ITERATIVE SELF-LEARNING SPEAKER AND CHANNEL ADAPTATION UNDER VARIOUS INITIAL CONDITIONS

Yunxin Zhao
Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign,
405 N. Mathews Ave., Urbana, IL 61801   E-mail: yxz@ifp.uiuc.edu

ABSTRACT

A self-learning adaptation technique is presented which handles the speaker and channel induced spectral variations without enrollment speech. At the acoustic level, the distortion spectral bias is estimated in two steps using the unsupervised maximum likelihood estimation: in the first step, the probability distributions of the speech spectral features are assumed uniform for severely mismatched channels; in the second step, the spectral bias is reestimated assuming Gaussian distributions for the spectral features. At the phone unit level, unsupervised sequential adaptation is performed via Bayesian estimation from the on-line, bias-removed speech data, and iterative adaptation is further performed for dictation applications. Over four 198–sentence test sets, on a continuous speech recognition task with vocabulary size = 853 and grammar perplexity = 105, the largest increase of average word accuracy is 85.2% from the baseline accuracy of –0.3%, and the maximum average word accuracy is 89.4% from the baseline accuracy of 56.5%.

1. INTRODUCTION

The current work presents a self-learning technique for improving speaker-independent continuous speech recognition, which performs speaker and channel adaptation while a speech recognizer is used for a certain application, without taking enrollment speech from the speakers. The adaptation is based on a decomposition of the speech spectral variation sources into two categories: one acoustic and the other phone-specific [1,2]. The acoustic source is attributed to speakers' physical individualities and linear channel distortions that cause spectral variations independent of phone units; the phone-specific source is attributed to speakers' idiosyncrasies and piecewise linear channel distortions that cause spectral variations dependent on phone units. The baseline speaker-independent continuous speech recognition system is based on the hidden Markov models of phone units: each phone unit has three states, and each state is modeled by a Gaussian mixture density [3].

A general concern on incorporating self-learning adaptation into continuous speech recognition is whether it would lead to diverged results under adverse conditions. In the current study, experiments have been performed on speech data collected under various conditions, where the worst case has the double problems of an unmatched microphone and a too-close distance between the microphone and the mouth. Since the probability distributions estimated from the training data is grossly inadequate under the severely mismatched and unknown channel conditions, the spectral bias of the acoustic source is first estimated under the assumption that the distorted speech spectral features are uniformly distributed. Based on the knowledge of this spectral bias, the trained speech models are used to reestimate the spectral bias in order to improve the estimate accuracy. After removing the bias, unsupervised sequential adaptation on the phone model parameters is performed via Bayesian estimation from the increasing amount of on-line speech data. For certain applications where delaying recognition output is permissible, for example, voice dictation, the recognition accuracy can be improved by iteratively adapting the phone model parameters and decoding the speech, making use of all the available speech data at the end of a session.

Relating to other efforts in the field, using unsupervised maximum likelihood estimation for spectral bias estimation was proposed in [4]; two independent efforts on unsupervised phone model adaptation were reported in [5,6]. The current work is presented in four sections: the self-learning adaptation method is described in Section 2; the experimental results are presented in Section 3; a summary is given in Section 4.

2. SELF-LEARNING ADAPTATION

The self-learning speech recognition system is illustrated in Fig. 1. Assuming the speaker q speaks one sentence at a time, the adaptation is implemented in two sequential steps for each sentence. The first step is performed before decoding the sentence, where the spectral bias is estimated from the current sentence and these spectra are subsequently normalized. The second step is carried out after decoding the sentence, where the parameters of the phone models are adapted via Bayesian estimation. In the second step, the adaptation data are prepared via the Viterbi segmentation of the sentence into phone segments according to the recognized word string. The adapted phone models are then used to recognize the next sentence utterance. When delaying the decoding output is permissible, at the completion of a session, the recorded sentences are reused to iteratively adapt the phone model parameters and to recognize the sentences until the recognition results are converged.

2.1. Acoustic Normalization

Assume that the phone model parameters of a standard speaker are estimated from a speaker-independent training set, where the phone models are unimodal Gaussian densities $\mathcal{N}(\mu_i, C_i)$, $i = 1, 2, \cdots, I$. For a speaker q, the spectral sequence of a sentence is denoted as $x^{(q)} = \left(x_t^{(q)}, t = 1, \cdots, T^{(q)}\right)$. In the EM algorithm, the spectral vectors $x_t^{(q)}$ are the observable data; their phone labels i_t are the unobservable data; the complete data set is $\left(x_1^{(q)}, x_2^{(q)}, \cdots, x_T^{(q)}, i_1, i_2, \cdots, i_T\right)$. The EM formulation for the spectral bias $h^{(q)}$ is derived as

$$\hat{h}_{n+1}^{(q)} = \arg\max_{h^{(q)}} \sum_{t=1}^{T^{(q)}} \sum_{i=1}^{I} [P\left(i_t = i | x_t, \hat{h}_n^{(q)}\right) \cdot \log f\left(x_t^{(q)} | i_t = i, h^{(q)}\right) P(i_t = i)]$$

with

$$f\left(x_t^{(q)} | i_t = i, h^{(q)}\right) \sim \mathcal{N}\left(\mu_i + h^{(q)}, C_i\right),\ P(i_t = i) \sim \frac{N_i}{N} \quad (1)$$

0-7803-2431-5/95 $4.00 © 1995 IEEE

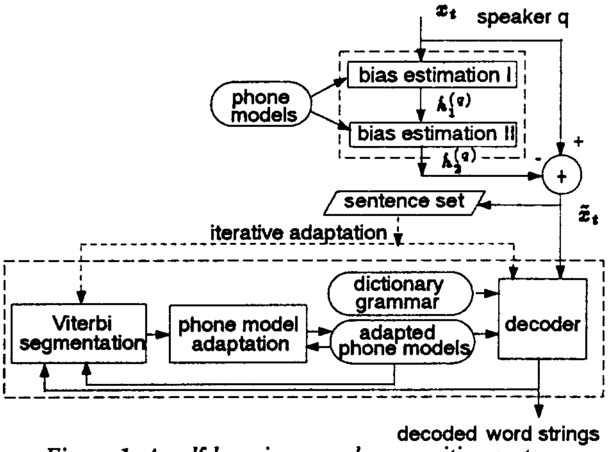

Figure 1 A self-learning speech recognition system

$$P\left(i_t = i | x_t^{(q)}, \hat{h}_n^{(q)}\right) = \frac{f\left(x_t^{(q)} | i_t = i, \hat{h}_n^{(q)}\right) P(i_t = i)}{\sum_{i=1}^{I} f\left(x_t^{(q)} | i_t = i, \hat{h}_n^{(q)}\right) P(i_t = i)}$$

where N_i is the sample size of the ith phone unit, and $N = \sum_{i=1}^{I} N_i$.

The initial bias $\hat{h}_0^{(q)}$ is unknown at the first iteration $n = 0$. Under good recognition conditions, it is reasonable to assume that $\hat{h}_0^{(q)} \approx 0$, i.e. the spectral bias is small. Under severely mismatched recording conditions, the spectral bias could be rather large such that starting with $\hat{h}_0^{(q)} = 0$ could trap the EM into a local maximum point which is far from the true bias. Taking into account the unknown distortion $\hat{h}_0^{(q)}$ under mismatched recording conditions, the likelihoods $f\left(x_t^{(q)} | i_t = i, \hat{h}_0^{(q)}\right)$ are computed from a uniform distribution, which lead to

$$P\left(i_t = i | x_t^{(q)}, \hat{h}_0^{(q)}\right) = P(i_t = i) = \frac{N_i}{N}.$$

The estimate $\hat{h}_1^{(q)}$ is derived as

$$\hat{h}_1^{(q)} = \frac{1}{T^{(q)}} \sum_{t=1}^{T^{(q)}} x_t^{(q)} - \sum_{i=1}^{I} \frac{N_i}{N} \mu_i = \bar{x}^{(q)} - \bar{\mu} \qquad (2)$$

where $\bar{x}^{(q)}$ and $\bar{\mu}$ are the mean spectra of the sentence utterance and the entire training set, respectively. In Fig. 1, the estimation of $\hat{h}_1^{(q)}$ is shown as "bias estimation I."

For $n \geq 1$, the likelihood functions $f\left(x_t^{(q)} | i_t = i, \hat{h}_n^{(q)}\right)$ are computed from the Gaussian densities as in Eq.(1). If the posterior probabilities $P\left(i_t = i | x_t^{(q)}, \hat{h}_n^{(q)}\right)$ are each approximated by the decision operation $i_n^*(t) = \arg\max_i P\left(i_t = i | x_t^{(q)}, \hat{h}_n^{(q)}\right)$ and the covariance matrices are the unit matrix, the spectral bias becomes

$$\hat{h}_{n+1}^{(q)} = \frac{1}{T^{(q)}} \sum_{t=1}^{T^{(q)}} \left(x_t^{(q)} - \mu_{i_n^*(t)}\right), \qquad (3)$$

where $\mu_{i_n^*(t)}$ is the mean vector of the phone unit labeled for x_t. In order to save computation, only one iteration is carried out using Eq.(3), which yields the refined bias estimate $\hat{h}_2^{(q)}$. In Fig. 1, the estimation of $\hat{h}_2^{(q)}$ is shown as the "bias estimation II." Acoustic normalization is defined as $\tilde{x}_t = x_t - \hat{h}_2^{(q)}$, $\forall t$.

2.2. Phone Model Adaptation

Considering a size-M Gaussian mixture density, the mean vectors and the covariance matrices of the component densities are denoted by $\theta_i = (\mu_i, C_i)$, $\forall i$. The mixture weights are $\alpha_i \geq 0$ and $\sum_{i=1}^{M} \alpha_i = 1$. The likelihood of a spectral vector x_t (the notation $\tilde{x}_t^{(q)}$ is dropped for simplicity) is $\sum_{i=1}^{M} \alpha_i f(x_t|\theta_i)$, with $f(x_t|\theta_i) \sim \mathcal{N}(\mu_i, C_i)$, $\forall i$. The prior distributions of θ_i's are assumed to be i.i.d., and the mixture weights α_i's are taken as constant. The prior mean and the prior covariance, $\mu_o^{(i)}$ and $C_o^{(i)}$ are the speaker-independent estimates from a training set with a sample size n_i. Defining the precision matrix $r_i = C_i^{-1}$, the joint distribution of (μ_i, r_i) is taken as a conjugate prior distribution. Based on the EM algorithm, the Bayesian estimation on the parameters of the Gaussian mixture density is

$$\hat{\mu}_i^{(n+1)} = \left(1 - \lambda_i^{(n)}\right)\mu_o^{(i)} + \lambda_i^{(n)} \mu_x^{(i)(n)}$$
$$\hat{C}_i^{(n+1)} \approx \left(1 - \lambda_i^{(n)}\right) C_o^{(i)} + \lambda_i^{(n)} C_x^{(i)(n)}$$
$$+ \lambda_i^{(n)} \left(1 - \lambda_i^{(n)}\right) \left(\mu_x^{(i)(n)} - \mu_o^{(i)}\right) \left(\mu_x^{(i)(n)} - \mu_o^{(i)}\right)',$$

where $\lambda_i^{(n)}$ is the interpolation parameter, $\mu_x^{(i)(n)}$ and $C_x^{(i)(n)}$ are the sample mean and the sample covariance of the adaptation data. Define $\gamma_{t,i}^{(n)} = \alpha_i f\left(x_t|\theta_i^{(n)}\right) / \sum_{j=1}^{M} \alpha_j f\left(x_t|\theta_j^{(n)}\right)$, then

$$\lambda_i^{(n)} = \sum_{t=1}^{T} \gamma_{t,i}^{(n)} / \left(\sum_{t=1}^{T} \gamma_{t,i}^{(n)} + n_i\right),$$
$$\mu_x^{(i)(n)} = \sum_{t=1}^{T} \gamma_{t,i}^{(n)} x_t / \sum_{t=1}^{T} \gamma_{t,i}^{(n)},$$
$$C_x^{(i)(n)} = \sum_{t=1}^{T} \gamma_{t,i}^{(n)} \left(x_t - \mu_x^{(i)(n)}\right) \left(x_t - \mu_x^{(i)(n)}\right)' / \sum_{t=1}^{T} \gamma_{t,i}^{(n)}.$$

2.3. Iterative Phone Model Adaptation

Denote the recorded sentences in one session as $\{S_k, k = 1, 2, \cdots, K\}$, and the decoded word strings as $\left\{A_k^{(1)}, k = 1, 2, \cdots, K\right\}$, with k the sequential index on the sentences. In the sequential adaptation, the phone models that are used to recognize the jth sentence are adapted from the previous sentences $\{S_1, S_2, \cdots, S_{j-1}\}$, and the Viterbi segmentation is guided by the word strings $\left\{A_1^{(1)}, A_2^{(1)}, \cdots, A_{j-1}^{(1)}\right\}$. In the iterative phone model adaptation, the word strings that are generated in the first iteration are denoted as $A_k^{(2)}$, and those generated in the nth iteration as $A_k^{(n+1)}$. When recognizing the jth sentence, the phone models are adapted from all the available sentences excluding the jth sentence: $\{S_1, \cdots, S_{j-1}, S_{j+1}, \cdots, S_K\}$, and in the first iteration the Viterbi segmentation is guided by the word strings $\left\{A_1^{(2)}, \cdots, A_{j-1}^{(2)}, A_{j+1}^{(1)}, \cdots, A_K^{(1)}\right\}$, for $j = 1, 2, \cdots, K$. As such, the sentences that were uttered before the jth sentence are segmented according to the newly decoded word strings, and those after the jth sentence are segmented according to the word strings generated from the sequential adaptation. To generalize, at the nth iteration, when adapting the phone models for recognizing the jth sentence, the Viterbi segmentation is guided by the word strings $\left\{A_1^{(n+1)}, \cdots, A_{j-1}^{(n+1)}, A_{j+1}^{(n)}, \cdots, A_K^{(n)}\right\}$.

Table 1. Recognition word accuracy using different adaptation methods

	baseline	acu0	PAU-I (acu0)	PAS-I (acu0)	acu1	acu2	PAU-I(acu2)	PAU-II(acu2)	PAS-I(acu2)
F-CT	74.6	83.0	86.1	86.1	81.1	81.9	84.0	-	84.3
M-CT	56.5	77.6	85.2	85.7	82.8	84.5	86.9	89.4	87.2
F-RD	44.2	64.4	72.8	76.3	69.9	70.7	78.2	81.0	80.6
M-RD	-0.3	54.5	73.3	78.1	75.1	77.9	83.1	85.2	84.3

3. EXPERIMENTS

The baseline speaker-independent HMM phone models were trained from 717 sentences by 325 speakers in the TIMIT database. The cepstrum coefficients of the PLP analysis (8th order), log energy, and their first-order 50–msec temporal regression coefficients were taken as features. The task vocabulary size was 853, and the grammar perplexity was 105. The test set has one female (F) and one male (M), each reading 198 sentences, and the speech were simultaneously recorded using two microphones: a close-talking microphone (CT) and a dynamic microphone (RD). Compared to the TIMIT data, the test data were collected under a higher level of ambient noise (approximately 15 dB higher). Recognition performances were evaluated for the following cases:

1. speaker-independent continuous speech recognition (baseline)
2. acoustic normalization assuming $\hat{h}_0^{(q)} = 0$ and using only the second bias-estimation block in Fig. 1 (acu0)
3. unsupervised phone model adaptation after acu0 (PAU-I (acu0))
4. supervised phone model adaptation after acu0 (PAS-I (acu0))
5. acoustic normalization using only $\hat{h}_1^{(q)}$ (acu1)
6. acoustic normalization using $\hat{h}_2^{(q)}$ (acu2)
7. unsupervised phone model adaptation after acu2 (PAU-I(acu2))
8. iterative unsupervised phone model adaptation after acu2 (PAU-II(acu2))
9. supervised phone model adaptation after acu2 (PAS-I(acu2)).

The average recognition word accuracies are summarized in Table 1 for each case. The recognition word accuracy as a function of the on-line speech data are further illustrated in Figures 2 and 3 for the two speakers, where (a) and (b) show the results on the CT data, and (c) and (d) show the results on the RD data. In (a) and (c) of Figures 2 and 3, the word accuracies were averaged over the number of sentences indicated on the horizontal axis (cumulative-averaged results), i.e. if $x = 60$, the result was averaged over the past 60 sentences. In (b) and (d) of the two figures, the word accuracies were averaged within each 20–sentence blocks (interval-averaged results), i.e. for $x = 60$, the result was averaged over the sentences 41 through 60. In these plots, the squares □ mark the baseline results; the diamonds ◇ mark the results from the acu2; the upward triangles △ mark the results from the PAU-I(acu2); the downward triangles ▽ mark the results from the PAU-II(acu2) (one iteration); the circles ○ mark the results from the PAS-I(acu2).

Acoustic Normalization The acoustic normalization improved the recognition accuracy significantly under the mismatched recording conditions. Under the best baseline condition F-CT (74.6%), the acu0 yielded the best result; for the rest cases, the acu1 yielded better results, and the acu2 improved acu1 for all the cases. Especially noticeable is the case M-RD: the baseline result was below 0% because the speaker put the dynamic microphone too close to his mouth which caused a significant boost to the low frequency components of the speech. In this case, the acu0 is inferior to acu1 because the posterior labeling probabilities based on the Gaussian models in Eq.(1) were very unreliable.

Phone Model Adaptation The phone model adaptations further improved the recognition accuracy: the trend is that the longer the learning period, the larger the gain of accuracy. If the amount of adaptation data is very small, the unsupervised sequential phone model adaptation could degrade the decoding accuracy. The unsupervised adaptation yielded similar improvements as the supervised adaptation if the recognition accuracy after the acoustic normalization is high. The main effect of the iterative adaptation is on the beginning sentences where the amount of adaptation data is too small for the sequential adaptation. The interval-averaged results indicate that the relative effects between the acoustic normalization and the phone model adaptation varied considerably depending on the subsets of sentences under evaluation; whereas the cumulative-averaged results indicate the general trends for the individual methods.

4. SUMMARY

The proposed self-learning adaptation technique has led to significant performance improvements for a speaker-independent continuous speech recognition system. The current work demonstrates that an *accurate* estimate of the distortion spectral bias is important for speech recognition under mismatched conditions. The proposed two-step bias-estimation method can be considered as an improvement to previous techniques of the mean-spectrum subtraction and the unsupervised maximum likelihood estimation. The unsupervised sequential phone model adaptation improves the decoding accuracy when the amount of speech data is sufficient. Iterative phone model adaptation further improves recognition accuracy when delaying the decoding output is permissible. Further study will be conducted on speech data emphasizing the speaker-induced spectral variations such as strong dialect/foreign accents or peculiar voice characteristics which account for various recognition conditions.

REFERENCES

[1]. Y. Zhao, "An Acoustic-Phonetic based Speaker Adaptation Technique for Improving Speaker-Independent Continuous Speech Recognition," *IEEE Trans. on Speech and Audio Processing,* Vol. 2, No. 3, pp. 380–384, July, 1994.

[2]. Y. Zhao, "Self-Learning Speaker Adaptation based on Spectral Variation Source Decomposition," *Proc. EuroSpeech 93,* Berlin, Germany, Sept. 1993, pp. 359–362.

[3]. Y. Zhao, "A Speaker-Independent Continuous Speech Recognition System Using Continuous Mixture Gaussian Density HMM of Phoneme-sized Units," *IEEE Trans. on Speech and Audio Processing,* Vol. 1, No. 3, pp. 345–361.

[4]. S. J. Cox and J. S. Bridle, , "Unsupervised Speaker Adaptation by Probabilistic Fitting," *Proc. ICASSP,* pp. 294–297, Glasgow, Scotland, May, 1989.

[5]. J.-L. Gauvain and C-H Lee, "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Communication,* Vol. 11, Nos 2–3, pp. 205–213, June, 1992.

[6]. D. B. Paul and B. F. Necioglu (1993), "The Lincoln large-vocabulary stack-decoder HMM CSR," *Proc. ICASSP,* Vol. II, Minneapolis, MN, pp. 660–664, Apr. 1993.

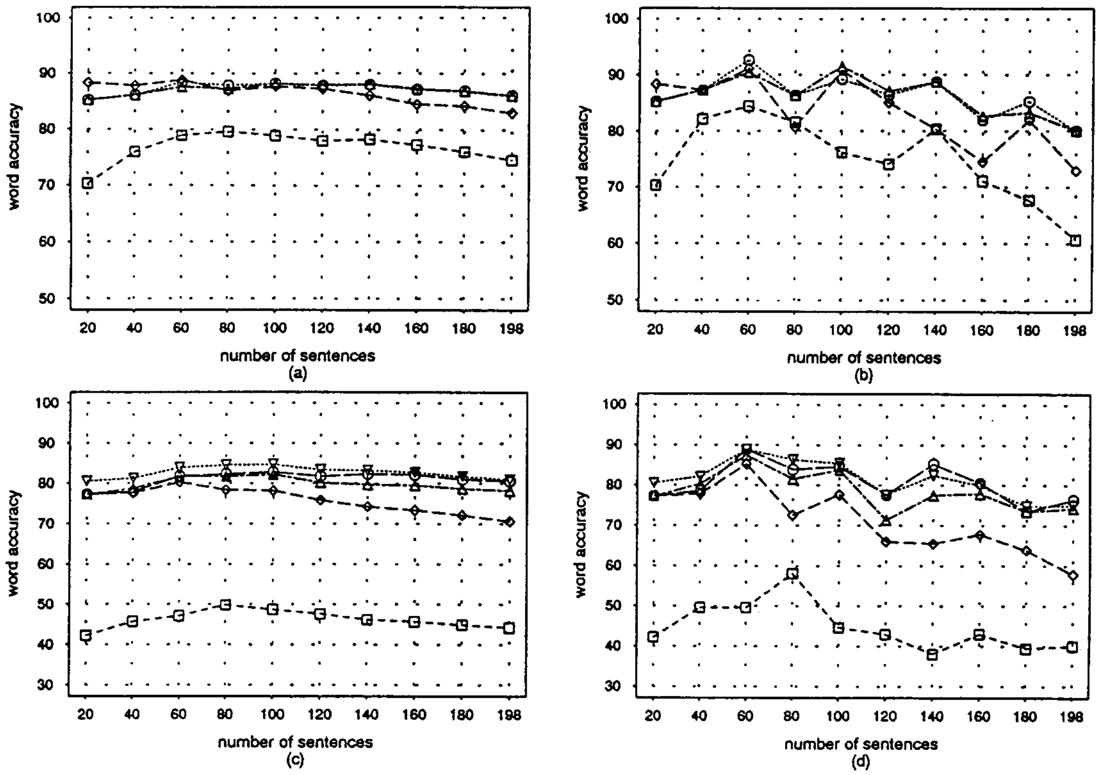

Figure 2. Experimental results on the female speaker:
(a) CT, cumulative average; (b) CT, interval average; (c) RD, cumulative average; (d) RD, interval average.

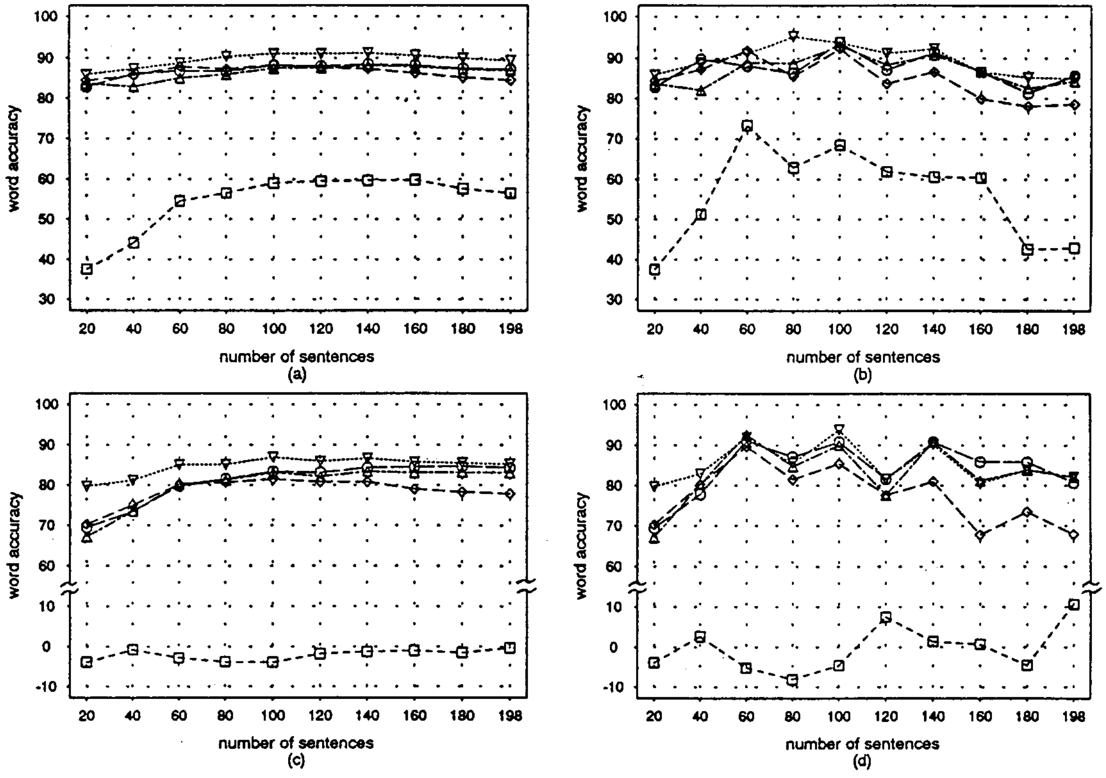

Figure 3. Experimental results on the male speaker:
(a) CT, cumulative average; (b) CT, interval average; (c) RD, cumulative average; (d) RD, interval average.