Correspondence

An RNN-Based Preclassification Method for Fast Continuous Mandarin Speech Recognition

Sin-Horng Chen, Yuan-Fu Liao, Song-Mao Chiang, and Saga Chang

***Abstract*—A novel recurrent neural network-based (RNN-based) front-end preclassification scheme for fast continuous Mandarin speech recognition is proposed in this paper. First, an RNN is employed to discriminate each input frame for the three broad classes of initial, final, and silence. A finite state machine (FSM) is then used to classify the input frame into four states including three stable states of Initial (I), Final (F), and Silence (S), and a Transient (T) state. The decision is made based on examining whether the RNN discriminates well between classes. We then restrict the search space for the three stable states in the following DP search to speed up the recognition process. Efficiency of the proposed scheme was examined by simulations in which we incorporate it with a hidden Markov model-based (HMM-based) continuous 411 Mandarin base-syllables recognizer. Experimental results showed that it can be used in conjunction with the beam search to greatly reduce the computational complexity of the HMM recognizer while keeping the recognition rate almost undegraded.**

I. Introduction

The recognition process of continuous speech recognition is essentially a search procedure to determine the optimal matching path that maps the testing utterance to a string of reference word (or subword) models. A basic problem is that there is typically a huge number of possible paths, so that intensive computations are needed. Usually, the path pruning approach is used to solve the problem. It uses a mechanism to prune some unlikely paths for reducing the computational complexity. The beam search and the A^* search with a tree-based lexicon [1]–[4] are two well-known methods. Recently, some phoneme level pruning techniques, which utilize the local probability estimates generated by the detailed recognizers themselves for path pruning, have also been studied. The phoneme look-ahead method [3] estimates the likelihood of each phoneme ahead of the current time frame. Only the succeeding phonemes with likelihood falling within the preset envelope remain to survive in the following search. The phone deactivation method [4] first estimates the local *posteriori* probabilities of phonemes by using a recurrent neural network (RNN), and then prunes all words containing those unlikely phonemes with low *posteriori* probabilities.

An alternative approach uses a simple front-end processor to preclassify the current frame or to presegment the input speech for reducing the search space of the following recognition process. A method of this approach is to classify each input frame into voiced, unvoiced, or silence [5], [6] and then to compress the search space by restricting the frame to stay on some legal states. But this approach is rarely used in the current existing continuous speech recognition systems. A fundamental problem comes from the fact that any error in the front-end processor may result in a fatal error in the following search procedure to seriously degrade the recognition performance.

Manuscript received October 19, 1995; revised March 4, 1997. This work was supported by the National Science Council, Taiwan, R.O.C., under Contract NSC85-2213-E009-110. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Douglas D. O'Shaughnessy.

The authors are with the Department of Communication Engineering, National Chiao Tung University, Hsinchu 300, Taiwan, R.O.C.

Publisher Item Identifier S 1063-6676(98)00635-X.

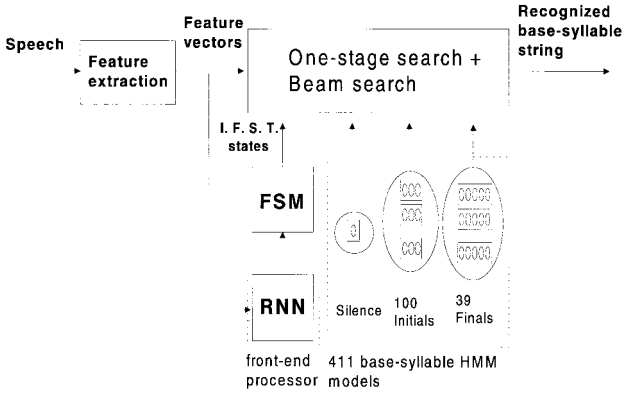

Fig. 1. Block diagram of the proposed fast CDHMM continuous Mandarin syllable recognizer.

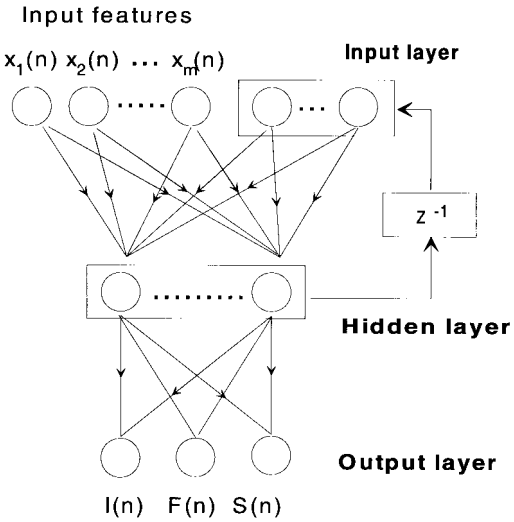

Fig. 2. Structure of the recurrent neural network.

In this correspondence, a novel preclassification scheme for fast continuous Mandarin speech recognition is proposed. A small RNN classifier is firstly used to discriminate each input frame into several broad classes of speech signal. Then a finite state machine (FSM) reflecting "the domain of knowledge" is used to examine whether the responses of the RNN are good enough to make a reliable classification. When the RNN discriminates well between classes, the FSM will make a firm classification to label the input frame into the corresponding stable state associated with the class with best response. Otherwise, it simply puts the input frame into a transient state. Different search spaces are then set in the following DP search for these states in order to reduce the computational complexity. Generally speaking, a more restricted search is used in these stable states, and an unrestricted search is used in the transient state. By this arrangement, most classification errors of the RNN can be tolerated in the sense of causing no trouble in the following recognition

1063–6676/98$10.00 © 1998 IEEE

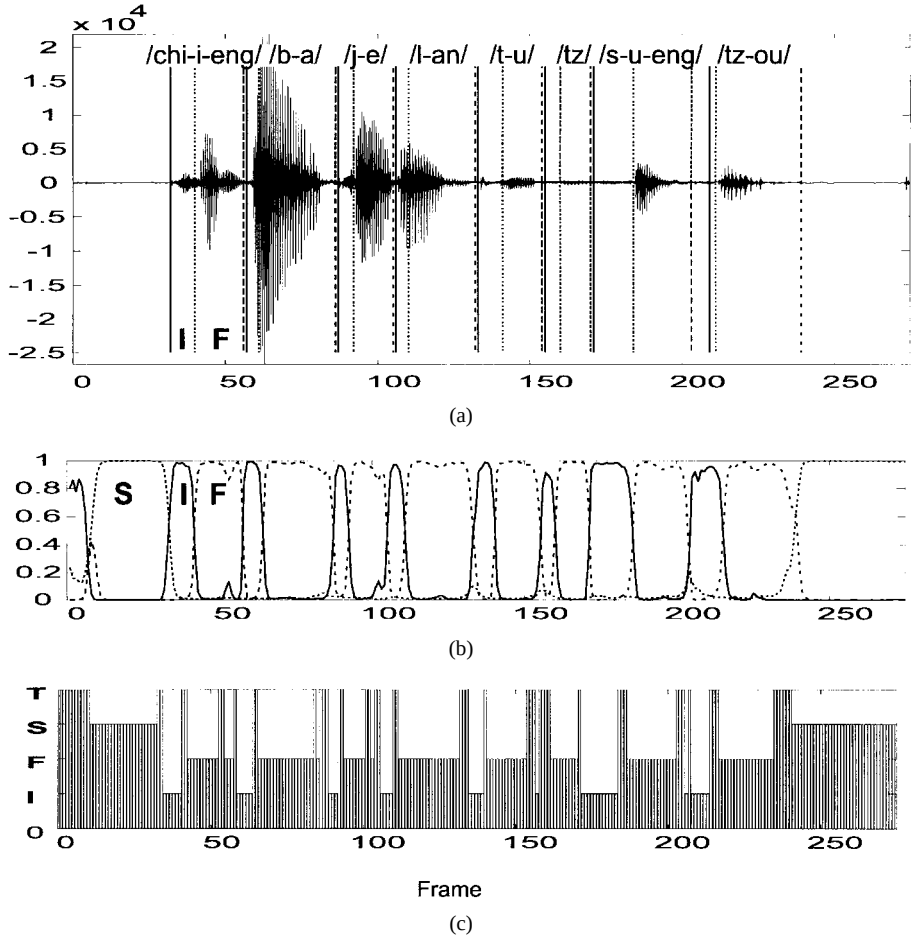

Fig. 3. Typical example showing the three responses of the front-end processor to the input utterance "/*chi-i-eng*/ /*b-a*/ /*j-e*/ /*l-an*/ /*t-u*/ /*tz*/ /*s-u-eng*/ /*tz-ou*/". (a) Waveform and the segmentation positions determined by the baseline CDHMM base-syllable recognizer (solid line: syllable beginning; dotted line: I-F boundary; dashed line: syllable ending). (b) Three responses of the RNN (solid line: $\boldsymbol{I}$; dashed line: $\boldsymbol{F}$; dotted line: $\boldsymbol{S}$). (c) Output of the FSM with $TH_H = 0.9$ and $TH_L = 0.1$.

search to degrade the recognition rate. In other words, we can speed up the DP search while still keeping the recognition performance almost undegraded. It is noted that the preclassification in the FSM is a partial-hard-decision-and-partial-soft-decision scheme from the viewpoint of the DP search. Several advantages of the method can be found as compared with the above-mentioned fast speech recognition methods. First, it is more robust to the preclassification errors than the previous front-end processor-based methods. Second, in addition to making the DP search more efficient like the phoneme level pruning methods, the computations of the likelihood or the *posteriori* probabilities for some unlikely reference word (or subword) models can also be eliminated. Third, it can be used in conjunction with some path pruning techniques, such as the beam search, to further improve the recognition efficiency.

II. The Proposed Preclassification Scheme

Fig. 1 shows the proposed preclassification scheme for fast continuous Mandarin speech recognition. The front-end processor consists of two main parts: an RNN classifier and an FSM. The function of the RNN is to discriminate each input frame for the three classes of silence, initial, and final. It is noted that the last two classes are chosen because initials and finals are commonly used as the basic recognition units in Mandarin speech recognition for taking advantages of the simple initial-final structure of Mandarin syllables (see Table I) [7]. The function of the FSM is to label each input frame into one of three stable states or a transient state based on examining whether the three responses of the RNN are good enough to make a reliable classification.

The RNN is a three-layer network with all outputs of the hidden layer feeding back to the input layer (see Fig. 2). An RNN of this type is a dynamic system with the outputs of its hidden layer at any time depending on a complex aggregate of all previous inputs. So it can easily catch dynamic information of the input speech signal for discriminating some speech patterns [4], [8]–[10]. Here, the RNN is chosen to provide the frame-synchronized preclassification scores with low overhead. The RNN is first trained by the "output delayed" backpropagation (BP) algorithm [9] with all targets being set according to the delayed segmentation positions of the training utterances determined by using an initial-final based continuous density hidden Markov model (CDHMM) recognizer [7]. A typical example of the three responses of the RNN is shown in Fig. 3. It can be seen from the figure that the RNN responds very well to make reliable classifications for most parts of the input speech (see Section IV for quantitative information). In fact, only some short initials surrounded by two vowel finals may cause the RNN fail to respond quickly and correctly. This is mainly owing to the suffering of contextual coarticulation on those short initials.

Based on the three responses of the RNN, a four-state FSM is constructed (Fig. 4). The FSM is designed to conform to the initial-final structure of Mandarin base-syllables. When the RNN discriminates well between classes, we make a hard-decision to move

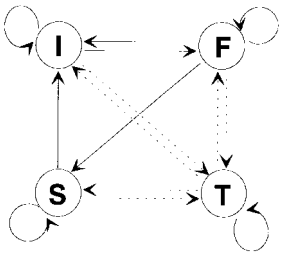

Fig. 4. State diagram of the FSM.

TABLE I
PHONETIC STRUCTURE OF MANDARIN SYLLABLES. THERE ARE IN TOTAL 22 INITIALS (INCLUDING A DUMMY) AND 39 FINALS

TONE			
INITIAL	FINAL		
(Consonant)	(Medial)	Vowel	(Nasal ending)

TABLE II
CONFUSION MATRIX OF THE CLASSIFICATION BY THE RNN

Result/Desired	I	F	S
I	47685	7027	726
F	4915	89014	524
S	564	629	32705

TABLE III
CONFUSION MATRIX OF THE CLASSIFICATION BY THE FSM

Result/Desired	I	F	S	T
I	33579	1235	178	20446
F	636	70340	48	23429
S	106	41	29435	4316

the FSM into one of the three stable states of Initial (I), Final (F), and Silence (S). Otherwise, we make a soft-decision to let the FSM stay in the Transient (T) state in order to tolerate the classification errors of the RNN. To realize the FSM, two thresholds, TH_H and TH_L, are first empirically determined. We then compare the outputs of the RNN with these two thresholds. When one output of the RNN is higher than TH_H and the other two outputs are all lower than TH_L, we move the FSM into the corresponding stable state. Otherwise, the FSM stays at the T state. A typical example of the responses of the FSM is shown in Fig. 3(c).

III. INTEGRATING THE FSM WITH THE DP SEARCH

The proposed preclassification scheme can be incorporated into any initial-final based continuous Mandarin speech recognition system to speed up its recognition process. Mandarin Chinese is a tonal and syllabic language. There exist more than 80 000 words, each composed of from one to several characters. There are more than 10 000 commonly used characters, each pronounced as monosyllable with one of five tones. There are in total 411 base-syllables, disregarding the tones required to cover all necessary pronunciations for Mandarin speech. A complete continuous Mandarin speech recognition system is generally composed of two components: 1) acoustic processing for syllable identification and 2) lexical decoding for word (or character) string recognition [7]. In this study, we only consider the part of acoustic processing. Effectiveness of the proposed method is thus demonstrated via incorporating it with a CDHMM-based continuous 411 Mandarin base-syllables recognizer. The recognizer uses 411 eight-state base-syllable HMM models and a one-state silence HMM model in the recognition search. The 411 base-syllable models are formed by using 100 three-state right-context-dependent initial HMM models and 39 five-state context-independent final HMM models. The number of mixtures in each state of a subsyllable HMM model varies from one to eight depending on the amount of training data. A conventional recognition procedure uses the well-known one-stage DP search embedded with a beam search to find out the best base-syllable sequence for the input testing utterance.

In the proposed fast recognition method, different search spaces in the DP search are set for those four states. Specifically, for frames with I, F, and S states, we let the search space be restricted to stay only in the states of 100 initial HMM's, 39 final HMM's, and the silence HMM, respectively. On the contrary, for frames with T state, unrestricted search is used. By this method, the recognition process can be greatly speeded up with almost no degradation on the recognition rate. In practical implementation, we may slightly relax the search space for frames of F state to include HMM states of 25 short initials for compensating the previously mentioned weak responses of the RNN to them.

IV. EXPERIMENTAL RESULTS

Efficiency of the proposed method was examined by simulations using a continuous Mandarin speech data base uttered by a single male speaker. The database contains in total 35 231 syllables including 28 197 training syllables and 7034 testing syllables. All speech signals were sampled at a rate of 10 kHz and preemphasized with a digital filter, $1 - 0.95z^{-1}$. It was then analyzed for each Hamming-windowed frame of 20 ms with 10 ms frame shift. The recognition features consist of 12 mel-cepstral coefficients, 12 delta mel-cepstral coefficients, and the delta energy. The following definition of syllable accuracy was used to evaluate the performance of the recognition system:

$$\text{syllable accuracy} = 1 - \frac{\text{Substitutions} + \text{deletions} + \text{Insertions}}{\text{number of testing syllables}}. \quad (1)$$

The average number of states to search per frame and the average number of Gaussian components to calculate per frame by the search procedure are used to measure the computational cost in order to avoid any implementation bias.

First, the proposed preclassification front-end processor was tested. An RNN with 25 hidden nodes achieved 92.9% classification rate calculated based on taking the segmentations of all testing utterances by using an initial-final based continuous CDHMM recognizer as reference. The confusion matrix is shown in Table II. After using the FSM to put some marginal frames into the T state, we found that the classification becomes very reliable. For the case of $TH_H = 0.90$ and $TH_L = 0.10$, the classification rate raises to 98.2% with a cost of 26.2% of frames being classified as T state. The confusion matrix is shown in Table III. So most classification errors of the RNN have been absorbed by the transient state of the FSM.

The baseline CDHMM base-syllable recognizer using the one-stage DP search with an embedding beam search was then tested. Several values of beamwidth were tested. The recognition results

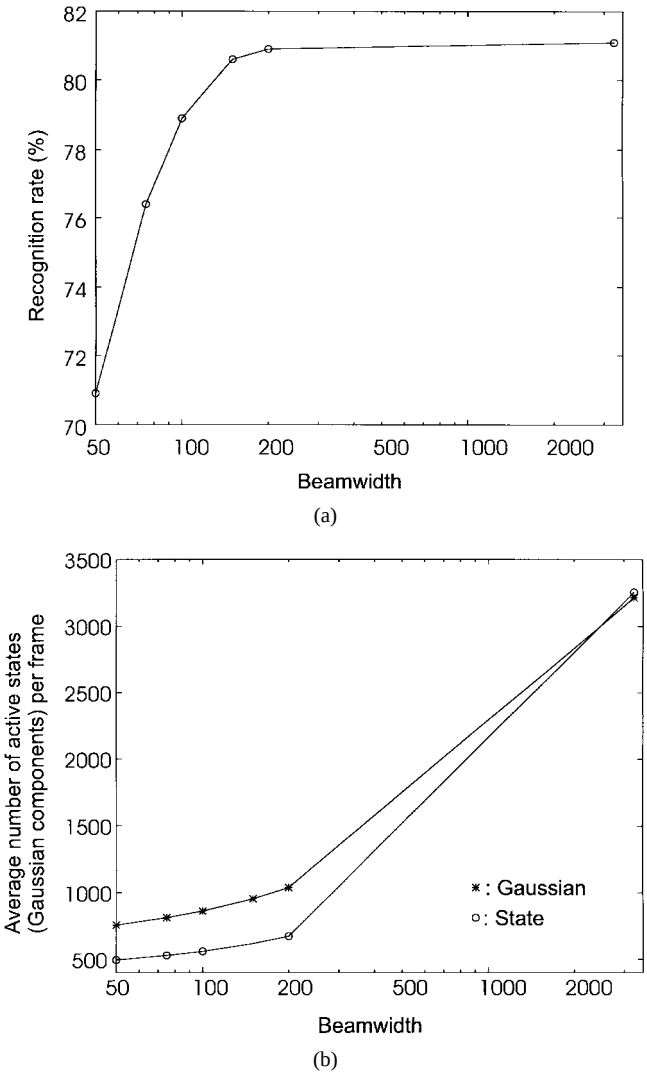

Fig. 5. Recognition results of the baseline CDHMM base-syllable recognizer with several values of beamwidth. (a) Recognition rate. (b) Average number of active states (Gaussian components) per frame.

and the computational costs are plotted in Fig. 5. It is found from the figure that the recognition speed of the one-stage DP can be greatly increased by engaging the beam search. But the recognition rate drops rapidly when the beamwidth is less than 100.

The proposed fast recognition method was then tested. An RNN with 25 hidden nodes was used. Its overhead is approximately equal to the computational load of calculating 30 Gaussian components. Several values of beamwidth and TH_H $(TH_L = 1 - TH_H)$ were tested. The recognition results are plotted in Fig. 6. It is found from the figure that the recognition speed of the DP search with an embedding beam search can be further improved with a negligible small degradation on the recognition rate. For the case of beamwidth = 100 and $TH_H = 0.95$ $(TH_L = 0.05)$, the recognition rate decreases by 0.1% only. But the computational cost is further improved by dropping away additional 38.7% of searching states and by eliminating the likelihood calculations for additional 35.1% of Gaussian components. This confirms the efficiency of the proposed fast recognition method.

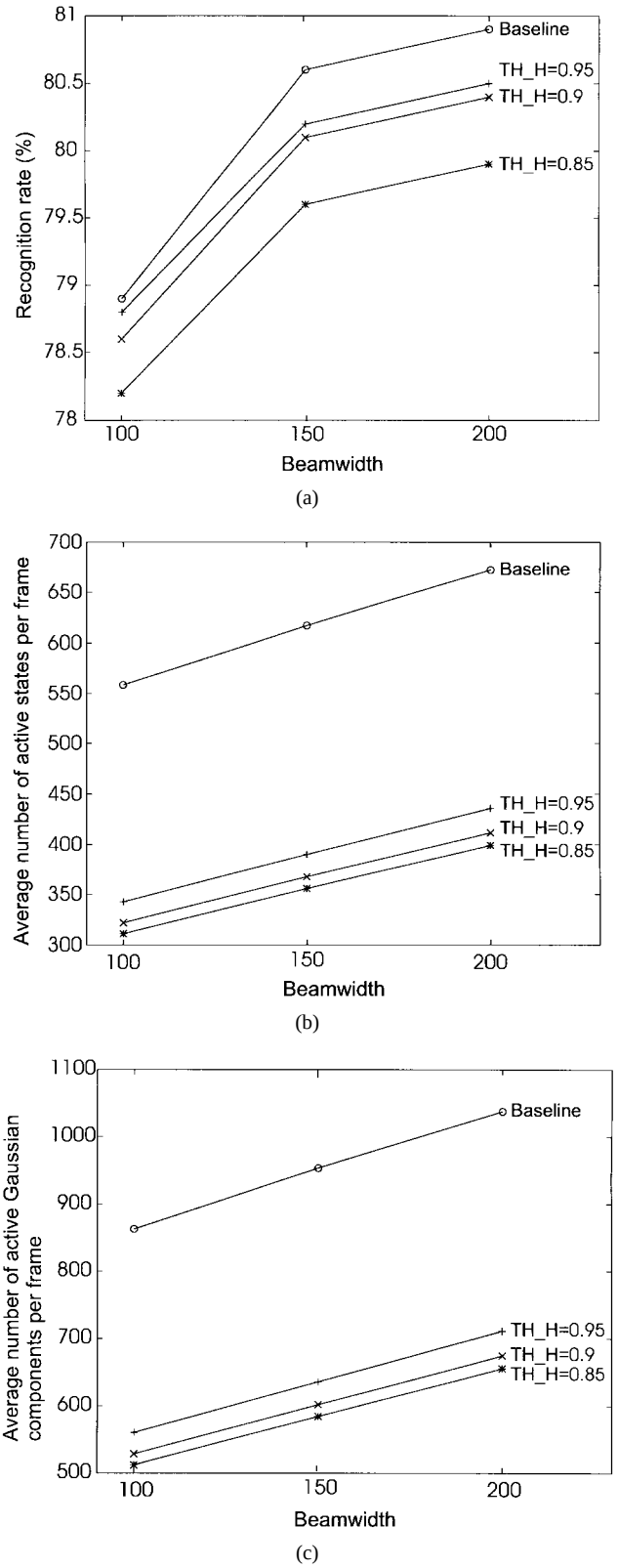

Fig. 6. Recognition results of the proposed fast recognition scheme with the search space for frames of F state being slightly relaxed to include the HMM states of 25 short initials. (a) Recognition rate. (b) Average number of active states per frame. (c) Average number of active Gaussian components per frame.

V. Discussions and Conclusions

In this work, an RNN-based front-end preclassification scheme for fast continuous Mandarin speech recognition has been discussed. Its effectiveness has been demonstrated by simulations to incorporate it into an HMM-based continuous 411 Mandarin base-syllables recog-

nizer. Experimental results showed that it can be used in conjunction with the beam search to greatly reduce the computational complexity of the HMM recognizer while still keeping the recognition rate almost undegraded. Obviously, it is also suitable to be incorporated with other subsyllable-based Mandarin speech recognizers.

An additional advantage of the proposed method was also found. Instead of making a decision at the last frame of the testing utterance done by the conventional one-stage DP search, an early decision can be made once the FSM enters an S state. We can therefore decompose a large complex DP search for the whole utterance into several simpler DP searches for the partitioned voice segments. It is of benefit in reducing system complexity when we consider the incorporation of a language model with the continuous Mandarin base-syllable recognizer.

ACKNOWLEDGMENT

The authors thank Telecommunication Laboratories, MOTC, Taiwan, R.O.C., for kindly supplying the data base.

REFERENCES

[1] H. Ney, D. Mergel, A. Noll, and A. Paeseler, "A data-driven organization of the dynamic programming beam search for continuous speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1987, pp. 833–836.

[2] P. S. Gopalakrishnan, L. R. Bahl, and R. L. Mercer, "A tree search strategy for large-vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1995, pp. 572–575.

[3] R. Haeb-Umbach and H. Ney, "Improvements in beam search for 10000-word continuous-speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 353–356, 1994.

[4] S. Renals and M. Hochberg, "Efficient search using phone probability estimates," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1995, pp. 596–599.

[5] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 201–212, June 1976.

[6] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classification of speech using hybrid feature and a network classifier," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 250–255, 1993.

[7] H. M. Wang, *et al.*, "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1995, vol. I, pp. 61–64.

[8] A. J. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Trans. Neural Networks*, vol. 5, pp. 298–305, Mar. 1994.

[9] A. Hunt, "Recurrent neural network for syllabification," *Speech Commun.*, no. 13, pp. 323–332, 1993.

[10] Y. F. Liao, W. Y. Chen, and S. H. Chen, "Continuous mandarin speech recognition using hierarchical recurrent neural network," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1996, vol. 6, pp. 3371–3374.

Speech Analysis and Recognition Using Interval Statistics Generated from a Composite Auditory Model

H. Sheikhzadeh and L. Deng

***Abstract*—A modeling approach to auditory speech analysis and recognition is proposed and evaluated, where a composite auditory model is used to generate parallel sets of auditory-nerve instantaneous firing rates (IFR's) along the spatial dimension, followed by a processing stage that constructs from the IFR's an interval statistics in a form called the *interpeak interval histogram* (IPIH). A speech preprocessor is designed that performs transformation on the auditory IPIH's and interfaces the IPIH-based auditory representation with a hidden Markov model-based (HMM-based) speech recognizer. The results demonstrate that the new preprocessor consistently outperforms the conventional mel frequency cepstral coefficient-based (MFCC-based) preprocessor for the signal-to-noise ratio (SNR) level up to at least 16 dB.**

I. INTRODUCTION

From evolutionary considerations of the speech production system and of the auditory (hearing) system, it is tempting to take the view that the acoustic properties of speech are uniquely structured so that phonologically meaningful speech sounds can be appropriately represented at all levels of the auditory pathway, given all the physiological constraints imposed by the auditory system. The purpose of this paper is to report an effort intended to reduce this highly speculative view to two concrete issues: first, what exactly is the actual representation of various classes of phonologically defined speech sounds at a lowest level (auditory nerve or AN) of the auditory pathway; and second, can (and how can) a detailed exploration of the nature of such a representation aid the engineering design of practical speech processing systems?

Section I of this correspondence is devoted to addressing the first issue above, for which a wealth of experimental data have been collected since 1979 on AN responses to speech sounds [1]–[8]. The wide scope of these experimental data and detailed analysis of them have made it possible to build a computer model capable of faithfully simulating major characteristics of the AN responses to several isolated speech tokens [9]–[14]. In particular, the temporal aspects of the AN responses to isolated speech tokens and to other complex sounds have been analyzed [15]–[17] and simulated in computer models to such a detailed degree that one can begin to draw conclusions on functional roles of the AN temporal response properties [18]–[21]. Based on all these previous experimental findings and computer simulation results, we are now in a position to address the issue of the nature of the AN representation of a comprehensive set of sounds in fluent speech streams (rather than just a limited set of isolated speech tokens as examined in the past).

Section II of this correspondence addresses the second issue—namely the application of the auditory temporal representation of speech to the design of the front-end (i.e., feature analysis) component of speech recognition systems. The philosophical motivation for possible success of this application is the amazing human speech recognition capabilities and the postulated linkage between acoustic properties of speech and human auditory constraints. The practical

Manuscript received May 10, 1994; revised April 17, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. James H. Snyder.

The authors are with the Department Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont., Canada N2L 3G1.

Publisher Item Identifier S 1063-6676(98)00585-9.

1063–6676/98$10.00 © 1998 IEEE