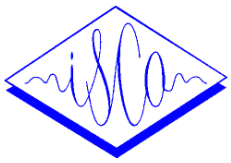

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

EXPERIMENTS WITH A SPEAKER-INDEPENDENT CONTINUOUS SPEECH RECOGNITION SYSTEM ON THE TIMIT DATABASE

Yunxin Zhao and Hisashi Wakita

Speech Technology Lab, Panasonic Technologies Inc.
Santa Barbara, CA 93105, USA

ABSTRACT

This paper describes a speaker-independent, continuous speech recognition system that we designed and implemented, and reports some of the major features of this system with experimental results on a subset of the TIMIT database. The system is based on hidden Markov modeling of phoneme-sized acoustic units using continuous mixture Gaussian densities. The mixture densities are generated using an algorithm which minimizes the average trace of the mixture components and makes use of the segmental structure of the speech signals. Methods of preparing a dictionary for decoding and controlling the perplexity of grammars are also elaborated. On a subset of TIMIT database with 443 words and a grammar perplexity of 49, the system achieved decoding rates of 87.4% sentence correct, 97.9% word correct and 97.5% word accuracy on the training set; on the test set, the rates are 72.4% sentence correct, 93.9% word correct and 92.4% word accuracy.

I. INTRODUCTION

In recent years, a number of HMM based speaker-independent continuous speech recognition systems have been reported with success [1,2]. In this paper, we introduce a speaker-independent, continuous speech recognition system that we designed and implemented. On the acoustic signal level, continuous mixture Gaussian densities are used to model phoneme-sized acoustic units which are generated from an algorithm consisting of the processes of segmentation, estimation, and merging. On the word level, the phonemic labels in the TIMIT database are extracted, and the transcriptions of words are compressed into a dictionary. On the sentence level, context-sensitive grammatical parts are defined to generate grammars of different levels of perplexities. The paper is organized into four sections. In Section II, an overview is given on the structure and features of the system, where we develop an algorithm for estimating the mixture Gaussian densities of phone models, a technique of preparing a dictionary for sentence decoding, and a method of controlling grammar perplexity. Experimental results on a subset of the TIMIT database are provided in Section III, and some discussions and a conclusion are in Section IV.

II. SYSTEM OVERVIEW

A block diagram of the system is shown in Fig. 1, which consists of modules for feature extraction, phone model training, dictionary preparation, grammar estimation, and sentence decoding. In the following we discuss some modules in more detail.

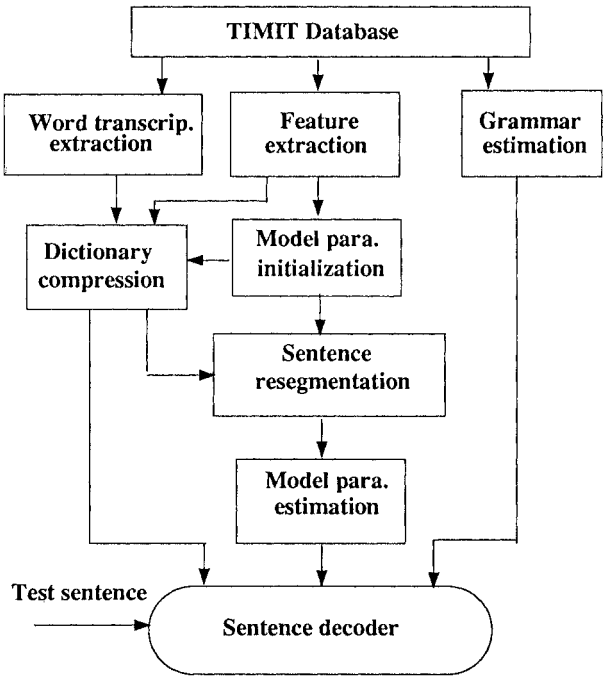

Fig. 1 The overview of the system

2.1 The TIMIT Database

The DARPA TIMIT continuous speech database was designed to provide acoustic phonetic speech data for the development and evaluation of automatic speech recognition systems [3]. A complete labeling of the sentences with acoustic-phonetic units is available in the database,

which provides us with a good source of studying continuous speech. Some training modules in our recognition system were designed to make use of these labels.

2.2 Feature Extraction

The analysis front-end uses perceptually based linear prediction (PLP), which transforms the spectra of speech signals into auditory spectra which are further fitted by an all pole model [4]. The cepstrum coefficients of PLP are weighted by a linear scale [5]. These weighted cepstrum coefficients and log power are used as instantaneous features, and the temporal regression coefficients of each instantaneous feature are then taken as dynamic features [6].

2.3 The HMM of Acoustic-Phonetic Units

The HMM phone models have 3 tied-states, each state being modeled by a mixture Gaussian density with a block diagonal covariance matrix for each mixture component. The topology of an HMM is shown in Fig.2. This topology limits the duration of a phone to be at least 2 frames long.

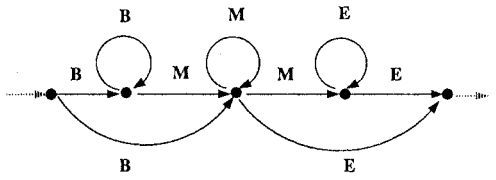

Fig. 2 The topology of an HMM phone model

The instantaneous and dynamic features are assumed independent. Let an instantaneous feature vector be $x = (c_1, ..., c_L, p)'$, where the c_i's are weighted cepstrum coefficients and p is the log power, and let a corresponding dynamic feature vector be $\Delta x = (\Delta c_1, ..., \Delta c_L, \Delta p)'$. Then a mixture Gaussian density is defined as

$$f(x, \Delta x) = \sum_j a_j f_{s,j}(x) f_{d,j}(\Delta x),$$

where $f_{s,j} \sim \mathcal{N}(\mu_{s,j}, C_{s,j})$, $f_{d,j} \sim \mathcal{N}(\mu_{d,j}, C_{d,j})$, and the a_j's are the weights of the mixture components. Note that the product of $f_{s,j}$ and $f_{d,j}$ is equivalent to a single Gaussian density with mean vector $\mu_j = (\mu'_{s,j}, \mu'_{d,j})'$ and covariance matrix $C_j = \begin{pmatrix} C_{s,j} & 0 \\ 0 & C_{d,j} \end{pmatrix}$.

From an autoregressive modeling point of view, each stationary segment of a phoneme can be considered as coming from a single Gaussian source; we therefore choose to train the HMM models by using the segmental structure of the speech signals. Since the sentences are completely labeled in the TIMIT database, it is straightforward to extract phonetic-acoustic segments from the sentences. To further divide each segment into finer units corresponding to states, a simple rule of 1/4, 1/2, and 1/4 is used to cut a segment into 3 subsegments. A maximum likelihood estimation of the mean and covariance matrix is then carried over on each subsegment.

Assume that for a specific state of a specific phoneme there are N segments corresponding to N Gaussian densities. To obtain estimates of a mixture density from the N densities, a merge algorithm is developed based on a minimum trace criterion which will be elaborated in the following subsection. The mixture Gaussian densities thus generated become the initial HMM models of the acoustic units. The initialized HMM models will be used to compress a dictionary directly extracted from the acoustic-phonetic labels of the training sentences. According to the new dictionary and the initial HMM models, each sentence will be automatically resegmented using the Viterbi decoding algorithm. The newly obtained segments are then used to estimate the individual Gaussian densities, and these densities are once again merged into mixture Gaussian densities using the same merge algorithm.

2.4 The Merge Algorithm for Mixture Densities

Assuming there are N Gaussian densities, $\mathcal{N}(\mu_i, C_i)$, each is estimated from data of a sample size $\#\Omega_i$, $i = 1, 2, ..., N$. A measure of concentration of the N Gaussian densities is the average trace of the N corresponding covariance matrices T_N:

$$T_N = \sum_{i=1}^{N} p_i \, trace(C_i),$$

where $p_i = \frac{\#\Omega_i}{\sum_{j=1}^{N} \#\Omega_j}$. Suppose the jth and kth Gaussian densities are merged into one Gaussian density indexed by l, then the total number of densities will be N–1, and the newly merged Gaussian density has the covariance matrix

$$C_l = p'_j C_j + p'_k C_k + p'_j p'_k (\mu_j - \mu_k)(\mu_j - \mu_k)',$$

where $p'_j = \frac{\#\Omega_j}{\#\Omega_j + \#\Omega_k}$ and $p'_k = \frac{\#\Omega_k}{\#\Omega_j + \#\Omega_k}$. The average trace of the N–1 remaining covariance matrices then becomes

$$T_{N-1} = \sum_{i=1}^{N} p_i \, trace(C_i) + (p_j + p_k) p'_j p'_k \|\mu_j - \mu_k\|^2,$$

thus

$$\begin{aligned} \Delta T(j,k) &= T_{N-1} - T_N \\ &= (p_j + p_k) p'_j p'_k \|\mu_j - \mu_k\|^2 \\ &\geq 0. \end{aligned}$$

At each step of merge, the pair of Gaussian densities which minimizes the increment of the average trace ΔT are selected, i.e.,

$$(j^*, k^*) = \underset{(j,k)}{argmin}\ \Delta T(j,k).$$

Since a feature vector consists of components of different scales, the measure of the Euclidian distance between two mean vectors needs to be properly weighted for each component. Thus the modified average trace is defined as

$$T_N = \sum_{i=1}^{N} p_i\, trace(AC_i),$$

where $A = diag(a_{s,1}, ..., a_{s,L+1}, a_{d,1}, ..., a_{d,L+1})$, and each diagonal element of the matrix is a weight of the corresponding feature in distance calculation.

The weight matrix A is phoneme specific. For a particular phoneme, let the total number of frame samples be K, 4 scale quantities r_1, r_2, r_3, r_4 are then calculated for $(c_1, ..., c_L)$, p, $(\Delta c_1, ..., \Delta c_L)$, and Δp, where $r_1 = \sum_{i=1}^{L}\sum_{k=1}^{K} c_{i,k}^2$, $r_2 = \sum_{k=1}^{K} p_k^2$, $r_3 = \sum_{i=1}^{L}\sum_{k=1}^{K} (\Delta c_{i,k})^2$, and $r_4 = \sum_{k=1}^{K} (\Delta p_k)^2$. Introducing a weighting factor w_d for the dynamic features, and letting $S = \frac{1}{r_1} + \frac{1}{Lr_2} + \frac{1}{w_d r_3} + \frac{1}{w_d L r_4}$, the elements of A are then defined as

$$a_{s,i} = \frac{1}{r_1} S^{-1},\ \ i = 1, ..., L$$
$$a_{s,L+1} = \frac{1}{Lr_2} S^{-1}$$
$$a_{d,i} = \frac{1}{w_d r_3} S^{-1},\ \ i = 1, ..., L$$
$$a_{d,L+1} = \frac{1}{w_d L r_4} S^{-1},$$

where the weight w_d was experimentally chosen as 2.

2.5 Context Dependency Modeling

When only a small amount of training data is available for each context, tuning weights of mixture Gaussian densities with respect to contexts is more realistic and has been shown effective [2]. To estimate the weights of mixture densities, we use a simple maximum likelihood estimation approach. Let the data sample size of a context be K, the objective function is defined as

$$J = \log \prod_{k=1}^{K} \left(\sum_i a_i f_i(x_k, \Delta x_k) \right) + \lambda \left(\sum_i a_i \right).$$

Maximizing J with respect to each a_i leads to the recursive estimate of a_i as

$$a_i^{(n+1)} = \frac{1}{K} \sum_{k=1}^{K} \frac{a_i^{(n)} f_i(x_k, \Delta x_k)}{\sum_i a_i^{(n)} f_i(x_k, \Delta x_k)}.$$

One iteration has been shown sufficient in our experiments.

2.6 Model Smoothing

For robustness in modeling, the covariance matrices of mixture components are smoothed by the unimodal Gaussian density of the corresponding phoneme using a simple interpolation. Let the covariance matrix of the unimodal Gaussian density be C and that of a mixture component be C_i, then the smoothed covariance matrix is $\hat{C}_i = \lambda C_i + (1-\lambda)C$. Similarly, the context dependent mixture weights are smoothed by the context independent mixture weights. Let $a_{m,i}$ be a weight of a context m, then the smoothed weight is $\hat{a}_{m,i} = \lambda a_{m,i} + (1-\lambda)a_i$.

2.7 Dictionary Preparation

The phoneme labels of the sentences in the TIMIT database can yield many transcriptions for a single word. The variety of transcriptions of a word could reflect dialect, context, as well as some possible mislabelings. It is desirable for sentence decoding to compress the directly extracted dictionary into a more succinct one through an automatic procedure which will advocate the transcriptions that occur frequently, while allowing some less frequent ones if these words cannot be fitted by other transcriptions.

Specifically, let the index of the transcription of a word w_k be i, i.e., $T(w_k) = i$, and let the HMM model associated with the transcription (a concatenation of phoneme HMM models according to the transcription) be $M^{(i)}$. The cost of changing the transcription from i to j is defined as

$$C(i \rightarrow j) = \sum_{k:T(w_k)=i} \frac{1}{N_k} \left\{ \log \frac{p(w_k/M^{(i)})}{p(w_k/M^{(j)})} \right\}$$

Note that the likelihood ratio is normalized by the length, N_k, of the word, and the cost of changing one transcription to another is weighted by the number of words labeled by the corresponding transcription. At each step of compression, a pair of transcriptions is selected which will incur the minimum cost from changing one into the other. The process stops when the minimum cost exceeds a threshold or only one transcription is left. In the experiments, we found that a majority of transcriptions thus generated is satisfactory, yet some transcriptions still need to be hand corrected.

2.8 Grammar

Word pair grammar has been used in a number of continuous speech recognition systems for controlling the task difficulty [7]. For well structured continuous speech database such as the DARPA Resource Management Database, the perplexity of a word pair grammar is reasonable. The TIMIT database, however, does not seem to have a very well constrained sentence structure, and the perplexity of a word pair grammar is quite high.

To obtain a better control of grammar perplexity, the word pair grammar is modified such that context dependent grammatical parts are defined for smoothing the transition between words instead of using the grammatical parts directly. Specifically, consider a 2–sided 1st order context dependency. Let the center grammatical part be a, its left neighbor be b, and its right neighbor be c, then a context dependent grammatical part is defined as $g_{(b,a,c)}$. The transition from a word w_j to a word w_k is bridged by the transition of their corresponding grammatical parts. Specifically, if the quantity

$$\sum_{g_{(b,a,c)}} \sum_{g_{(b',a',c')}} p(w_j/g_{(b,a,c)}) p\Big(g_{(b,a,c)}-g_{(b',a',c')}\Big) p\Big(w_k/g_{(b',a',c')}\Big)$$

is nonzero, the word w_k can follow w_j, and the transition from w_j to all its followers will be assigned to an equal probability. Grammars at different levels of perplexity can be obtained by varying the degree of context dependency of the grammatical parts.

2.9 Sentence Decoding

Sentence decoding is based on the Viterbi decoding algorithm, which integrates the HMM phone models, the dictionary (allowing multiple pronunciation per word), and the grammar into a constrained, optimized path search. The amount of nodes expanded at each step is controlled by a threshold of beam search. The statistics of word durations are modeled by Gaussian densities and were used in decoding, which are useful for reducing insertion errors. The path that has the highest likelihood score is back tracked to produce the decoded word string of the sentence.

III. EXPERIMENTS

Experiments have been performed on a subset of the TIMIT database. As we only have the released training set of the database, the experiments were done by splitting the available data into a training set and a test set. The speech signals are down sampled from 16 KHz to 10.67 KHz. Features are computed using a PLP analysis of order 8, every 100 samples step, and 200 samples long per frame. There are 98 distinct sentences (SX3 — SX100) with a vocabulary size of 443. The training set has 356 sentences, spoken by 325 speakers; the test set has the same 98 sentences, spoken by 89 speakers different from those in the training set. The sentences were labeled by 19 grammatical parts, and 2 types of context dependent grammatical parts were generated, one with the 1st order dependency on both left and right neighbors, giving a grammar perplexity of 12, and the other with the 1st order dependency on the left neighbor only, giving a grammar perplexity of 49. The decoding rates are scored using the standard routine provided by National Institute of Standards and Technology, and are shown in Table 1.

Table 1. Decoding rates

		Sent. correct	Word correct	Word accuracy
perp. ≈ 12	Training set	93.5%	98.9%	98.9%
	Test set	82.7%	95.8%	94.8%
perp. ≈ 49	Training set	87.4%	97.9%	97.5%
	Test set	72.4%	93.7%	92.4%

IV. CONCLUSION

In this paper, we presented some features of a speaker-independent continuous speech recognition system that we designed and implemented. Some modules in the system have been designed to make use of the information available in the TIMIT database, and these modules can be easily modified to deal with data from other sources. Some parts of the system need to be further tuned, such as the termination threshold of the merge algorithms. Our current experiments on the TIMIT database showed promising results. Work is currently in progress to further improve the system performance.

ACKNOWLEDGMENT

Some programming help by Brian Mak, labeling by Lisa King and Lillian Stuman, are sincerely acknowledged.

REFERENCES

[1]. K. F. Lee, Speaker-Independent Continuous Speech Recognition Using Hidden Markov Models, PhD dissertation, Carnegie-Mellon University, 1988.

[2]. C. H. Lee, "Acoustic Modeling of Subword Units for Speech Recognition," Proc. ICASSP, pp. 721–724, Albuquerque, Apr., 1990.

[3]. L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," Proceedings of Speech Recognition Workshop (DARPA), 1986.

[4]. H. Hermansky, B. A. Hanson, H. J. Wakita, "Perceptually Based Linear Predictive Analysis of Speech," Proc. ICASSP, pp.509–512, Tampa, Florida, 1985.

[5]. B. Hanson and H. Wakita, "Spectral Slope Distance Measures with Linear Prediction Analysis for Word Recognition in Noise," IEEE Trans. ASSP, ASSP-35, pp. 968–973, 1987.

[6]. S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," IEEE Trans. ASSP,ASSP-34, pp. 52–59, 1986.

[7]. F. Kubala, Y. Chow, A. Derr, M. Feng, O.Kimball, J. Makhoul, P. Price, J. Rohlicek, S. Roucos, R. Schwartz, and J. Vandegrift, "Continuous Speech Recognition Results of the BYBLOS System on the DARPA 1000–Word Resource Management Database," Proc. ICASSP, pp. 291–294, New York, Apr. 1988.