current state and not on past states. More formally, a stochastic process x_n is a Markov process if for every n the probability density $v(x_n)$ has the property

$$v(x_n \mid x_{n-1} \cdots x_1) = v(x_n \mid x_{n-1})$$

The use of Hidden Markov Models in speech recognition is based on the assumption that speech is produced by an underlying Markov process. The states themselves are not observable, hence the name "hidden". The observed speech s_n is actually a stochastic function of the Markov process, determined by its density function $f(s_n \mid x_n)$. (Note that we assume here that the output density depends only on the current state.)

A Hidden Markov Model **M** is characterized by its (finite) number of states N, its transition matrix V and the probability density function for the observations $f_i(s)$.

$$v_{ij} = Prob(x_t = j \mid x_{t-1} = i)$$

$$f_i(s_t) = f(s_t \mid x_t = i)$$

Speech recognition using the Hidden Markov Model technique consists of two phases : training and recognition. During the training phase, we must compute a model for each word (or segment) in the vocabulary. This is usually done by finding the transition matrix V and density functions $f_i(s)$ which maximize the a-posteriori probability of producing the training utterances. Recognition is performed by choosing the word whose model most likely produced the test utterance.

EXPERIMENT 1

A first experiment using HMM's has been conducted. In this experiment, HMM's were obtained for each segment in the data-base. A HMM was then computed for each word in the test vocabulary by concatenating segment models. These models were then used for recognizing the test utterances. This procedure is very similar to the experiment using dynamic-time warping and whole word templates obtained by concatenation. An important difference is that only one model was computed for each word instead of five references in the previous system. In this experiment, vector quantization has been used. Therefore the observations all belong to a finite dictionary so the probability densities $f_i(s)$ are discrete. The number of elements in the dictionary was 64. Increasing the size of the dictionary did not improve the recognition. The dictionary was obtained using the dynamic clusters algorithm on the set of reference segments that had been computed in our previous experiments.

The HMM for each segment has two states. The transition matrix is a 2×2 lower triangular matrix of the form

$$\begin{bmatrix} \nu & 0 \\ 1-\nu & 1 \end{bmatrix}$$

The model for each segment is obtained with the Forward-Backward algorithm. The model for each word is obtained by concatenating the segment models. The second column of each segment transition matrix is replaced by the column-vector

$$\begin{bmatrix} \alpha \\ 1-\alpha \end{bmatrix}$$

where α has experimentally been tuned to 0.5. This yields a transition matrix of the form

$$\begin{bmatrix} \nu_1 & \alpha & 0 & 0 & 0 & 0 & \cdots \\ 1-\nu_1 & 1-\alpha & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & \nu_2 & \alpha & 0 & 0 & \cdots \\ 0 & 0 & 1-\nu_2 & 1-\alpha & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \nu_3 & \alpha & \cdots \\ 0 & 0 & 0 & 0 & 1-\nu_3 & 1-\alpha & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

The recognition is then performed with the Viterbi algorithm.

EXPERIMENT 2

In this experiment no vector quantization was used. The probability densities $f_i(s)$ were assumed to be normal, with diagonal covariance matrices. This is only an initial experiment, since it is known that gaussian mixtures should give better results [9]. The main differences with the previous experiment concern the training procedure where new estimates for the means and for the covariance matrices are computed instead of recording the probability of each symbol in the finite alphabet. See [8] for more information on this computation.

RESULTS

The number of errors obtained with dynamic programming, no vector quantization and 5 references for each word was 16 ; with vector quantization the system tested yielded 21 errors. The number of recognition errors in the first experiment was 27. This is approximately 25% more errors than the experiment using dynamic time warping. The number of errors for the second experiment was 30. These rather poor results are probably due to the initial guesses used to start the Forward-Backward algorithm. This issue will be investigated.

FUTURE WORK

Better probability densities, such as gaussian mixtures will be used for the continuous density HMM's. This should greatly improve the performances of our system.

The connected segments algorithm used in our previous experiments will also be redesigned to use the hidden Markov models for segments, in a manner similar to the connected-word approach described in [10]. We believe this should also improve the recognition scores.

REFERENCES

[1] L. Sauter, "Segment-based Speaker Independent Isolated Word Recognition," *Proc. Congrès AFCET Informatique*, Paris, Jan 1985.

[2] L. C. Sauter, "Isolated Word Recognition using a Segmental Approach,", *Proc. ICASSP 85*, pp. 850-853, Mar 1985.

[3] L. R. Rabiner and J. G. Wilpon, "Speaker Independent Isolated Word Recognition for a Moderate Size (54 Word) Vocabulary," *IEEE Trans. Acoust. Speech and Sign. Proc.*, Vol. ASSP-27, No. 6, pp. 583-587, Dec. 1979.

[4] O. Fujimura, "Syllable as a Unit of Speech Recognition," *IEEE Trans. Acoust. Speech and Sign. Proc.*, Vol. ASSP-23, No. 1, pp. 79-82, Feb. 1975.

[5] Th. Schotola, "On the use of demisyllables in automatic word recognition," *Speech Communication*, Vol. 3, No. 1, pp. 63-86, April 1984.

[6] L. E. Baum, T. Petrie, G. Soules and N. Weiss, "A Maximization Technique Occuring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The Ann. of Math. Stat.*, Vol. 41, No. 1, pp. 164-171, 1970.

[7] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. of the IEEE*, vol. 64, no. 4, pp. 532-556, April 1976

[8] L. A. Liporace, "Maximum Likelihood Estimation for Multivariate Observations of Markov Sources," *IEEE Trans. on Inform. Theory*, Vol. IT-28, No. 5, pp. 729-734, Sep. 1982.

[9] L. R. Rabiner, B.-H Juang, S. E. Levinson and M. M. Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities," *AT&T Tech. Journ.*, VOl. 64, No. 6, Part 1, pp. 1211-1234, July-Aug. 1985.

[10] L. R. Rabiner and S. E. Levinson, "A Speaker-Independent, Syntax-Directed, Connected Word Recognition Sytem Based on Hidden Markov Models and Level Building," *IEEE Trans. on Acoust., Speech and Sign. Proc.*, Vol. ASSP-33, No. 3, pp. 561-573, June 1985.