# Modeling microsegments of stop consonants in a hidden Markov model based word recognizer[a)]

L. Deng,[b)] M. Lennig, and P. Mermelstein
*INRS-Télécommunications, Montreal, Quebec H3E 1H6, Canada*

(Received 1 May 1989; accepted for publication 17 January 1990)

The motivation of this study is the poor performance of speech recognizers on the stop consonants. To overcome this weakness, word initial and word final stop consonants are modeled at a subphonemic (microsegmental) level. Each stop consonant is segmented into a few relatively stationary microsegments: silence, voice bar, burst, and aspiration. Microsegments of certain phonemically different stops are trained together due to their similar spectral properties. Microsegmental models of burst and aspiration are conditioned on the adjacent vowel category: front versus nonfront vowels. The resulting context-dependent microsegmental hidden Markov models (HMMs) for six stops possess the desired properties for a compromise between modeling accuracy and modeling robustness. They allow the recognizer to focus discrimination onto those regions of a stop that serve to distinguish it from other stops. Use of these models in recognition experiments for word lists consisting of CVC words reduces the error rate by 35% compared with the result obtained by using one HMM for each stop phoneme.

PACS numbers: 43.72.Ne, 43.72.Ar, 43.70.Fq

## INTRODUCTION

In many large vocabulary speech recognizers, each phoneme is represented by a single hidden Markov model (HMM),[1] and words are modeled by concatenating phonemic HMMs (Bahl *et al.*, 1980; Merialdo, 1987; Lee and Hon, 1988; Murveit and Weintraub, 1988; Gupta *et al.*, 1988; Deng *et al.*, 1988).[2] This phonemic modeling approach, evaluated in a 75 000-word speaker-dependent recognition system (Gupta *et al.*, 1988; Deng *et al.*, 1988), showed a weakness in discriminating among stop consonants, especially in monosyllabic consonant–vowel–consonant (CVC) words that contain stop consonants. In order to provide a diagnostic tool for investigating the source of stop confusions, we designed a list of mostly CVC words, rich in stop consonants and containing many minimal pairs. The error rate for acoustic recognition of this word list was nearly three times as high as that for normal prose texts. This result indicates that the phonemic HMM does not adequately represent the temporally local context-dependent acoustic features that distinguish one stop phoneme from another.

The techniques of diphone and triphone modeling have been successfully used in speech recognizers with vocabulary sizes on the order of 1000 words (Schwartz *et al.*, 1984; Paul and Martin, 1988; Lee *et al.*, 1989). In larger vocabulary systems such as our 75 000-word recognizer, the number of triphones required is of the order of 15 000. The amount of speech required to train such a large number of triphones would make speaker-dependent recognizers impractical. Derouault (1987) and Deng *et al.* (1990) have attempted to reduce the number of triphone models by merging triphones according to acoustic-phonetic knowledge about expected coarticulatory effects. However, this technique has not proved useful when the amount of training data is limited to about 1000 words (Deng *et al.*, 1990). Lee *et al.* (1989) have proposed an automatic triphone clustering algorithm based on information-theoretic criteria, but their algorithm requires presence in the training data of all the triphones to be clustered, and hence is not directly applicable to our very large vocabulary recognition task. Therefore, reducing phonemic confusions in large vocabulary speech recognition requires new strategies.

The strategy that we propose in this article is to exploit the microsegmental structure of stops (Fant, 1973). The basis for this strategy is that different stop phonemes and different allophones of the same stop phoneme often share common microsegments. Since common microsegments occurring in different phonemes do not provide discriminative information, they can be represented by a single model. Discrimination then focuses on the small number of microsegments that differentiate stop phonemes. Because training data for common microsegments are pooled across tokens of different phonemes and of different allophones, model robustness is enhanced.

Moore *et al.* (1983) have proposed a technique for focusing recognition in whole-word pattern matching in a dynamic time warping based speech recognizer. For example, in discriminating the two words *stalagmite* and *stalactite*, it is desirable to eliminate confusions caused by irrelevant differences in the regions *stala-* and *-ite*. This is accomplished by constructing a network where these regions are integrated into a common path. Such discriminative focusing at the phoneme level is already inherent in phonemic recognizers. The present study is intended to extend this concept to a subphonemic level. This extension is desirable since many contrasting phoneme pairs, especially stops, differ acoustically in only a small part of their overall extents.

In Sec. I of this paper, we introduce our microsegmental

[a)] This work was supported by the Natural Sciences and Engineering Research Council of Canada.

[b)] Present address: Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

analysis of stop consonants and describe the construction of microsegmental models. Section II presents isolated word recognition results. Section III is a summary and discussion of our new stop modeling approach.

## I. MICROSEGMENTAL MODELING OF STOP CONSONANTS

### A. Inventory of microsegments for stop consonants

The stops /ptkbdg/ are produced by complex movements in the vocal tract. Acoustic analysis shows that, when a stop is adjacent to a vowel, several segments of rather distinct spectral properties can be identified (Halle *et al.*, 1957; Fant, 1973). Figure 1 is a spectrogram illustrating these microsegments for the voiced stop /b/ in the word *bib*. For the initial[3] /b/ in this token, it is easy to identify a *voice bar* microsegment, whose spectrum is dominated by low-frequency energy, and a very short stop *burst* microsegment, whose energy is more spread out over frequency. The initial *voice bar* can sometimes be weak or absent. The final stop /b/ in this example can be decomposed into *voice bar*, *burst*, and *final voiced stop aspiration* microsegments. *Voice bars* in initial and final positions tend to have similar spectral characteristics. The same can be said of stop *bursts* of a given phoneme.

For the voiceless stops, different microsegments can be identified. Figure 2 is a spectrogram of the word *pep*. For the final /p/, there is a silent closure period before the burst starts. We will call this the *silence* microsegment. Unlike initial /b/, initial /p/ has a rather long and easily identified aspiration after the burst, making the burst and aspiration look similar to those in final position.

The above observations largely hold for stops having other places of articulation as well. Each of the microsegments discussed above will be represented by a single HMM. Because microsegments are shared across different allophones and phonemes, the number of microsegmental HMMs required is not large. This is important because it will improve the robustness of the trained models. In this study, the following nine microsegments are used in various combinations to represent the six stops in initial and final positions:

(1) *voice bar*, which occurs in final /bdg/ and optionally in initial /bdg/; a two-state HMM is used.
(2) *silence*, which occurs optionally in final /ptk/; a two-state HMM is used.
(3) *final voiced stop aspiration*, which occurs optionally in final /bdg/; a three-state HMM is used.
(4) /b/ *burst*, which occurs in initial /b/ and optionally in final /b/; a two-state HMM is used.
(5) /d/ *burst*, which occurs in initial /d/ and optionally in final /d/; a two-state HMM is used.
(6) /g/ *burst*, which occurs in initial /g/ and optionally in final /g/; a two-state HMM is used.
(7) /p/ *burst* + *aspiration*, which occurs in initial /p/ and optionally in final /p/; a four-state HMM is used.
(8) /t/ *burst* + *aspiration*, which occurs in initial /t/ and optionally in final /t/; a four-state HMM is used.
(9) /k/ *burst* + *aspiration*, which occurs in initial /k/ and optionally in final /k/; a four-state HMM is used.

The *burst* and *burst* + *aspiration* microsegments (4)–(9) are referred to as the release microsegments, since they include the acoustic result of the release of intraoral pressure accumulated during the closure portion of the stop.

Medial voiceless stops in /s/-stop-vowel context, e.g., /k/ in the word *sky*, have some special microsegmental properties. As an example, a spectrogram of the word *sky* is shown in Fig. 3. Due to the presence of the preceding /s/, the stop /k/ is unaspirated. Its burst looks different from the burst + aspiration portion of initial /k/. Instead, it resembles the burst of initial /g/. Therefore, /k/ in *sky* is segmented into *silence* followed by /g/ *burst*. In general, the voiceless stop in an /s/-stop-vowel context is modeled as *silence* followed by the *burst* of the homorganic voiced stop.

While the above models represent stops in initial, final,

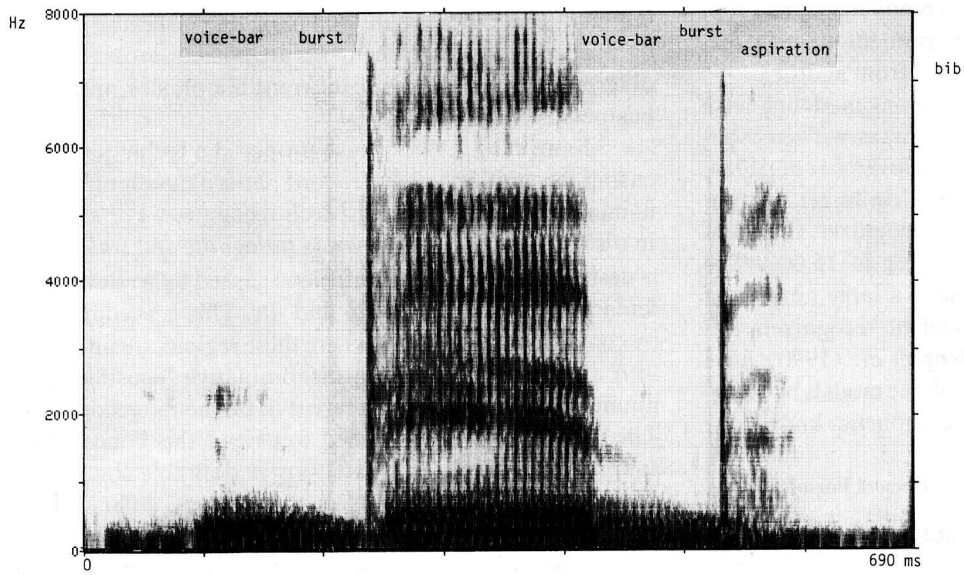

FIG. 1. Spectrogram of the word *bib* /bɪb/, illustrating *voice bar* and *burst* microsegments of initial voiced stop /b/ and *voice bar*, *burst*, and *final aspiration* microsegments of final voiced stop /b/. These two voice bars are trained jointly, as are the two bursts.

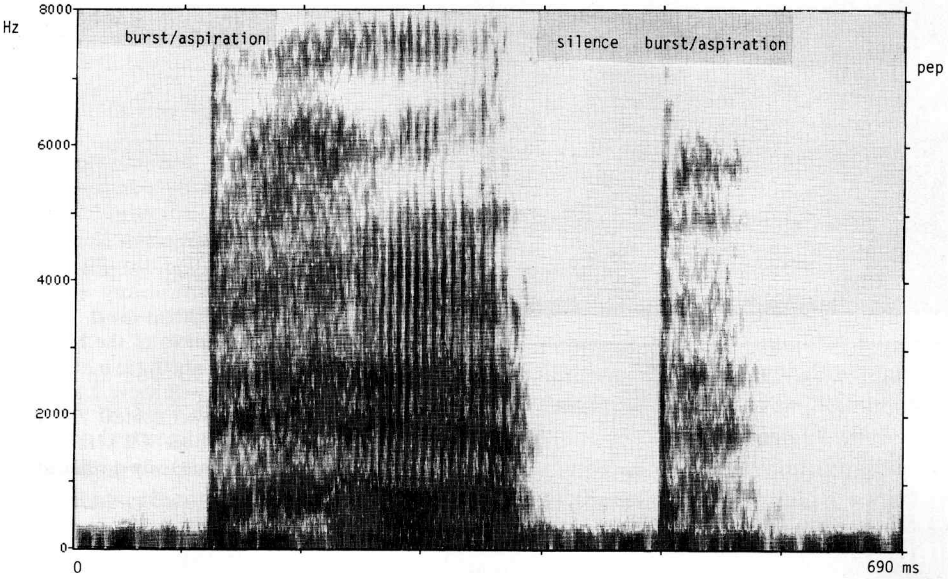

FIG. 2. Spectrogram of the word *pep* /pɛp/, illustrating the *burst* + *aspiration* microsegment of initial voiceless stop /p/, and *silence* and *burst* + *aspiration* microsegments of final voiceless stop /p/. The two *burst* + *aspiration* microsegments are trained jointly.

and /s/-stop-vowel positions, representation of other medial stops, including flaps and medial clusters, requires more complex analysis. At this preliminary stage, we have not attempted to model medial stops. Except for stops in /s/-stop-vowel position, medial stops are simply represented by one model per phoneme. That is, each medial stop is represented by one HMM trained from all the tokens of that phoneme occurring medially in the training set.

In the microsegmental modeling of stops described so far, two complications arise. First, the *voice bar* of an initial voiced stop may or may not be present in a given token. Second, final stops are sometimes unreleased. If we force the occurrence of a microsegmental model for a microsegment that is missing in the actual utterance, a low recognition score will result. To avoid this, we introduce a null transition (Jelinek, 1976) in parallel with the HMM representing the optional microsegment (*voice bar* in initial position and *burst*, *final voiced stop aspiration*, and *burst* + *aspiration* in final position). The relative weights of the null transition and of the optional microsegment are determined by the relative frequency with which the optional microsegment is observed in the training data.

### B. Context dependence of the release microsegments

The microsegmental models described above have significantly improved the discrimination of stops, as will be shown in Sec. II. Yet, further improvement is obtained by taking account of the fact that the spectral properties of certain stop microsegments are strongly affected by the adjacent vowel. Coarticulatory effects can be seen most clearly for stops in front versus nonfront vowel contexts. As a first step in context-dependent modeling, the current study examines these contexts. Figures 4–6 illustrate these coarticulatory effects through spectrographic examples.

Figure 4(a) and (b) shows spectrograms of the word *kick* /kɪk/ and *coke* /kok/, showing the different spectral effects of front /ɪ/ and back /o/ on the burst + aspiration portion of the /k/ in both initial and final positions. Depending on whether the adjacent vowel is /ɪ/ or /o/, the frequency location of the burst + aspiration energy in /k/ is signifi-

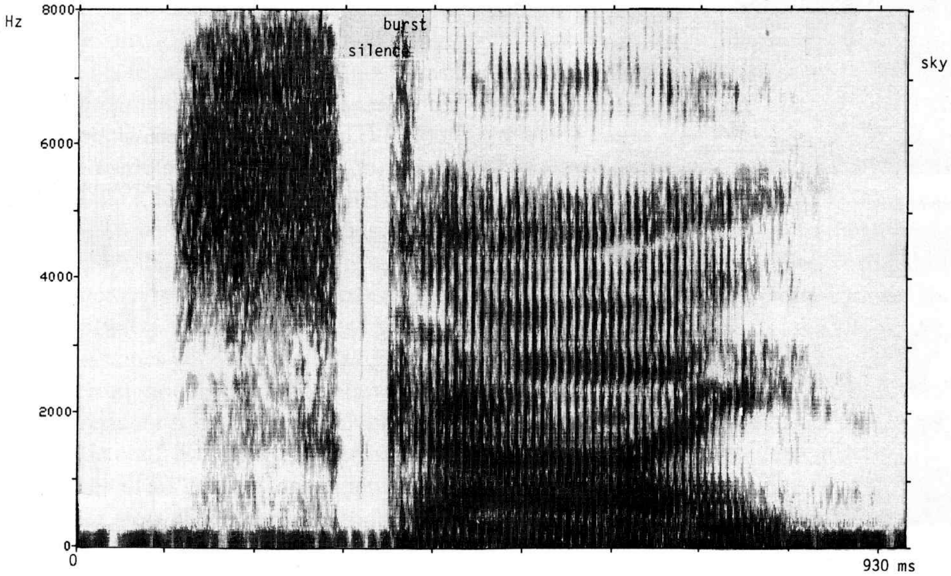

FIG. 3. Spectrograms of the word *sky* /skaj/, illustrating the *silence* and *burst* microsegments of voiceless stop /k/ in an /s/-stop-vowel context. Note that, in such a context, the voiceless stop is unaspirated.

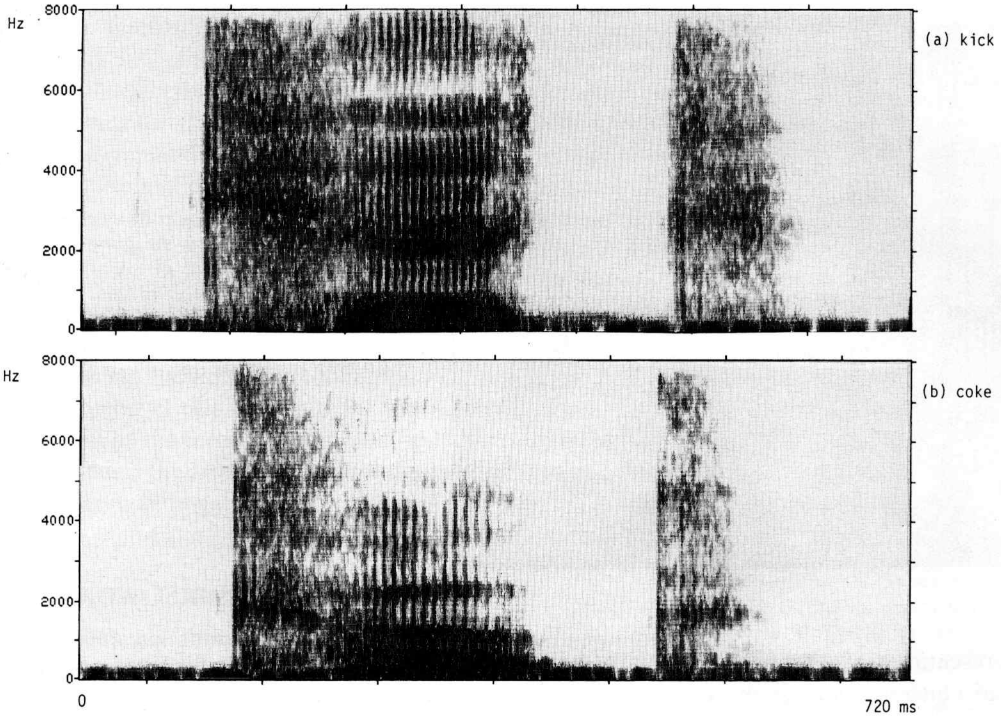

FIG. 4. Spectrograms of words (a) *kick* /kɪk/ and (b) *coke* /kok/, showing coarticulatory effects on /k/ by the adjacent vowel. The spectral prominences of the burst and aspiration for both the initial and final /k/ are well above 2 kHz with the front vowel context /ɪ/, while they are well below 2 kHz with the back (nonfront) vowel context /o/.

cantly different: One is widely spread around 3000 Hz, while the other is concentrated near 1700 Hz.

Two more examples are provided in Figs. 5 and 6. Figure 5 shows the coarticulatory effects of front and back vowels on initial /p/ in words *pea* /pi/ and *poke* /pok/. It is clear that spectral energy of the burst + aspiration portion of /p/ is concentrated above 2000 Hz with the presence of the following /i/, while if the following vowel is /o/, the spectral prominences of the burst + aspiration portion of /p/ are of lower frequency. A similar vowel-dependent effect on the initial /g/ is shown in Fig. 6 in the example words *geese* /gis/ and *go* /go/.

These observations on coarticulation motivated our development of context-dependent models of the release microsegments. Two vowel contexts are used in the modeling: front versus nonfront. The diphthong /aw/ is always associated with the nonfront-vowel group, while /ɔj/ and /aj/ are associated with the nonfront-vowel group when they influence a preceding stop and with the front-vowel group when they influence a following stop. The release microseg-

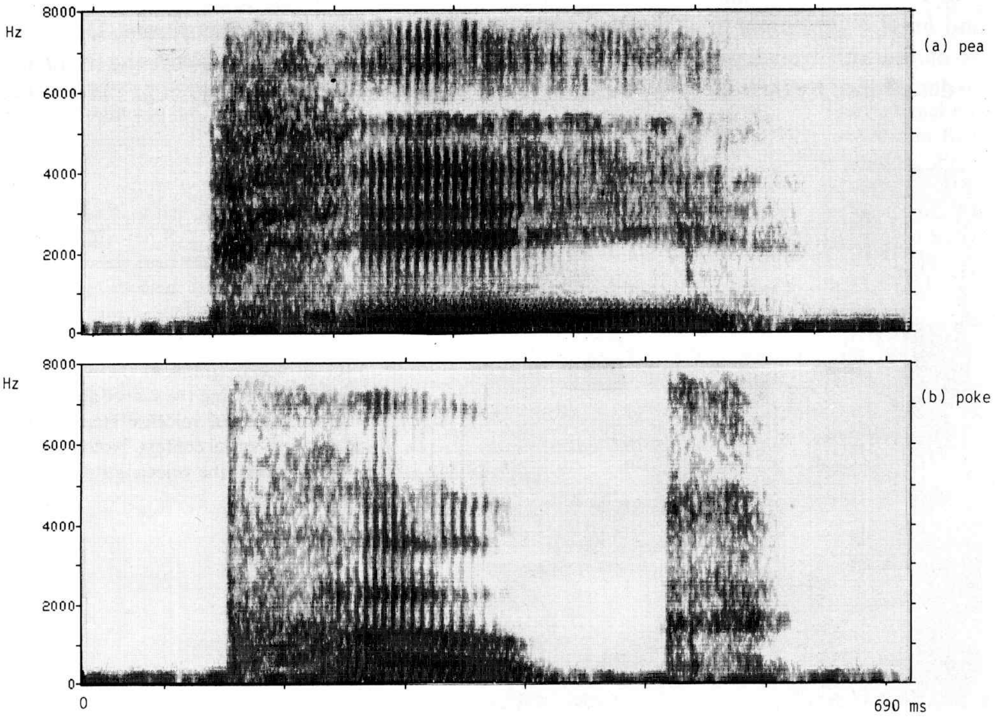

FIG. 5. Spectrograms of words (a) *pea* /pi/ and (b) *poke* /pok/, showing coarticulatory effects of vowels on initial /p/.

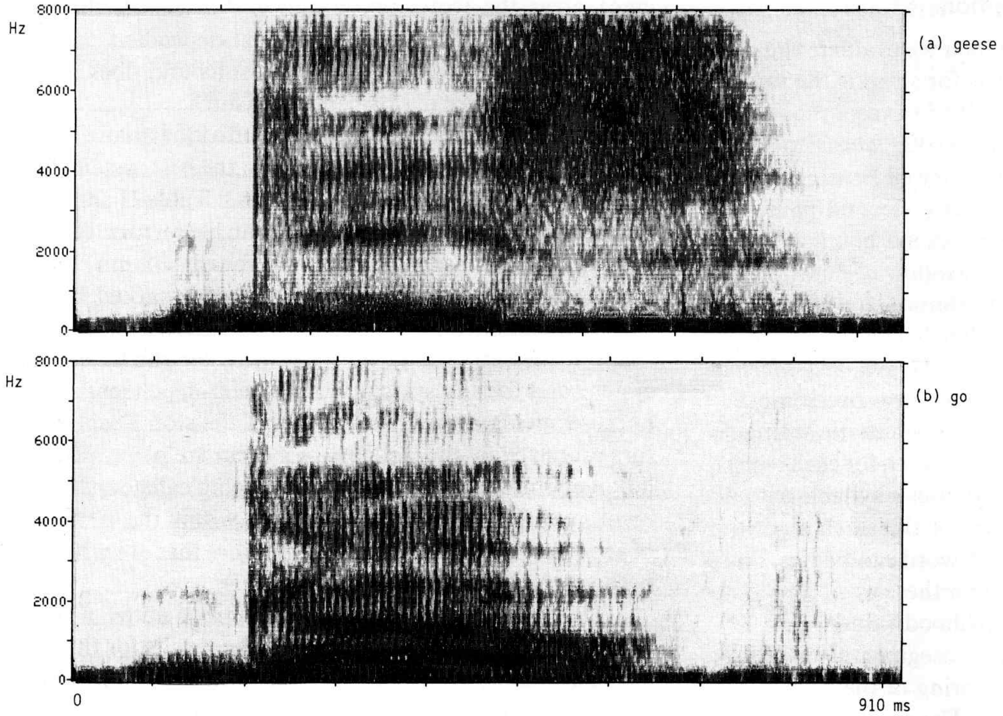

FIG. 6. Spectrograms of words (a) *geese* /gis/ and (b) *go* /go/ showing coarticulatory effects on initial /g/.

ments are conditioned on the front versus nonfront vowel context, thereby adding 6 microsegments into the 9 microsegments described previously for a total of 15 microsegmental HMMs.

## II. SPEECH RECOGNITION EXPERIMENTS

Prior to the present work, our large vocabulary speaker-trained isolated word recognizer used one HMM to represent each phoneme of English (except for the phonemes /l/ and /r/, for each of which two allophonic HMMs were used). The recognition performance obtained by this system, based essentially on one model per phoneme, is the benchmark with which we will compare the effect of using the more sophisticated stop models presented in this paper. We have also trained and tested a system based on three allophone models per stop phoneme, corresponding to initial, medial, and final positions. Finally, we have tested two microsegment-based systems (cf. Sec. I). The first uses nine microsegmental models. In the other, the 6 release microsegments are modeled context dependently, resulting in a total of 15 microsegmental HMMs.

Two male native speakers of English, one American and one Canadian, read the training and test sets once each. All HMMs were trained using the standard forward–backward algorithm (Baum, 1972; Liporace, 1982). The HMMs used multivariate Gaussian density functions with full covariance matrix for the output probability densities on each transition. One pooled covariance matrix was used per HMM.

### A. Training set

The training text consists of two parts: a prose part and a word list part. The prose part contains 925 words of arbitrarily selected passages from English language newspapers, books, and magazines. The word list part consists of 278 monosyllabic English words. Most of these are CVC words containing stop consonants.

The subset of the training data used to train the microsegment models consists of all 278 words in the word list part together with 300 of the 925 words in the prose part of the training set. The microsegments in these 578 tokens were manually segmented whenever the microsegments could be clearly identified through inspection of the spectrogram. Ambiguous, uncertain, or unclear microsegments were discarded from the training set.

The entire 1203-word training set was used to train the whole-phoneme HMMs, used to represent the stops in the benchmark experiment, and to represent phonemes other than stops in all experiments. In the experiment using three allophones per stop phoneme, stop tokens in the 1203-word training set were partitioned into initial, medial, and final occurrences to train their respective allophone models. The medial models were reused in the microsegment experiments to represent medial stops not modeled microsegmentally.

### B. Test set

Recognition experiments were performed using two test sets. The first test set is a word list whose vocabulary is disjunct from the word list used for training. The second test set consists of prose passages and does not include any of the prose selections used in training. The word list test set consists of 312 English words. These are mostly CVC words, as well as some CVCV words and words containing /s/-stop vowel. The CVC(V) words are selected to be highly confusable, e.g., *bit*, *bid*, *bib*, *big*, and *bitter*. The prose test set consists of 396 words of texts selected arbitrarily from books, magazines, and newspapers. The average rate of occurrence of initial and final stops in the prose test set is 0.487 times per word.

### C. Recognition system description

The recognition system in which we evaluate the effectiveness of microsegmental models for stops is the same as the one described in Gupta *et al.* (1988) except the vocabulary size has been increased from 60 000 words to 75 000 words. The system will be briefly reviewed here.

Speech material, read with quarter-second pauses between words, is recorded in a quiet sound booth at a sampling rate of 16 kHz. A Hamming window of duration 25.6 ms is applied every 10 ms. For each Hamming window, a 15-dimensional feature vector is calculated. The vector consists of mel-frequency cepstral coefficients (Davis and Mermelstein, 1980), augmented by their differences over time.

The recognition phase starts with automatic segmentation of words from the input sentence. Next, for each word to be recognized, a fast search strategy uses a syllable network (Gupta *et al.*, 1988) and a variant of the stack algorithm (Jelinek, 1976) to generate about 70 word candidates. Finally, likelihood scoring is carried out for these word candidates to determine the final acoustic likelihood values.

In the work presented here, microsegmental modeling is used only during the likelihood scoring of the word candidates (not during the fast search). The results reported assume that the list of word candidates generated by the fast search always contains the correct word.[4]

### D. Recognition results

For each input word, the output of the acoustic recognizer is an ordered list of word hypotheses together with their likelihood scores.[5] The performance of the acoustic recognizer is evaluated using three measures: the percentage of words correctly identified as the top (highest likelihood) word choice, the average rank of the correct word in the ordered candidate list, and the average difference between the log likelihoods of the correct word and of the highest scoring incorrect word. Evaluation measures other than the top choice are useful since, even if the percentage top choice accuracy is the same for different experimental conditions, a lower average rank or a higher average log likelihood difference would produce better performance when additional information (e.g., from a language model) is introduced. Table I shows these three performance measures on the word list test set for the four types of stop models under discussion.

As can be seen, use of three allophones (initial, medial, and final) for each stop phoneme gives similar recognition accuracy to use of nine context-independent stop microsegments. The benchmark system (one HMM per stop phoneme) gives the worst performance. Among all the techniques, the one using context-dependent release microsegments produces the best results and does so uniformly for all three performance measures.

We have examined in detail individual word errors made in the benchmark system and in the best system (context-dependent release microsegments). Table II shows the results of this analysis. The first column lists errors that have been corrected (marked by √ in the fourth column) and on which new errors have been introduced (marked by ×). The phonemic transcription of the recognizer's top choice is listed in the second and third columns for the benchmark system and for the system with context-dependent release microsegments, respectively. Most of the stop phoneme errors produced by the benchmark system are place of articulation confusions within the same voicing category. The remaining stop errors are voicing errors within the same place of articulation. (There are no stop errors that are wrong for both voicing and place of articulation.)

The same set of experiments described above was performed on the prose test set. Unlike the results for the word list, the overall error rates are nearly the same for the four types of stop models, as shown in Table III.

Table IV is analogous to Table II for the prose test set, examining the differential errors with respect to the benchmark recognizer. Out of a total of 61 word errors, only 9 errors are corrected. Nine additional errors are introduced, leading to identical recognition accuracies. Note error corrections and new errors have occurred for several test words, such as *this*, *new*, *in*, *hole*, and *here*, which do not contain initial and final stops and hence are not scored by the microsegmental models. Although identical likelihood scores are obtained by the benchmark models and by the microsegmental models for the correct word hypothesis, the competing candidate words are scored differently if they contain initial or final stops. Also note that four out of nine newly created errors involve words with stops in consonant clusters (other than /s/-stop vowel). Stops in these positions have not been modeled carefully in this study. If a stop in a consonant cluster happens to be initial or final, then it is treated in the same way as the initial or final stop in a CVC context. This simplistic treatment of stops in consonant clusters appears to be inadequate.

To improve our confidence in the above results, all the experiments were repeated using the speech of a second male speaker. The results are consistent with those of the first speaker. That is, for the prose test set, various types of stop models produce similar recognition performance, while, for

TABLE I. Comparative recognition performance of different stop model types on the word list test set (speaker 1).

| | Performance on word list test set | | |
|---|---|---|---|
| Type of model for stop consonants | Percent correct | Average rank | Average diff. of log scores |
| One model per stop phoneme | 68.6 | 2.30 | 11.0 |
| Three allophonic models per stop phoneme | 76.6 | 1.77 | 18.7 |
| Microsegments with context-independent releases | 75.6 | 1.96 | 14.7 |
| Microsegments with context-dependent releases | 80.4 | 1.61 | 20.7 |

TABLE II. Tokens in the word list test set for which recognition errors were corrected or newly introduced (speaker 1).

| Test Words | Top choice using one HMM for each stop phoneme | Top choice using context-dependent microsegments | Error correction(√) or new error (×) |
|---|---|---|---|
| *leaky* | /gliti/ *gleety* | /liki/ | √ |
| *league* | /lid/ *lead* | /lig/ | √ |
| *deep* | /tit/ *teet* | /dip/ | √ |
| *keep* | /tʃip/ *cheap* | /kip/ | √ |
| *teak* | /titʃ/ *teach* | /tik/ | √ |
| *lit* | /lɪd/ *lid* | /lɪt/ | √ |
| *bib* | /did/ *deed* | /bɪb/ | √ |
| *diddle* | /tɪtl/ *tittle* | /dɪdl/ | √ |
| *pick* | /pɪt/ *pit* | /pɪk/ | √ |
| *pity* | /pɪni/ *pinnae* | /pɪti/ | √ |
| *date* | /get/ *gate* | /det/ | √ |
| *take* | /tet/ *teth* | /tek/ | √ |
| *tape* | /tik/ *teak* | /tep/ | √ |
| *goat* | /bot/ *boat* | /got/ | √ |
| *pose* | /hoz/ *hose* | /poz/ | √ |
| *toad* | /kod/ *code* | /tod/ | √ |
| *tote* | /tɛnt/ *tent* | /tot/ | √ |
| *leg* | /lɛgd/ *legged* | /lɛg/ | √ |
| *debt* | /gɛt/ *get* | /dɛt/ | √ |
| *debtor* | /gɛtr/ *getter* | /dɛtr/ | √ |
| *pet* | /pɛd/ *ped* | /pɛt/ | √ |
| *ten* | /tæn/ *tan* | /tɛn/ | √ |
| *dies* | /bajz/ *buys* | /dajz/ | √ |
| *fight* | /faj/ *fie* | /fajt/ | √ |
| *height* | /haj/ *high* | /hajt/ | √ |
| *pipe* | /hajp/ *hype* | /pajp/ | √ |
| *bet* | /bɛd/ *bed* | /bɛt/ | √ |
| *dear* | /gɪr/ *gear* | /dɪr/ | √ |
| *dug* | /bʌg/ *bug* | /dʌg/ | √ |
| *dies* | /bajz/ *buys* | /dajz/ | √ |
| *kill* | /tʃil/ *chiel* | /kɪl/ | √ |
| *Bob* | /blɑb/ *blob* | /bɑb/ | √ |
| *goat* | /god/ *goad* | /got/ | √ |
| *coat* | /kod/ *code* | /kot/ | √ |
| *beat* | /bikt/ *beaked* | /bit/ | √ |
| *cot* | /kɑpt/ *Copt* | /kɑt/ | √ |
| *reap* | /rik/ *reak* | /rip/ | √ |
| *spite* | /skrajb/ *scribe* | /spajt/ | √ |
| *bid* | /bɪd/ | /bild/ *bield* | × |
| *writer* | /rajtr/ | /rajdr/ *rider* | × |
| *boat* | /bot/ | /bod/ *bode* | × |
| *gate* | /get/ | /kit/ *keet* | × |
| *dit* | /dɪt/ | /dip/ *deep* | × |
| *peep* | /pip/ | /dʒip/ *jeep* | × |
| *leap* | /lip/ | /klip/ *clepe* | × |

the word list test set, microsegmental models with context-dependent releases produce a significantly higher recognition rate than the other types of stop models. A comparison of the results is shown in Tables V and VI.

## III. SUMMARY AND DISCUSSION

The modeling technique reported in this paper focuses on improving discrimination of the six stop phonemes. We make use of acoustic-phonetic knowledge about the stops to obtain composite models that represent a sequence of speech events. The specific knowledge used for positing the modeling approach is as follows: (1) A stop, although highly nonstationary in its overall extent, nevertheless consists of a few relatively stationary portions that we call microsegments; (2) some of these microsegments are shared across different stop phonemes and across different allophones of the same phoneme; (3) the release of the voiceless stop in /s/-stop-vowel context resembles the release of the homorganic

TABLE III. Comparative recognition performance of different stop model types on the prose test set (speaker 1).

| Type of model for stop consonants | Performance on prose test set | | |
|---|---|---|---|
| | Percent correct | Average rank | Average diff. of log scores |
| One model per stop phoneme | 84.6 | 1.72 | 31.7 |
| Three allophonic HMMs per stop | 84.5 | 1.72 | 31.8 |
| Microsegments with context-independent releases | 84.6 | 1.76 | 30.9 |
| Microsegments with context-dependent releases | 84.6 | 1.72 | 31.9 |

TABLE IV. Tokens in the prose test set for which recognition errors were corrected or newly introduced (speaker 1).

| **Test Words** | **Top choice using one HMM for each stop phoneme** | **Top choice using microsegments with context-dependent releases** | **Error correction(√) or new error (×)** |
|---|---|---|---|
| *said* | /stɛd/ *stead* | /sɛd/ | √ |
| *this* | /bɪs/ *bis* | /ðɪs/ | √ |
| *new* | /bu/ *boo* | /nu/ | √ |
| *in* | /ɪnd/ *ind* | /ɪn/ | √ |
| *hook* | /kʊk/ *cook* | /hʊk/ | √ |
| *tied* | /tajnd/ *tined* | /tajd/ | √ |
| *hole* | /pɔl/ *pall* | /hol/ | √ |
| *not* | /bɑnət/ *bonnet* | /nɑt/ | √ |
| *an* | /ænd/ *and* | /æn/ | √ |
| *described* | /dəskrajbd/ | /əstrajd/ *astride* | × |
| *but* | /bʌt/ | /bɑrɛt/ *barrette* | × |
| *community* | /kəmjunəti/ | /təmɪdəti/ *timidity* | × |
| *leapt* | /lipt/ | /lid/ *lead* | × |
| *greatest* | /gretɪst/ | /gredəs/ *gradus* | × |
| *here* | /hɪr/ | /pɪr/ *peer* | × |
| *grows* | /groz/ | /broz/ *brose* | × |
| *that* | /ðæt/ | /ðæn/ *than* | × |
| *in* | /ɪn/ | /ɪnd/ *ind* | × |

TABLE V. Comparative recognition performance of different stop models on the word list test set (speaker 2).

| Type of model for stop consonants | Performance on word list test set | | |
|---|---|---|---|
| | Percent correct | Average rank | Average diff. of log scores |
| One model per stop phoneme | 67.6 | 2.47 | 10.1 |
| Three models per stop phoneme | 74.3 | 2.04 | 14.7 |
| Microsegments with context-independent releases | 74.0 | 1.98 | 14.3 |
| Microsegments with context-dependent releases | 77.9 | 1.62 | 18.1 |

TABLE VI. Comparative recognition performance of different stop model types on the prose test set (speaker 2).

| | Performance on prose test set | | |
|---|---|---|---|
| Type of model for stop consonants | Percent correct | Average rank | Average diff. of log scores |
| One model per phoneme | 76.4 | 1.79 | 30.5 |
| Three allophones per phoneme | 76.8 | 1.82 | 31.3 |
| Microsegments with context-independent releases | 76.9 | 1.78 | 30.5 |
| Microsegments with context-dependent releases | 76.5 | 1.78 | 31.1 |

voiced stop; and (4) the spectra of the release microsegments are strongly affected by place of articulation and labialization features of the adjacent vowel.

Based on the above knowledge, we have constructed a set of microsegmental HMMs for the stops. Use of these models for word list test data has reduced the error rate for speaker 1 from 31.4% (benchmark result, obtained by using one model per stop phoneme) to 19.6% (best result, obtained by using microsegmental models with context-dependent releases). For speaker 2, the corresponding error reduction is from 32.4% to 22.1%. At present, stops are modeled microsegmentally only when they are initial, final, or when they are in /s/-stop-vowel context.

Three significant advantages arise from microsegmental stop models. First, discriminative focusing: These models focus on the information-bearing elements of stops, i.e., the release microsegments. Because the same *voice bar* model is used in all three voiced stops, random variation in the voice bar region cannot adversely affect the discrimination among the three voiced stop phonemes. The same is true for discrimination among the three voiceless stops, because the same *silence* model is used in all three of them. In contrast, when a single HMM is used to represent the whole stop phoneme, random differences during the period of voice bar or silence can mask the discriminative information differentiating place of articulation. This effect is accentuated by the fact that voice bar or silence typically occupies a significantly longer duration than the release. Second, since the *voice bar* model and the *silence* model are trained jointly from the three voiced stops and the three voiceless stops, respectively, the resulting models possess stronger discriminability of the voicing feature. Third, the spectral similarity of release microsegments in initial versus final positions allows pooling of the microsegment tokens, thereby enhancing the robustness of the models. It is largely due to this pooling that the context-dependent microsegmental approach only minimally increases the total number of models needed.

Using three allophonic HMMs per stop phoneme also performs significantly better than using a single model per stop phoneme: It reduces the word error rate by 21% (averaged over two speakers; see Tables I and V). However, use of three allophones per stop lacks two of the advantages of microsegmental models described above: discriminative focusing and reduction of the number of models required. For example, in contrast to the 15 microsegmental HMMs needed to represent six context-dependent stops, using three allophones per stop, if further conditioned on the front versus nonfront vowel context, would require 36 HMMs. In addition, each allophonic HMM requires a greater number of states to accommodate the higher complexity time-frequency patterns of whole stop phonemes. This makes such models less robust than the microsegmental models when limited training data are available.

Several speech recognition groups have shown that context-dependent Markov modeling is valuable at the phonemic level (Schwartz *et al.*, 1984; Derouault, 1987; Paul and Martin, 1988; Lee *et al.*, 1989; Deng *et al.*, 1989a,b, 1990). The present study extends these results from the phonemic level to the microsegmental level. Although in the present work we have used only two contexts, i.e., front versus nonfront vowels, a reduction in word error rate of 17% (from 25.2% error to 20.8% error when averaged over two speakers; see Tables I and V) has been achieved.

In contrast to the significant error rate reduction obtained using microsegmental stop models for the word list test set, no corresponding error rate reduction is obtained for the prose test set. This result appears to be caused by several factors. (1) Speakers use different speaking styles when reading word lists as opposed to prose (Labov, 1972, pp. 80–85). In the prose reading style, we often observe unreleased or weakly released final stops. In contrast, the word list style tends to elicit citation forms, in which most of the stops are released. Since our microsegment models are trained using only the strongly released final stops in the training data, they do not adequately represent the microsegments in unreleased or weakly released final stops, which occur frequently in the prose test set. Although heuristic use of a null transition (described in Sec. I A) has partially remedied the problem, this simplistic approach is inadequate for coping with the complex behavior of the stops in reading style speech. (2) Medial stops are not modeled microsegmentally in the current study: They are modeled by one HMM per phoneme. (3) Microsegmental modeling of initial and especially final stops in consonant clusters is inexact (cf. discussion in Sec. II D). (4) The average rate of occurrence of initial and final stops in the prose test set is less than 0.5 times per word, while, in the word list test set, it is greater than 1.0 times per word. Furthermore, the prose set contains proportionally fewer monosyllabic words and potential stop consonant minimal pairs. Any advantage that the microsegmental approach might have therefore shows up less clearly in the prose set than in the word list set. (5) The size of the training set used to train the microsegmental models is less than half that used to train the whole phoneme models.

We propose the following solutions to the first three problems listed above. (1) Increase the percentage of prose in the microsegment training set. (2) Develop microsegmental analyses of the stops in medial position. For fully articulated intervocalic stops this is straightforward; for flaps and for stops occurring in clusters, judicious analysis will be required. (3) Develop a more realistic microsegmental analysis of stops in final position occurring in consonant clusters. This analysis should allow for stochastic deletion of these stops. A better microsegmental analysis of initial stops in clusters other than /s/-stop-vowel may also be required.

We believe that the large effect observed on the word list test set is very real and is a potentially useful technique for improving the acoustic recognition of prose input as well. The power of this technique for prose should become more evident as the three solutions outlined above are implemented. This is the direction of our current research in microsegmental modeling.

## ACKNOWLEDGMENTS

Financial support for this work was graciously provided by the Natural Sciences and Engineering Research Council of Canada under their Strategic Grants program. The authors would also like to thank their colleagues Vishwa Gupta, Patrick Kenny, and Franz Seitz for their helpful comments.

[1] In some systems, phonetically well-understood allophones of certain phonemes, such as flap-t [ɾ], are represented as distinct HMMs.

[2] In subsequent publications, the works of Lee and Hon (1988) and of Deng *et al.* (1988) have been extended to include context-dependent models of allophones (Lee *et al.*, 1989; Deng *et al.*, 1990).

[3] In this paper, the terms *initial*, *medial*, and *final* refer to *word-initial*, *word-medial*, and *word-final*, respectively.

[4] In practice, the candidate list contains the correct choice about 97% of the time. For the purposes of the experiments presented here, when the correct word was absent from the candidate list, it was added to the list.

[5] The output of the acoustic recognizer normally serves as the input for the language model module. In the experiments reported here, no language model is used. All words are considered *a priori* equally probable. The recognition accuracy obtained in this way is purely due to the acoustic information in the input speech.

Bahl, L. R., Bakis, R., Cohen, P. S., Cole, A. G., Jelinek, F., Lewis, B. L., and Mercer, R. L. (**1980**). "Further results on the recognition of a continuously read natural corpus," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **2**, 872–875.

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

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. Electr. Comput. Eng. **1**, 315–318.

Deng, L., Lennig, M., Seitz, F., and Mermelstein, P. (**1990**). "Large vocabulary word recognition using context-dependent allophonic hidden Markov models," Computer Speech Lang. **4** (to be published).

Deng, L., Kenny, P., Lennig, M., Gupta, V., and Mermelstein, P. (**1989a**). "A locus model of coarticulation in an HMM speech recognizer," Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. **1**, 97–100.

Deng, L., Lennig, M., and Mermelstein, P. (**1989b**). "Use of vowel duration information in a large vocabulary word recognizer," J. Acoust. Soc. Am. **86**, 540–548.

Derouault, A. (**1987**). "Context-dependent phonetic Markov models for large vocabulary speech recognition," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 360–363.

Fant, G. (**1973**). *Speech Sounds and Features* (MIT, Cambridge, MA).

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

Halle, M., Hughes, G. W., and Radley, J. P. A. (**1957**). "Acoustic properties of stop consonants," J. Acoust. Soc. Am. **29**, 107–116.

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

Labov, W. (**1972**). *Sociolinguistic Patterns* (University of Pennsylvania, Philadelphia).

Lee, K. F., and Hon, H. W. (**1988**). "Large vocabulary speaker-independent continuous speech recognition," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 123–126.

Lee, K. F., Hon, H. W., Hwang, M. Y., Mahajan, S., and Reddy, R. (**1989**). "The SPHINX speech recognition system," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 445–448.

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

Merialdo, B. (**1987**). "Speech recognition with very large size dictionary," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 364–367.

Moore, R. K., Russell, M. J., and Tomlinson, M. J. (**1983**). "The discriminative network: A mechanism for focusing recognition in whole-word pattern matching," Proc. IEEE Conf. Acoust. Speech Signal Process. **2**, 1041–1044.

Murveit, H., and Weintraub, M. (**1988**). "Speaker-independent connected-speech recognition using hidden Markov models," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 115–118.

Paul, D. B., and Martin, E. A. (**1988**). "Speaker stress-resistant continuous speech recognition," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 283–286.

Schwartz, R. M., Chow, Y. L., Roucos, S., Krasner, M., and Makhoul, J. (**1984**). "Improved hidden Markov modeling of phonemes for continuous speech recognition," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **2**, 35.6.1–35.6.4.