Use of vowel duration information in a large vocabulary word recognizer

L. Deng, M. Lennig,[a] and P. Mermelstein[a]
INRS-Télécommunications, 3 Place du Commerce, Montreal, Quebec H3E 1H6, Canada

(Received 6 July 1988; accepted for publication 21 March 1989)

In this paper, a method is developed to employ vowel duration properties in a hidden Markov model (HMM)-based large vocabulary speaker trained recognition system. It is found that each of the vowel phonemes spoken in isolated words can be divided into three allophones, each corresponding to a largely distinctive range of vowel durations. Such a division is based upon the phonetic context where the vowel occurs. In order to incorporate the durational information, each vowel's HMM is trained using a maximum-likelihood method with three separate sets of transition probabilities, corresponding to the three allophones. The output distributions of the HMM are assumed to be the same for all three allophones and trained jointly, to make best use of the limited number of available training tokens. The duration-specific HMMs for vowel allophones have been evaluated in isolated word recognition experiments for two male speakers. The results show that the performance of the recognizer is improved, reducing the error rate by approximately 14% compared with recognition results without the use of the vowel durational models. The performance improvement resulting from use of the vowel durational models is due to reduction of postvocalic consonant errors arising from their contextual correlation with vowels of different durations, as well as to improved discrimination between vowel phonemes.

PACS numbers: 43.72.Ne, 43.70.Fq

INTRODUCTION

It is well known that vowel duration is context dependent (Lehiste, 1970; Klatt, 1976). However, in attempts to recognize speech using hidden Markov models (HMMs), this fact has not been exploited. The objective of this paper is to demonstrate the usefulness of context-dependent durational allophones of vowels in a large vocabulary speech recognition system. Use of the fine-grained acoustic-phonetic information is particularly important as the size of the vocabulary, and thus potential confusability among words, grows.

The recognition system in which we evaluate the effectiveness of durational allophones for the vowels has a vocabulary of 60 000 English words. The system is trained for each new speaker by asking the speaker to read out loud a text consisting of about 1000 isolated word tokens. About 1000 word tokens are sufficient to train the 60 000-word recognizer because each vocabulary item is modeled as a sequence of phonemes. Prior to the use of durational allophone models for vowels reported in this paper, the top choice word accuracy for our best speaker using a uniform language model (i.e., all words are considered *a priori* equally probable) was 84% for natural text and 64% for a separate list of words purposely chosen to be confusable. This result was obtained by using one HMM to represent each phoneme, except for /l/ and /r/, which were represented by two HMMs each, and using mel-frequency cepstral coefficients (Davis and Mermelstein, 1981) and their differences over time as feature parameters.

In order to improve the performance of the phoneme-based HMM recognizer at the acoustic level, we explore the use of the durational properties of context-dependent vowel allophones in their HMM representations. The standard HMM of a phoneme consists of a set of output distributions that tend to model spectral characteristics and of a set of transition probabilities that model temporal characteristics of the phoneme (Jelinek, 1976; Bahl *et al.*, 1983). To model durational allophones of a given vowel phoneme, our approach is to provide a distinct set of transition probabilities for each duration-sensitive allophone but to retain the same output distributions. This yields allophone models of a vowel that have similar spectral characteristics, and thus retain robustness of parameter estimations for output distributions, but that differ in expected duration.

This context-dependent durational modeling approach differs fundamentally from other attempts to incorporate duration information into HMMs. All work on duration in the HMM framework has focused on the following deficiency: The expected length of time spent in a particular state in a standard HMM decreases exponentially. When states correspond to phonetic units, which is often the case when a whole word is represented by a single HMM, this aspect of the HMM is unrealistic. In fact, the duration of phonetic units tends to be distributed in a gamma-like fashion (Crystal and House, 1982). This has led Russell and Moore (1985), Levinson (1986), Russell and Cook (1987), and Codogno and Fissore (1987) to propose and experiment with the variable-duration HMM (also called the semi-HMM), which allows probability distributions of state occupancy durations to be modeled explicitly.

However, when an HMM is used to represent a vowel, as is the main concern of this paper, the states do not usually represent individual phonetic events. The overall vowel du-

[a] Also with Bell-Northern Research, Montreal, Canada.

 0001-4966/89/080540-09$00.80

ration as modeled by the HMM can be shown to possess a gamma-like, rather than an exponential, distribution due to the concatenation of states. Thus the form of the overall duration distribution in the standard HMM for a vowel does not present a serious problem, as it does in the HMM for a word. Therefore, the variable-duration HMM does not offer any particular advantage over the standard HMM in terms of modeling the distribution of the *vowel duration as a whole.*

The variable-duration HMMs proposed so far (Russell and Moore, 1985; Levinson, 1986) have not taken account of known systematic regularities in speech segment duration variations. Except for specifically chosen test data sets, semi-Markov modeling has not been shown to outperform the standard HMM in general cases. In contrast, the approach proposed here directly utilizes knowledge about how vowel duration variation is conditioned by phonetic contexts, thus modeling the vowel duration information more closely than all previous context-independent HMMs, either standard HMMs or semi-HMMs. It is due to this more precise duration modeling that the models proposed here outperform standard HMMs consistently for the data we have tested, as will be shown in this paper.

This paper is organized as follows. In Sec. I, to provide a background for incorporating vowel duration into the HMMs, we describe the context-dependent characteristics of vowel duration in isolated words as observed in the training data set. Section II shows how these durational characteristics can be incorporated into HMMs and describes the training algorithm for HMMs having multiple sets of transition probabilities. Section III presents the results of isolated word recognition experiments using vowel durational models and demonstrates improvements in recognition accuracy. Finally, in Sec. IV, we summarize and discuss our findings and results.

I. CONTEXT-DEPENDENT CHARACTERISTICS OF VOWEL DURATION

It is well known that the number of syllables in a word is the major factor influencing vowel duration (Crystal and House, 1982; Ladefoged, 1982). For isolated words, vowel duration decreases as the number of syllables in the word increases (Barnwell, 1971). For words equally long in number of syllables, vowel duration is influenced by phonemes that follow the vowel. The vowel is lengthened in syllables closed by voiced consonants relative to syllables closed by voiceless consonants (House and Fairbanks, 1953; Raphael *et al.*, 1975; Ladefoged, 1982). In turn, the duration of the vowel is a major cue for distinguishing between voiced and voiceless stop consonants in the vowel–consonant context (Zue, 1985).

We examined the dependence of vowel duration on the above contextual factors in our training data set comprising 1102 isolated words spoken by a native American English speaker. About 80% of words in the training set are derived from sentences read from texts selected randomly from magazines, books, newspapers, and office correspondence. The remaining words were chosen to contain phonemes in a variety of consonant clusters and CVC contexts. Each word in the training set is automatically segmented into a sequence of phoneme-sized units. The units are expressed in the surface form, derived from the baseform after applying a set of speaker-dependent phonological rules. The rules for vowels, determined by a phonetician, concern mainly the vowel reduction, the merge of /ɑ/ and /ɔ/, and the determination of lax or tense /æ/. The segmentation method uses the Viterbi algorithm (Jelinek, 1976), which aligns a word to a sequence of phone-sized HMMs obtained from a small amount of hand-segmented training data. The Viterbi-segmented phone boundaries of each word were carefully checked using spectrograms and were manually adjusted when necessary.

To minimize complexity and to make best use of the limited number of word tokens available, we attempted to capture only the most distinct durational differences. First, we grouped all the words with the number of syllables greater than one as *polysyllabic* words. Second, vowels (except schwa) in monosyllabic words, which are closed by a voiced consonant, and those in open syllables were grouped as a single category, since the durations for both are signifi-

TABLE I. Durational statistics of vowels in training data (1102 words). The significant level of the differences in the duration means for the three allophones is tested by the t statistic.

	Voiced coda or open syllable in monosyllabic words			Voiceless coda in monosyllabic words			In polysyllabic words		
Vowel	mean (ms)	s.d. (ms)	No. of tokens	mean (ms)	s.d. (ms)	No. of tokens	mean (ms)	s.d. (ms)	No. of tokens
aj	328	60	57	216	25	32	202	47	54
e	306	58	21	213	18	13	176	34	23
aw	294	61	7	260	1	2	252	35	35
ɔj	390	29	6	255	25	2	276	69	8
i	256	60	23	164	25	7	137	61	142
ɪ	218	59	46	156	29	19	104	41	172
ɛ	214	36	21	185	24	15	133	38	63
æ	303	51	39	234	22	19	159	36	61
ɑ	311	47	8	234	29	13	178	33	87
ʌ	266	53	81	174	25	19	129	38	28
u	294	81	42	186	28	8	168	46	20
ʊ	193	26	6	157	21	10	121	37	10
o	308	68	29	203	24	14	193	57	33
ɔ	262	54	20	203	9	3	167	44	46

cantly greater than the durations of vowels closed by a voiceless consonant. In short, the three phonological environments that determine the three allophones of each vowel are as follows: (1) in monosyllabic words with a voiced coda or in open syllables, (2) in monosyllabic words with a voiceless coda, and (3) in polysyllabic words. For each allophone, the mean duration, standard deviation, and the number of training tokens available are shown in Table I. It is apparent from Table I that, for nearly all the vowels, the three allophones possess systematically different durations. Statistical t tests (Freund and Walpole, 1980) show that most of these differences in duration are highly significant (generally, $p < 0.01$).

The vowel schwa /ə/ is an exception to the above durational pattern. Most isolated words tend to be stressed, preventing the production of schwas in monosyllabic words. For example, the word *the* has a phonetic transcription /ðʌ/ instead of /ðə/, where /ʌ/ represents the stressed midcentral vowel /ə/. For schwas in polysyllabic words, however, significant durational variations are still evident. For 103 schwa tokens in bisyllabic words closed by a voiced coda or in open syllables, the duration is 107 ± 30 ms (mean and standard deviation), while for 20 schwa tokens in bisyllabic words closed by a voiceless coda, the duration is 92 ± 20 ms, which is significantly shorter ($t = 2.142, p < 0.025$). The duration of 159 schwa tokens in words with more than two syllables is 67 ± 23 ms, which is shorter still ($t = 4.925$, $p < 0.005$).

Figures 1 and 2 show two examples of spectrograms of words with different durations of the same vowel occurring in three different phonetic contexts. Figure 1 shows spectrograms of words *code*, *coat*, and *coating*, on the same time scale. The duration of vowel /o/ in word *code* (monosyllabic word with a voiced coda) can be seen to be longer than that in word *coat* (monosyllabic word with a voiceless coda), which in turn is longer than that in word *coating* (polysyllabic word). Similar observations for vowel /ɪ/ as in words *pig*, *pick*, and *piggin* are illustrated in Fig. 2.

II. INCORPORATION OF VOWEL DURATION INTO THE RECOGNIZER

A. Employing durational features in Markov models

The vowel duration statistics described in the last section suggest that each vowel (except schwa) can be divided into three allophones according to the phonetic context where the vowel occurs, each allophone reflecting a distinct distribution of vowel durations. The phonetic contexts that define the three allophones are, in order of increasing vowel duration, (1) vowels in polysyllabic words, (2) vowels in monosyllabic words with a voiceless coda, and (3) vowels in monosyllabic words with a voiced coda or in open syllables. Although the duration of a vowel can still differ appreciably

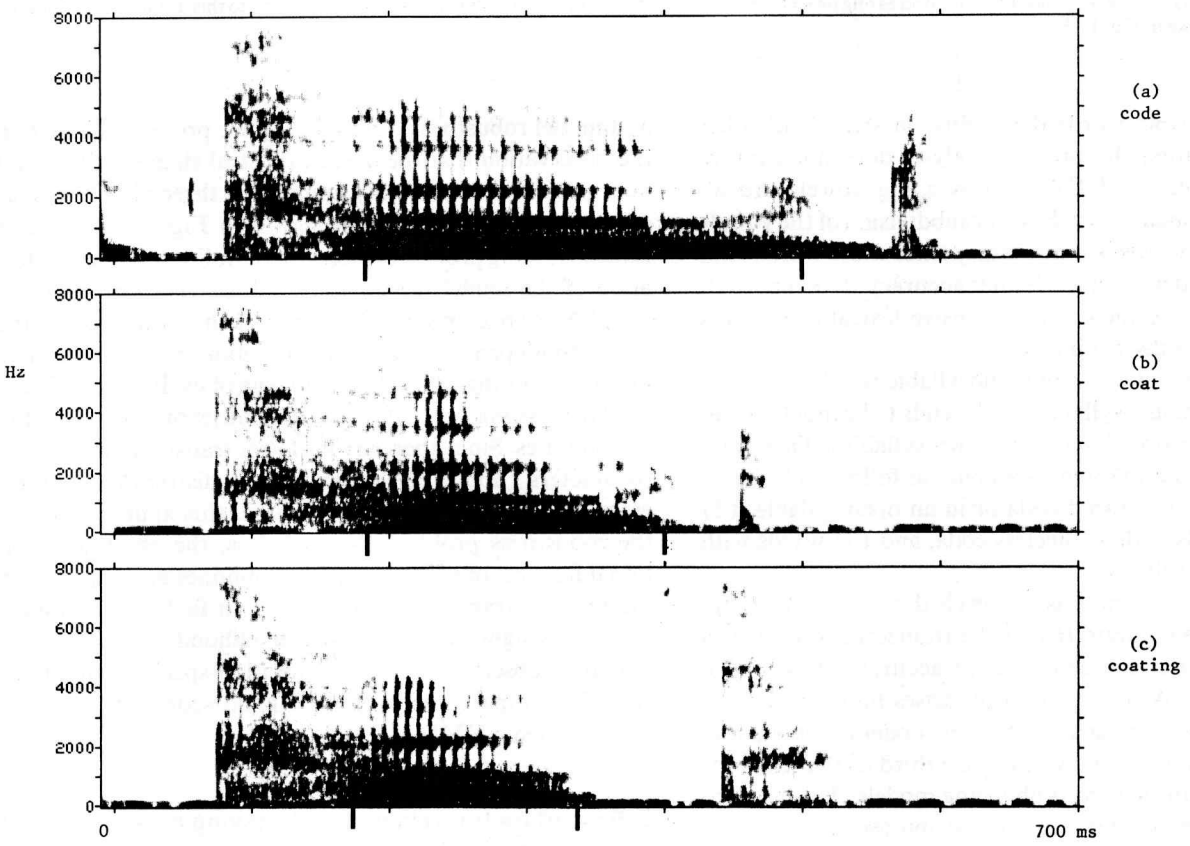

FIG. 1. Wideband spectrograms of words (a) *code*, (b) *coat*, and (c) *coating*, respectively, all with the time scale from 0 to 700 ms and with the frequency scale from 0 to 8000 Hz. The vertical bars under the abscissas delimit the vowels under consideration. Note that the vowel/o/ allophone in the monosyllabic word with an (a) voiced coda is the longest, the allophone in the monosyllabic word with a (b) voiceless coda is somewhat shorter, and the allophone in the polysyllabic word (c) is the shortest.

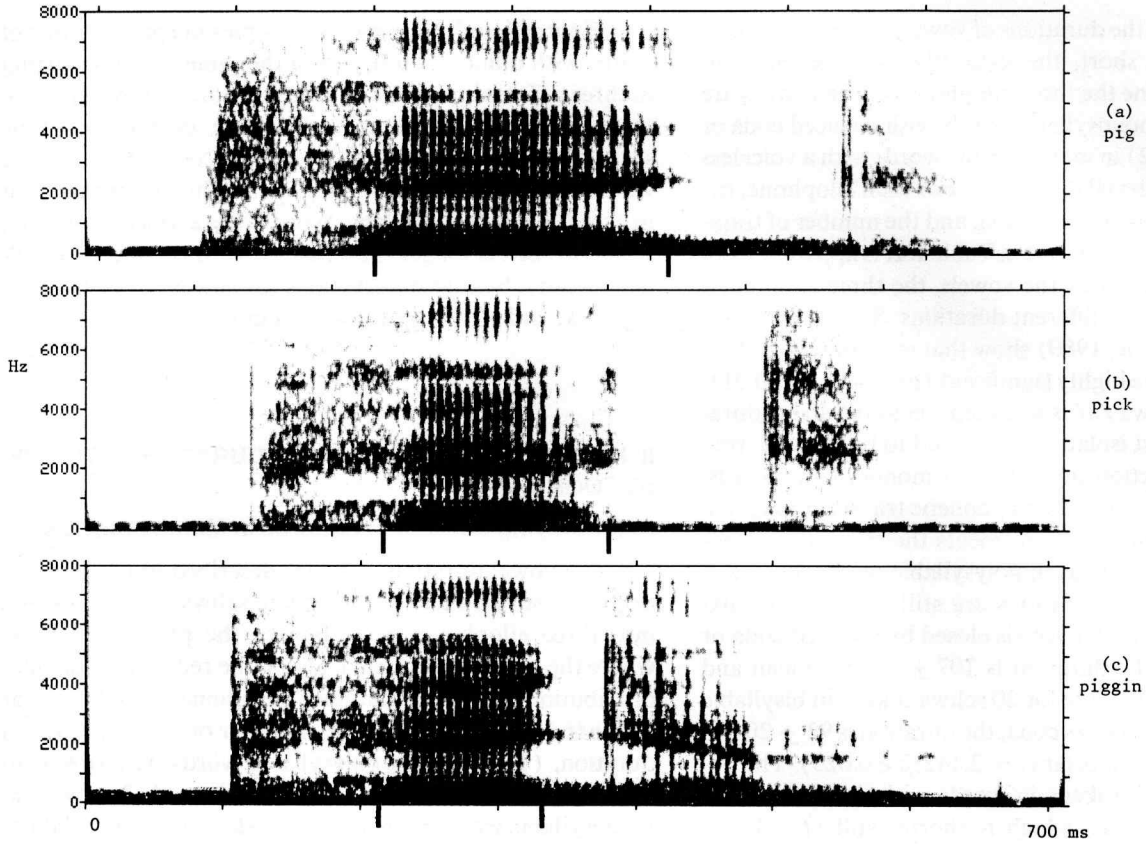

FIG. 2. Wideband spectrograms of words (a) *pig*, (b) *pick*, and (c) *piggin*, respectively, with the same time and frequency scales as those in Fig. 1. Three allophones of vowel /ɪ/ have their durations in a decreasing order from (a)–(c). The phonological environments corresponding to this durational variation are the same as those in Fig. 1.

among polysyllabic words depending on stress and other contextual factors, the present analysis does not capture such distinctions, and the corresponding vowels are all grouped into one category. Further subdivision of the vowels in polysyllabic words is unnecessary since these words tend to be recognized with much higher accuracy than monosyllabic words due to their more extensive lexical constraints and thus greater discriminability.

Schwa does not occur in monosyllabic words spoken in isolation. Schwa in bisyllabic words tends to be much longer than schwa in words longer than two syllables. Therefore, schwa is divided into three allophones as follows: (1) bisyllabic words with a voiced coda or in an open syllable, (2) bisyllabic words with a voiceless coda, and (3) words with more than two syllables.

The allophonic analysis of vowels described above suggests that three separate HMMs be trained for each vowel phoneme in order to achieve more accurate modeling of vowel duration. A serious problem arises in doing so: The amount of training data available per model for estimating HMM parameters is, on average, one third of that used for the training of the standard phoneme models. This leads to nonrobust estimates of the output distribution parameters of the HMMs. As will be described in Sec. III, our experiments show that less robust models lead to a serious degradation of the recognizer's performance.

We are faced with a trade-off between model accuracy and model robustness. The solution we propose is based on the observation that the overall spectral shape of the vowel does not differ appreciably among the three allophones, as can be seen in the examples shown in Figs. 1 and 2. The HMM training process exploits this as follows. For each iteration of the model reestimation, all tokens from the three allophones are combined to reestimate the parameters of the context-independent output distribution (multivariate Gaussian) in the HMMs. The tokens of each allophone are used to reestimate a context-dependent set of state transition probabilities. Since there are far fewer transition probability parameters (about 5%) than output distribution parameters (about 95%) to be reestimated, this approach solves the robustness problem. Nevertheless, the essential durational features of the contextual allophones are captured by the model's transition probabilities. In fact, during model training, a significant increase in likelihood scores was consistently observed for the duration-specific allophone HMMs, compared with the likelihood scores for the standard phoneme HMM.

B. Procedure for training HMMs having multiple sets of transition probabilities

This section describes the new training procedure for the HMMs that incorporates the durational features by reestimating multiple (three in the present work) sets of state

transition probabilities. The derivation of the reestimation formulas rigorously follows the general Baum–Welch reestimation algorithm (Baum, 1972; Levinson *et al.*, 1983; Liporace, 1982), as summarized below. Since a large portion of our derivation is similar to that of Liporace (1982) (for the reestimation of context-independent Gaussian HMMs), only the final reestimation formulas for the context-dependent vowel HMMs are presented here.

The Baum–Welch algorithm starts with an initial guess of the parameters of the model (left-to-right model in the present application) for multiple training tokens representing each of the allophones. The tokens are denoted as $\mathbf{O}_l^{(k)}$, $l = 1, 2$, and 3 (the allophone index) and $k = 1,2,...,K_l$ (the token index), where K_l is the total number of tokens of the lth allophone. Each token k in the allophone l, of length $T_l^{(k)}$, is a sequence of observation vectors; i.e., $\mathbf{O}_l^{(k)} = (\mathbf{o}_{l1}^{(k)}, \mathbf{o}_{l2}^{(k)}, ..., \mathbf{o}_{lT_l^{(k)}}^{(k)})$.

The reestimation algorithm is a transformation that maps the parameter space into itself based on the model. Each transformation updates the model consisting of the parameter set (A,B), where $A = [a_{ij}]$, $i,j = 1,2,...,N$ is the transition probability from state i to state j, and $B = [b_{ij}]$, $i,j = 1,2,...,N$ is the output distribution associated with the state transition from i to j. The key property of the transformation is that the updated model always has a higher score on the overall training token set than the previous iteration of the model, unless a critical point has been reached. The transformation procedure is iterated until such a critical point is reached.

The transformation of the context-dependent transition probability of the vowel HMM is as follows:

$$a_{ij}^{(l)} = \sum_{k=1}^{K_l} \sum_{t=1}^{T_l^{(k)}} \gamma_t^{k,l}(i,j) \Big/ \sum_{k=1}^{K_l} \sum_{t=1}^{T_l^{(k)}} \gamma_t^{k,l}(i), \tag{1}$$

where $l = 1,2,3$ is the index of the three separate sets of transition probabilities, and

$$\gamma_t^{k,l}(i,j) = P(\mathbf{O}_l^{(k)}, s_{t-1}^k = i, s_t^k = j | M_l) / P(\mathbf{O}_l^{(k)} | M_l), \tag{2}$$

is the conditional probability that for token k, state j is occupied at time t and state i is occupied at time $t-1$, given that the observation sequence is generated by the model. The model used to compute this conditional probability is the one relevant to the vowel duration context, denoted by M_l. Similarly,

$$\gamma_t^{k,l}(i) = P(\mathbf{O}_l^{(k)}, s_{t-1}^k = i | M_l) / P(\mathbf{O}_l^{(k)} | M_l) \tag{3}$$

is the conditional probability that for token k, state i is occupied at time $t-1$, given that the observation sequence is generated.

Note that, in Eq. (1), the lth set of transition probabilities is estimated only from the tokens belonging to the lth allophone of tokens.

The output distribution used in the present study is continuous multivariate Gaussian:

$$b_{ij}(\mathbf{O}) = N[\mathbf{O}, \Theta_{ij}, \Sigma],$$

whose parameters are the mean vectors Θ_{ij}, associated with each state transition, and the covariance matrix Σ pooled over all the state transitions of the phoneme for the robustness of its estimation.

The transformation of the mean vector corresponding to the transition from state i to state j is

$$\Theta_{ij} = \frac{\sum_{l=1}^{3} \sum_{k=1}^{K_l} \sum_{t=1}^{T_l^{(k)}} \gamma_t^{k,l}(i,j) \mathbf{O}_t}{\sum_{l=1}^{3} \sum_{k=1}^{K_l} \sum_{t=1}^{T_l^{(k)}} \gamma_t^{k,l}(i,j)}, \tag{4}$$

and the reestimate of the covariance matrix is

$$\Sigma = \frac{\sum_{l=1}^{3} \sum_{k=1}^{K_l} \sum_{t=1}^{T_l^{(k)}} \Sigma_{i,j} \gamma_t^{k,l}(i,j) (\mathbf{O}_t - \Theta_{ij})(\mathbf{O}_t - \Theta_{ij})^*}{\sum_{l=1}^{3} \sum_{k=1}^{K_l} \sum_{t=1}^{T_l^{(k)}} \Sigma_{i,j} \gamma_t^{k,l}(i,j)}. \tag{5}$$

Note that in the transformations (4) and (5) for both the mean vector and the covariance matrix, tokens of all allophones are pooled, thus effectively increasing the number of tokens available for parameter estimation.

The conditional probabilities involved in the transformations (1), (4), and (5) can be efficiently calculated by using the forward and backward probabilities (Baum, 1972; Jelinek, 1976; Bahl *et al.*, 1983; Levinson *et al.*, 1983). To simplify the notation, the superscript for the token index and the subscript for the allophone index will be dropped below.

The forward probability is

$$\begin{aligned} \alpha_t(j) &\equiv P(\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_t, s_t = j) \\ &= \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_{ij}(\mathbf{o}_t), \end{aligned} \tag{6}$$

with

$$\alpha_0(j) = \begin{cases} 1, & \text{for } j = 1, \\ 0, & \text{for } j > 1. \end{cases}$$

The backward probability is

$$\begin{aligned} \beta_t(j) &\equiv P(\mathbf{o}_{t+1}, \mathbf{o}_{t+2}, ..., \mathbf{o}_T | s_t = j) \\ &= \sum_{i=1}^{N} \beta_{t+1}(i) a_{ji} b_{ji}(\mathbf{o}_{t+1}), \end{aligned} \tag{7}$$

with

$$\beta_T(j) = \begin{cases} 1, & \text{for } j = N, \\ 0, & \text{for } j < N. \end{cases}$$

Once the forward and backward probabilities are computed, the probability appearing in (1), (4), and (5) that the sequence $\mathbf{O}$ is observed and that the state transition from i to j occurs from time $t-1$ to t, can be calculated by

$$P(\mathbf{O}, s_{t-1} = i, s_t = j) = \alpha_{t-1}(i) a_{ij} b_{ij}(\mathbf{o}_t) \beta_t(j), \tag{8}$$

and the probability that the sequence $\mathbf{O}$ is observed and that state i is occupied at time t is calculated by

$$P(\mathbf{O}, s_t = i) = \alpha_t(i) \beta_t(i). \tag{9}$$

The probability of the observation sequence for each token is (for the left–right HMM model)

$$P(\mathbf{O}|M) = \alpha_T(N). \tag{10}$$

In summary, during each iteration of the model reestimation, the following steps are carried out: (1) Calculate the forward and backward probabilities by Eqs. (6) and (7);

(2) calculate the state transition probability, the state occupation probability, and the probability score by Eqs. (8)–(10), and from these calculate the conditional state transition and state occupation probabilities by Eqs. (2) and (3); (3) accumulate the quantities appearing in Eqs. (1), (4), and (5) over frames, over tokens, over the state transitions (only for covariance matrix estimation), and over allophone groups (only for mean vector and covariance matrix estimation); and (4) reestimate the model parameters according to Eqs. (1), (4), and (5).

III. EXPERIMENTS ON ISOLATED WORD RECOGNITION

An overview of our large vocabulary isolated word recognizer (without the use of vowel durational models presented here) can be found in Gupta *et al.* (1988) and Deng *et al.* (1988). Briefly, the recognition process consists of word end-point detection, a fast search algorithm to generate a list of most likely word choices, and the computation of exact likelihoods for these choices. The durational allophone vowel models presented in this paper are only introduced at the exact likelihood scoring stage. Since the phonetic transcription of each candidate word is known, so is the phonetic context of the constituent vowel during the exact likelihood computation. This allows one of the three vowel allophones to be deterministically selected to score each observation.

Training and test data were recorded in a quiet sound booth. Two sets of the training and test data were used, one consisted of texts comprising natural-language sentences and the other consisted of words with consonant clusters and of consonant–vowel–consonant (CVC) or consonant–vowel–consonant–vowel (CVCV) sequences. The training data had 1102 words in total. The test set consisted of 312 special words selected to be highly confusable (most of them were CVC and CVCV) and 782 words from natural text. The same two distinct sets of training and test data were used in the two sets of experiments using durational and standard vowel HMMs.

The performance of the recognizer is evaluated using the following three criteria. The first criterion is the percentage of words correctly identified as the top word choice by the recognizer (since no language model is used, homophone confusions are not counted as errors). The second criterion is the average rank of the correct word in the ordered list of word hypotheses. Even if recognition accuracy is unchanged, a lower average rank is evidence of better likelihood estimation that may lead to better recognition accuracy when additional information (e.g., a language model) is introduced. The third criterion is the average difference between the log probability of the correct word and the largest log probability among the incorrect words. A higher average difference reflects better discriminability.

A comparison of the recognition performance using vowel HMMs with and without incorporating durational information is shown in Table II for a 782-word natural text test set and in Table III for a 312-word CVC(V) test set. Use of the vowel durational HMMs (column B), compared with use of standard HMMs (column A), consistently improves

TABLE II. Comparison of recognition performance with and without vowel durational models for natural-text data set. Column A: one HMM per vowel; column B: three HMMs per vowel with tied output distributions; column C: three HMMs per vowel with separate output distributions.

Performance measure	Natural text set (782 words)		
	A	B	C
Percent correct	84.6	86.8	79.4
Average rank	1.72	1.57	1.79
Ave. diff. of log scores	31.7	34.8	30.2

the recognizer's performance for all three evaluation criteria for both the CVC(V) and the natural text sets. For the natural text set, the error rate is reduced by 14.3%, and, for the CVC(V) set, it is reduced by 13.3%.

We should emphasize that the vowel durational models improve recognition only when the model robustness is maintained by tying the output distributions for all three allophone models. To illustrate this, an experiment was performed where the output distributions were not tied. The recognition performance, shown in column C of Tables II and III, is significantly poorer.

To gain insight into the advantages of using the new vowel durational models, we examined the word recognition errors corrected or introduced by using the vowel durational allophones. Table IV shows these data. The first column lists the words in the test data on which errors have been corrected (marked by $\surd$ in the fourth column) or new errors have been introduced (marked by $\times$). The phonetic transcriptions of the recognizer's top choice output words are listed in the second and third columns, respectively, for the recognizers without and with vowel durational models.

An examination of the errors under the different modeling conditions shown in Table IV reveals that vowel durational models provide two different types of advantages. First, incorporation of vowel duration into the HMM simply makes the model a more precise discriminator of the vowel itself. This accounts for many error corrections such as [lɑks] to [læks], [bajt] to [bʌt], and so on. Second, vowel duration provides a cue to the syllabic coda (voiced or voiceless) and to the syllable type (open or closed). Thus better modeling of the vowel not only helps avoid vowel confusions in recognition, but also improves the discrimination of consonants in the syllabic coda. Evidence for this is found in correction, from [dajz] to [dajs]. Many errors due to weakly released voiceless stops have been corrected by using vow-

TABLE III. Comparison of recognition performance with and without vowel durational models for CVC(V) test set.

Performance measure	CVC(V) set (312 words)		
	A	B	C
Percent correct	63.8	68.6	62.0
Average rank	2.58	2.30	2.80
Ave. diff. of log scores	6.2	11.0	6.1

TABLE IV. List of test words on which recognition errors were corrected or newly introduced.

Test words	Top choice using standard HMMs	Top choice using duration HMMs	Error correction (✓) or new error (×)
a	[pei] *payee*	[e]	✓
are	[awr] *our*	[ɑr]	✓
background	[bækgrawnd]	[bækənt] *bacchant*	×
beg	[bed] *bade*	[bɛg]	✓
bid	[bid] *bead*	[bɪd]	✓
bitter	[bɪtr]	[bitr] *beater*	×
but	[bajt] *bite*	[bʌt]	✓
coat	[kot]	[krot] *Croat*	×
colon	[kɑləm] *column*	[kolən]	✓
deacon	[pɪkɪn] *Pekin*	[dikən]	✓
dice	[dajz] *dyes*	[dajs]	✓
digger	[ditr] *deter*	[dɪgr]	✓
eat	[i] *E*	[it]	✓
fan	[fæn]	[fɛn] *fen*	×
future	[tɪtr] *titter*	[fjutʃr]	✓
gate	[kit] *keet*	[get]	✓
gave	[gev]	[gɪv] *give*	×
guess	[gæs] *gas*	[gɛs]	✓
have	[hɑv] *halve*	[hæv]	✓
height	[haj] *high*	[hajt]	✓
is	[iz] *ease*	[ɪz]	✓
lax	[lɑks] *lox*	[læks]	✓
lost	[lʌst] *lust*	[lɔst]	✓
me	[mi]	[ni] *knee*	×
nineteen	[bajtɪŋ] *biting*	[najntin]	✓
one	[wɑn] *wan*	[wʌn]	✓
pip	[pɛp] *pep*	[pɪp]	✓
pipe	[hajp] *hype*	[pajp]	✓
pity	[pɪni] *pinnae*	[pɪti]	✓
pride	[kajnd] *kind*	[prajd]	✓
right	[raj] *rye*	[rajt]	✓
slight	[slaj] *sly*	[slajt]	✓
so	[sto] *stow*	[so]	✓
than	[ðæn]	[ðɛn] *then*	×
then	[ðæn] *than*	[ðɛn]	✓
tick	[tek] *take*	[tɪk]	✓
told	[kɔld] *called*	[told]	✓
tried	[tʃrajəd] *triad*	[tʃrajd]	✓
was	[wajz] *wise*	[wʌz]	✓
why	[wajni] *winy*	[waj]	✓
will	[wo] *woe*	[wɪl]	✓

el durational models, due to a cue provided by the vowel duration as to whether the vowel is in an open syllable or in a syllable closed by a voiceless consonant.

A similar experiment was performed with speech of a second male speaker (Montreal dialect) for 396-word natural-text test data. The recognition results are shown in Table V. Use of the vowel durational models improved the recognition rate from 77.0% to 79.3%, a 10% reduction in the error rate. Table VI lists the recognition errors corrected and newly introduced by using the vowel durational models in the same format as Table IV for the first speaker.

TABLE V. Comparison of recognition performance with and without vowel durational models for speaker 2. Column A: one HMM per vowel; column B: three HMMs per vowel with tied output distributions; column C: three HMMs per vowel with separate output distributions.

Performance measure	Natural text set (396 words)		
	A	B	C
Percent correct	77.0	79.3	68.7
Average rank	1.70	1.65	2.30
Ave. diff. of log scores	30.5	33.0	19.0

The standard calculation of the 95% confidence interval for proportion (Freund and Walpole, 1980) shows that the improvement of recognition accuracy by using vowel durational models is statistically significant for both speakers ($p < 0.025$). Although these results need to be confirmed on data from additional speakers, they already suggest that we are extracting durational cues utilized consistently by speakers of the language.

IV. SUMMARY AND DISCUSSION

The major contribution of this paper is to establish a method to exploit certain allophonic information in the framework of phonemic Markov modeling, yet maintain robustness in the estimation of the model parameters. The allophonic information exploited is that provided by vowel du-

TABLE VI. List of test words on which recognition errors were corrected or newly introduced for speaker 2.

Test words	Top choice using standard HMMs	Top choice using duration HMMs	Error correction (√) or new error (×)
about	[əbʌt] *abut*	[əbawt]	√
and	[ɛnd] *end*	[ænd]	√
background	[rækrɛnt] *rack-rent*	[bækgrawnd]	√
capital	[kæpətl]	[kæprl] *caporal*	×
do	[dju] *dew*	[du]	√
dug	[dʌg]	[dajnd] *dined*	×
good	[pɪn] *pin*	[gud]	√
his	[pez] *pays*	[hɪz]	√
immigration	[ɪntəgreʃən] *integration*	[ɪməgreʃən]	√
new	[mju] *mew*	[nju]	√
nineteen	[majtinəs] *mightiness*	[najntin]	√
not	[nɑt]	[blɑt] *blot*	×
oar	[ɔr]	[lɔr] *lore*	×
once	[wʌnz] *ones*	[wʌns]	√
presented	[prɪzɛntətɪv] *presentative*	[prizɛntɪd]	√
sure	[ʃr] *shirr*	[ʃur]	√
swinging	[swipɪŋ] *sweeping*	[swɪŋɪŋ]	√
Venis	[vɛndrz] *vendors*	[vɛnɪs]	√
why	[wajd] *wide*	[waj]	√
what	[wajt] *white*	[wʌt]	√
with	[wɪdθ] *width*	[wɪθ]	√
year	[jɪr]	[irə] *era*	×
years	[jɪrz]	[gɪrz] *gears*	×

ration concerning the identity of the vowel as well as features of the consonantal environment. The information is extracted by allowing multiple sets of transition probabilities in each of the vowel HMMs corresponding to different durational allophones, training them in an optimal way, and accessing them appropriately at recognition time. The robustness of such models is largely maintained since most of the model parameters, those of the output distributions, are tied across allophones.

Tying of the output distributions across allophones is reasonable because the overall spectral shape, which the output distributions intend to model, appears to be similar for allophones with different durations. To a first approximation, vowel duration is independent of the spectral shape. Thus it is appropriate to train the transition probabilities (determined mainly by the duration) independently of the output distributions. The significantly poorer recognition results obtained when all model parameters are training independently for each allophone due to undertraining are avoided by this approach.

We have shown that the HMMs incorporating context-dependent vowel duration information consistently improve the performance of our large vocabulary recognizer. About 14% reduction in error rate has been achieved both for the natural text set as well as for the more difficult CVC(V) set. In fact, in the natural text set, most errors corrected by the new vowel durational models are from the monosyllabic CVC words (see Table IV). However, for most of the misrecognized polysyllabic words that did not get corrected by the vowel durational models, we still observe that the differences are reduced between their scores and the scores of the words incorrectly identified as the top choices, as is reflected in the third measure of the recognizer's performance. The performance improvement by using the vowel durational models lies not only in the fact that such models represent vowels themselves more faithfully, but also that they provide cues to the nature of the following consonants. Due to lexical constraints of the language, modeling of vowels and consonants is highly interdependent.

As can be seen from the error rates in Table III, CVC-type words present the most difficulty for our recognizer because of their potentially high phonetic confusability, or, viewed another way, their weaker lexical constraints as compared with polysyllabic words. Although the error rate reduction is similar for CVC(V) and natural text sets in our present experiments, the proportion of error correction for only *monosyllabic* words is significantly larger for the natural text set than for the CVC(V) set. This can be accounted for by the fact that our CVC(V) set was chosen to be highly confusable. Such results, on the other hand, suggest that the power of vowel durational models in discriminating the highly confusable CVC(V) words is limited. More precise (and no less robust) modeling of consonants as well as vowels is needed to further improve the performance of the recognizer.

ACKNOWLEDGMENT

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

Bahl, L. R., Jelinek, F., and Mercer, R. L. (**1983**). "A maximum likelihood approach to continuous speech recognition," IEEE Trans. Pattern Analysis Machine Intelligence **PAMI-5**, 179–190.

Barnwell, T. P. (**1971**). "An algorithm for segment duration in a reading machine context," Tech. Rep. 479, Research Lab. Electron. MIT, Cambridge, MA.

Baum, L. E. (**1972**). "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," Inequalities **3**, 1–8.

Codogno, M., and Fissore, L. (**1987**). "Duration modeling in finite state automata for speech recognition and fast speaker adaptation," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **3**, 1269–1273.

Crystal, T. H., and House, A. S. (**1982**). "Segmental durations in connected speech signals: Preliminary results," J. Acoust. Soc. Am. **72**, 705–716.

Davis, S. B., and Mermelstein, P. (**1980**). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust. Speech Signal Process. **ASSP-28**, (4), 357–365.

Deng, L., Kenny, P., Lennig, M., Gupta, V., and Mermelstein, P. (**1988**). "Large vocabulary word recognition based on phonetic representation by hidden Markov models," Proc. Can. Conf. Electrical Comp. Eng. 131–134.

Freund, J. E., and Walpole, R. E. (**1980**). *Mathematical Statistics* (Prentice-Hall, Englewood Cliffs, NJ).

Gupta, V., Lennig, M., and Mermelstein, P. (**1988**). "Fast search strategy in a large vocabulary word recognizer," J. Acoust. Soc. Am. **84**, 2007–2017.

House, A. S., and Fairbanks, G. (**1953**). "The influence of consonant environment upon the secondary acoustical characteristics of vowels," J. Acoust. Soc. Am. **25**, 105–113.

Jelinek, F. (**1976**). "Continuous speech recognition by statistical methods," Proc. IEEE **64**(4), 532–556.

Klatt, D. H. (**1976**). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. **59**, 1208–1221.

Ladefoged, P. (**1982**). *A Course in Phonetics* (Harcourt Brace Jovanovich, New York).

Lehiste, I. (**1970**). *Suprasegmentals* (MIT, Cambridge, MA).

Levinson, S. E. (**1986**). "Continuously variable duration hidden Markov models for automatic spech recognition," Comp. Speech Lang. **1**, 29–45.

Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. (**1983**). "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," Bell Syst. Tech. J. **62**, 1035–1074.

Liporace, L. A. (**1982**). "Maximum likelihood estimation for multivariate observations of Markov sources," IEEE Trans. Inf. Theory **IT-28**, 729–734.

Raphael, L. J., Dorman, M. F., and Freeman, F. (**1975**). "Vowel and nasal duration as cues to voicing in word-final stop consonants: Spectrographic and perceptual studies," J. Speech Hear. Res. **18**, (3), 389–499.

Russell, M. J., and Cook, A. E. (**1987**). "Experimental evaluation of duration modeling techniques for automatic speech recognition," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **4**, 2376–2379.

Russell, M. J., and Moore, R. K. (**1985**). "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 5–8.

Zue, V. W. (**1985**). *Speech Spectrogram Reading: Lecture Notes and Spectrograms* (MIT, Cambridge, MA).