Computer Speech and Language (1992) **6**, 345–359

Use of minimum duration and energy contour for phonemes to improve large vocabulary isolated-word recognition

V. Gupta*, M. Lennig*, P. Mermelstein*, P. Kenny, P.F. Seitz† and D. O'Shaughnessy

INRS-Télécommunications, 16 Place du Commerce, Montreal, Quebec, Canada H3E 1H6

Abstract

Many acoustic misrecognitions in our 86 000-word speaker-trained isolated-word recognizer are due to phonemic hidden Markov models (phoneme models) mapping to short segments of speech. When we force these models to map to longer segments corresponding to the observed minimum durations for the phonemes, then the likelihood of the incorrect phoneme sequences drops dramatically. This drop in the likelihood of the incorrect words results in significant reduction in the acoustic recognition[1] error rate. Even in cases where acoustic recognition performance is unchanged, the likelihood of the correct word choice improves relative to the incorrect word choices, resulting in significant reduction in recognition error rate with the language model. On nine speakers, the error rate for acoustic recognition reduces from 18·6 to 17·3%, while the error rate with the language model reduces from 9·2 to 7·2%.

We have also improved the phoneme models by correcting the segmentation of the phonemes in the training set. During training, the boundaries between phonemes are not marked accurately. We use energy to correct these boundaries. Application of an energy threshold improves the segment boundaries between stops and sonorants (vowels, liquids and glides), between fricatives and sonorants, between affricates and sonorants and between breath noise and sonorants. Training the phoneme models with these segmented phonemes results in models which increase recognition accuracy significantly. On two speakers, the error rate for acoustic recognition reduces from 26·5 to 23·1%, while the error rate with the language model reduces from 11·3 to 8·8%. This reduction in error rate is in addition to the error rate reductions obtained by imposing minimum duration constraints. The overall reduction in errors for these two speakers using minimum durations and energy thresholds is from 27·3 to 23·1% for acoustic recognition, and from 14·3 to 8·8% with the language model.

* Also with Bell-Northern Research, Montreal.

† Currently at Center for Auditory & Speech Sciences, Gallaudet University, Washington, D.C., U.S.A.

[1] We use the term acoustic recognition error rate to mean the recognition error rate when every word in the vocabulary is considered *a priori* equally likely.

0885–2308/92/040345 + 15 $08.00/0

1. Introduction

The goal of our 86 000-word recognizer is to transcribe speech spoken as a sequence of isolated words. For each spoken word, the recognizer uses acoustic information and rough likelihoods in a fast search algorithm (Gupta, Lennig & Mermelstein, 1988) to narrow the possible word hypotheses from the 86 000 words (Seitz *et al.*, 1990) in the total vocabulary to a small list. It then refines the list by computing an exact likelihood score for each hypothesized word. The exact likelihood scores take into account acoustic information but not the syntactic, semantic and pragmatic characteristics of English. To take these into account, the exact likelihoods are further refined with the aid of a statistical language model to generate the most likely sequence of words (Gupta, Lennig & Mermelstein, 1992). In this paper our focus is on improving the accuracy of our recognizer through improvements in the acoustic recognition algorithm.

Our strategy is to improve recognition accuracy by eliminating weaknesses inherent in hidden Markov modelling algorithms for speech recognition. For instance, hidden Markov models incorporate only weak duration constraints in the phonemes they generate. During both the fast search algorithm and the exact likelihood scoring, the phoneme[2] models are mapped to acoustic segments in order to compute the likelihood of the acoustic data. When an incorrect phoneme sequence is mapped to the acoustic input, we frequently observe that one or more models are often mapped to acoustic segments shorter than the minimum possible duration for the phoneme. One such example can be seen in Fig. 1. Here, the word spoken is *veins*, while the best choice produced by the recognizer is *brings*. Notice that the model for /r/ is mapped to 20 ms of acoustic data, while the *duration minimum*[3] for /r/ observed in the training data is 40 ms. We consider mapping of phoneme models to acoustic segments shorter than those observed in the training data as a weakness in hidden Markov modelling. This weakness has been addressed by imposing minimum duration constraints on the phonemes in the HMM framework. These minimum durations are derived from the training data. Imposing duration minima on the phonemes results in significant reduction in the likelihoods of incorrect word hypotheses, and increases both acoustic recognition accuracy and recognition accuracy with the language model.

In previous work, Bush and Kopec (1987) apply minimum duration constraints on speech segments to improve digit-string recognition accuracy. They find that a minimum duration constraint of 50 ms for each acoustic segment is optimal for their digit-string recognizer. Soong (1989) has used minimum durations to improve phoneme recognition accuracy in spoken Japanese text. He assigns minimum allophone durations to 2084 allophonic HMMs to improve phoneme recognition accuracy. In our experiments with allophone modeling (Deng *et al.*, 1990), we have obtained best recognition accuracy using 44 phonemic mixture HMMs (see Table I for details on the 44 models used). In this paper we impose phoneme duration constraints during recognition using these 44 phoneme models. For the sonorants and affricates, one duration minimum per phoneme is used everywhere, while for the remaining phonemes, the duration minimum depends on whether the phoneme occurs in the initial, medial or final position.[4] The duration

[2] The models represent phonemes, except for /l/ and /r/, where we use two allophones (prevocalic and syllabic/postvocalic). Despite the fact that /l/ and /r/ each have two models, we refer to each of the 44 models in our system as a phoneme model.

[3] We refer to the minumum duration of a phoneme observed in the training set as its duration minimum.

[4] We use the terms initial, medial and final to mean word-initial, word-medial and word-final, respectively.

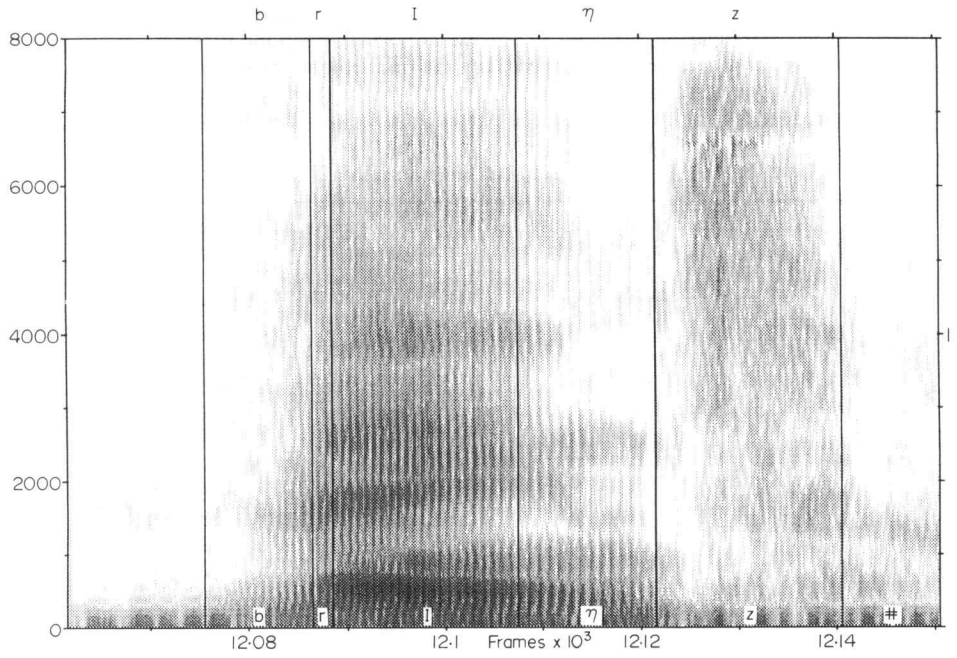

Figure 1. An example of an HMM mapping to a short duration acoustic segment. In this example, *veins* was misrecognized as *brings*. Notice that the phoneme model for /r/ gets mapped to a 20 ms-long acoustic segment. The minimum duration for /r/ we have observed in the training set is 40 ms.

minimum varies from 20 ms to 100 ms depending on the phoneme and the context in which it appears.

Another weakness in hidden Markov modelling that we have addressed is the erroneous segmentation of words into phonemes. Many segmentation errors in the training data are between high energy and low energy phonemes. A number of these segmentation problems have been corrected by constraining the energy contours to prevent phonemes with high energy from mapping onto phonemes with lower energy. Segment boundaries between stops and sonorants, between fricatives and sonorants, between affricates and sonorants and between breath and sonorants are most amenable to correction using energy constraints. Segment boundaries between vowels, liquids, glides and nasals can be corrected by using duration minima. Correction of the segment boundaries between phonemes in the training set leads to improved phoneme models, resulting in higher acoustic recognition accuracy and higher recognition accuracy with the language model.

Bush and Kopec (1987) and Kopec and Bush (1985) have also applied energy constraints to improve isolated digit and digit–string recognition accuracy. The energy constraints have been imposed on either the peak energy in the speech segment or on the minimum energy in the speech segment. For example, imposing the constraint that the peak energy in a stressed vowel segment be above a certain threshold reduces certain digit insertion errors (Bush & Kopec, 1987). Forcing the minimum energy in voiceless fricative segments to be below a certain threshold reduces voiceless fricative confusion

TABLE I. Duration constraints in milliseconds for different phonemes used in recognition. The first 44 phonemes listed correspond to phoneme models used in recognition. Note that the phoneme models also include models for initial and final breath. Initial sonorants can be optionally preceded by initial breath, while the final sonorants can be optionally followed by final breath. Phonemes in the last four rows have position-dependent duration minima

Allophone	Min. duration (ms)	Allophone	Min. duration (ms)
/aj/	100	/aw/	100
/ɔj/	100	/ɑ/ ~ /ɔ/	70
/i/	60	/ɪ/	40
/e/	70	/ɛ/	40
/æ/	60	/ɑr/	100
/ʌ/	60	/ʊ/	50
/u/	70	/o/	70
/ɔr/	80	/ə/	20
/j/	30	/w/	50
[l]	40	[l̥]	60
[r]	40	[r̥]	60
Initial breath	30	Final breath	30
/p/	60	/b/	40
/t/	40	/d/	40
/k/	70	/g/	40
/ʧ/	80	/dʒ/	70
/f/	70	/v/	40
/θ/	70	/ð/	50
/s/	90	/z/	80
/ʃ/	100	/ʒ/	60
/m/	40	/n/	40
/ŋ/	70	/h/	60
	Initial [t b d g ð]		20
	Initial [p f θ v]		40
	Initial [k]		60
	Final [p b t d k g f θ v ð]		20

with sonorants (Kopec & Bush, 1985). These constraints have been incorporated in their recognition algorithm in order to improve recognition accuracy of their isolated digit and digit–string recognizers.

Since our aim is to improve segment boundaries between phonemes during Viterbi training, the energy constraints take a very different form than those used by Bush and Kopec (1987). For example, to mark boundaries between vowels and fricatives accurately, we impose the constraint that every frame in the vowel segment has energy above a certain threshold. (Note that Bush and Kopec require only one of the frames in a stressed vowel segment to have energy above a certain threshold.) Similarly, we require every frame of a fricative to have energy below a certain threshold.

2. Exploiting phoneme durations to improve recognition accuracy

A detailed description of the recognition system is given in Gupta, Lennig and

Mermelstein (1988). Here, we outline details pertinent to this paper only. The 44 allophonic mixture HMMs used in the recognizer are described in detail in Deng *et al.* (1990). In this section we also outline weaknesses in hidden Markov modelling algorithms used for recognition, and show how these weaknesses have been addressed by imposing duration constraints in the recognition algorithm.

A block diagram of the recognition system is shown in Fig. 2. Recognition is performed in four stages. The first stage determines the end-points of the words using C_0, the log of a weighted spectral energy. The temporal sequence of feature vectors between these two end-points is employed by the succeeding stages to recognize the unknown word.

The second stage of recognition generates a number of hypotheses for the syllable count (total number of syllables) in the unknown. Estimation of the syllable count allows us to restrict our search to a subset of the vocabulary. For each syllable count estimate K, we form a graph consisting of a concatenation of K distinct syllable networks. Each arc of the graph corresponds to one of 44 phoneme models listed in Table I. Every phoneme sequence in the lexicon corresponding to a K-syllable word has a path through this graph. This graph is called the syllabic graph of count K.

The third stage of recognition is a syllabic graph search (or fast search algorithm) which computes the sequence of most likely phoneme strings through the syllabic graph. The syllabic graph search performs a fast search through all possible paths within the syllabic graph using a variation of the *stack algorithm* (Jelinek, 1976), or the *A* algorithm* as it is termed in the artificial intelligence literature (Nilsson, 1980). The output of the syllabic graph search is a list of the N most likely lexically valid phoneme strings and their associated rough likelihoods.

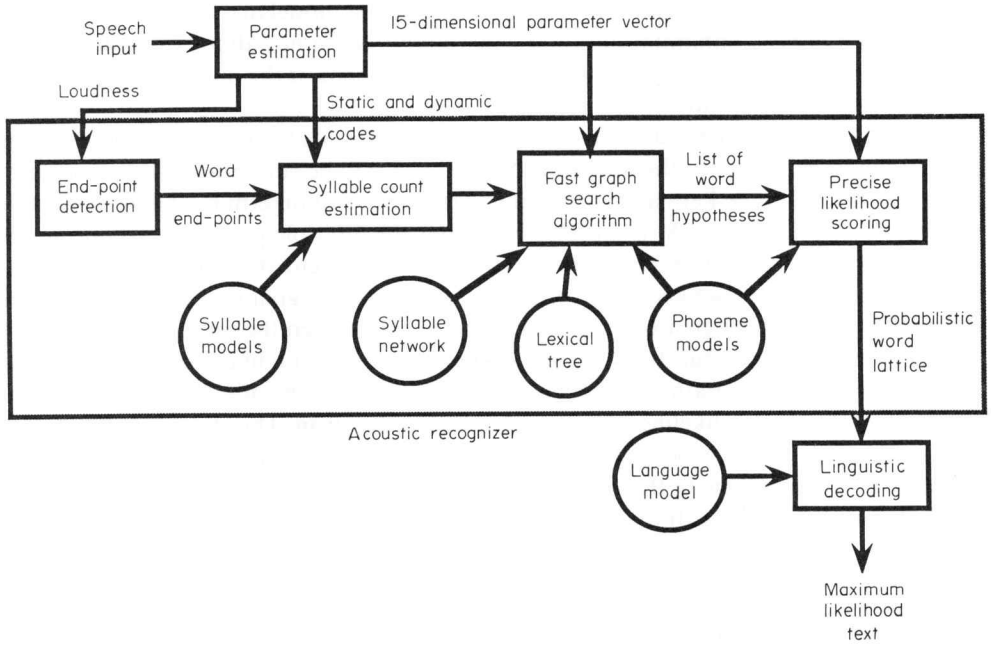

Figure 2. Block diagram of the large vocabulary recognition system.

The fourth stage of recognition computes the exact likelihood scores for the phoneme strings which were generated by the syllabic graph search, and re-orders the phoneme strings based on these scores. Finally, this word lattice, together with the associated likelihoods, is input to a trigram language model (Gupta, Lennig & Mermelstein, 1992) which outputs the most likely sentence.

Let us first outline one of the weaknesses observed in hidden Markov modelling. Many incorrect words cannot be differentiated from the correct word on the basis of the exact likelihoods. A number of incorrect words have high likelihoods due to phoneme models matching well with acoustic segments shorter than the duration minimum for the phoneme. Figure 1 shows an example of one such match. Notice that the model for [r] is mapped to 20 ms of the phoneme /e/, while the duration minimum for [r] is 40 ms.

Forcing phoneme models to match to acoustic segments exceeding the duration minima of the phonemes results in a dramatic drop in the likelihoods of many of these incorrect word choices. Such drops in likelihoods of incorrect words result in significant improvement in both the acoustic and language recognition accuracy. There are many possible strategies for incorporating duration[5] constraints.

We have applied duration constraints by restricting the possible state sequences through the phoneme model to correspond to the duration minimum for the phoneme.[6] Consider how these duration constraints can be incorporated into the HMM framework. We would like to find the most likely state sequence (through the phoneme models corresponding to the given phoneme sequence) which generates the given acoustic input and obeys the duration minima for the phonemes. To obtain such a state sequence, let us look at all possible paths which end at state S_i at time t (see Fig. 3). In the Viterbi formulation without duration constraints, we only keep the most likely state sequence to state S_i at time t. To incorporate duration constraints, we also compute the duration in frames of the current phoneme up to state S_i at time t. This duration d_j corresponds to the total number of transitions taken through the current model in order to reach state S_i at time t. To impose minimum duration constraints, we do not allow a transition from state S_i at time t to a state in the model for the next phoneme if the duration d_j is less than the minimum allowed duration.

If we keep only the best state sequence for each state S_i at time t, then the resulting path (or state sequence) to the final state S_F at time T will be suboptimal. Let us see how many paths need to be kept at each point (S_i, t) in order to determine the optimal state sequence to the final state S_F at time T. Note that the multiple paths correspond to the most likely state sequences to the point (S_i, t) with different durations in the current model. At every point we keep only one path with duration greater than the minimum. If a path with duration greater than the minimum exists, then the only additional paths we keep are those with duration less than the minimum having likelihoods greater than the minimum duration path. A path with duration less than the minimum is kept only if there is no other higher likelihood path with a longer duration. The maximum number of paths that may have to be kept at any point cannot be greater than the minimum duration of the phoneme in frames. This happens only when the likelihood of the paths reduces monotonically with increasing duration.

[5] The models already contain weak duration constraints through state transition probabilities. These duration constraints do not penalize short durations adequately.

[6] These phoneme models are left-to-right Gaussian mixture HMMs with self-loop, next state and skip transitions (Deng *et al.*, 1990).

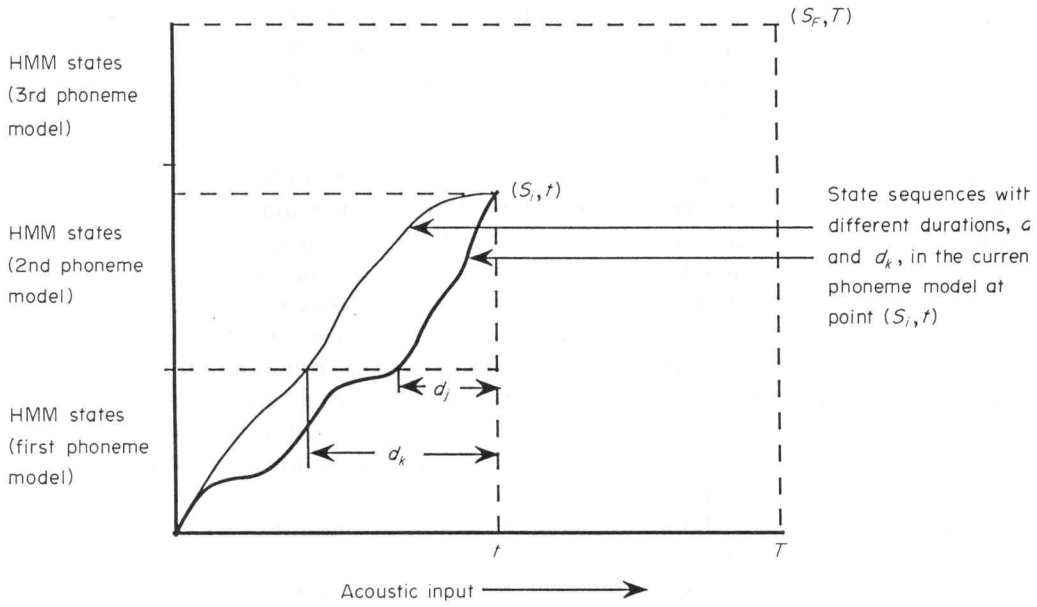

Figure 3. Trellis for Viterbi search showing how duration is incorporated in the search.

The implementation of duration constraints in the fast search algorithm is quite similar. In the fast search algorithm we generate a list of phoneme sequence hypotheses. These hypotheses are generated by finding the most likely phoneme sequences through a syllabic network, where each branch of the network corresponds to a phoneme model (Gupta, Lennig & Mermelstein, 1988). To find the most likely phoneme sequence through this syllabic network we estimate the likelihood of each branch in this network. The likelihood of a branch corresponds to the likelihood of all possible paths through this branch in the network. We apply minimum duration constraints during estimation of this likelihood by forcing the model corresponding to this branch to map to an acoustic segment longer than the specified duration minimum for the phoneme.

3. Experimental results using duration constraints

Before describing estimation of minimum phoneme durations we outline the experimental set-up. We have recorded speech from a total of nine native English speakers. Each speaker read betweeen 2000 and 3000 words, pausing at least 150 ms between words. The words correspond to paragraphs selected arbitrarily from magazines, books and newspaper articles. Each speaker took between two and five sessions to record the entire script. A part of the text was used for training the phoneme models, while the remaining text was used for estimating the recognition accuracy of the algorithm. No attempt was made to separate the training and test data according to recording sessions. Two of the nine speakers are co-authors of this paper (speakers ML and FS).

In deriving the duration minimum we have looked at the durations of phonemes in the training set for two speakers with fast speaking rate (among the nine speakers). To

derive duration minima for phonemes we observe a few tokens of each phoneme in contexts where they are expected to be short. The entire training set need not be manually segmented. We require manual segmentation since the Viterbi segmentation does not provide correct values for the duration minima. The duration minima derived from two speakers have been used for all nine speakers.

We have derived one duration minimum per phoneme, except for stops and fricatives, whose duration minimum depends on position. For each stop and fricative we use three possible duration minima corresponding to the phoneme occurring in initial, medial or final position. Stops in initial position can be shorter, since the silent stop gap or voice bar may not be present. Similarly, final stops may be unreleased or very weakly released, in which case the end-point detector may not classify the stop burst as part of the word. Also, initial and final weak fricatives /θðfv/ can be shorter. Therefore, we have reduced the minimum durations for initial and final weak fricatives as shown in Table I. The phoneme durations observed are generally longer than what would be expected in continuous speech. The minimum durations thus derived are used for all speakers.

Recognition results with and without duration constraints are compared for nine speakers in Table II. The duration constraints reduce the number of search errors in the fast search algorithm. A search error occurs when the correct word is absent from the list of possible word hypotheses. The number of search errors is reduced from 210 to 121 (3·1 to 1·8%). Application of duration constraints in the exact likelihood scoring stage results in a reduction of acoustic recognition error rate from 18·6 to 17·3%. However, the major effect of the duration constraints is evident after applying the language model. There, the duration constraints result in a reduction of 22% in word errors (from 9·2 to 7·2%). The reason for the reduced error rate is the relative lowering of the likelihoods of incorrect words as compared to the correct word. This is evident from Table III, which shows the average values of [log likelihood (correct word) − log likelihood (incorrect word)] with and without duration constraints.

In conclusion, forcing phoneme models to map to longer segments results in significant reduction in the likelihoods of incorrect segments. Such reductions lead to reduced search errors in the fast search algorithm, improved acoustic recognition and significant increase in word recognition accuracy after the language model. Duration minima derived from a few speakers with a fast speaking rate are effective on other speakers also. Even though we cannot claim these duration minima to apply to all possible speakers, we expect them to be applicable to a majority of speakers.

One question we have not answered is whether we can use more precise duration information instead of the duration minima to improve recognition accuracy further. In this case we would require reliable statistics of phoneme durations in various contexts. To collect statistics for phoneme durations we need a large amount of training data manually segmented into phonemes. At present, we do not have such a database. Also, such statistics may be speaker-dependent, requiring a large amount of manually segmented training data for each speaker.

4. Use of an energy measure to improve training of phoneme models

The phoneme models can be trained by a forward–backward algorithm or by a Viterbi algorithm (Levinson, Rabiner & Sondhi, 1983). In the context of our large vocabulary recognizer, the two training algorithms result in identical recognition accuracy. With Viterbi training, the implied segmentations can be evaluated to see how effectively the

TABLE II. Recognition results for nine speakers with and without duration constraints

Speaker (sex)	Total words		Acoustic recognition				Errors after lang. model	
			Search errors		Recog. errors			
	Test	Training	No dur. (%)	Dur. (%)	No dur. (%)	Dur. (%)	No dur. (%)	Dur. (%)
DS (m)	451	1900	4·9	3·1	24·0	21·9	14·4	10·4
AM (m)	565	2742	3·6	1·8	30·6	31·0	14·2	12·2
ML (m)	596	2000	1·7	1·2	14·5	12·6	6·7	5·4
JM (m)	587	1664	3·0	3·0	23·9	22·7	8·2	7·8
FS (m)	1014	2322	1·7	1·3	8·4	7·5	5·0	3·7
NM (f)	967	1299	4·0	1·7	19·4	15·0	11·0	6·0
CM (f)	1090	2343	3·5	2·1	16·9	16·5	8·9	7·8
MM (f)	586	2338	2·2	1·4	14·3	14·3	5·0	3·6
LM (f)	863	2353	3·8	1·7	23·7	22·6	12·1	9·8
Average	6719	2107	3·1	1·8	18·6	17·3	9·2	7·2

phoneme models segment words into phonemes. We have observed many phoneme segmentation errors in the training data, and we consider this a weakness in hidden Markov modelling. By correcting the segment boundaries between phonemes in the training set we can improve the phoneme models and enhance the accuracy of the recognizer. Both duration and energy constraints are helpful in correcting such segmentation errors.

In the training data many segmentation errors are between low energy and high energy phonemes. For example, we have observed incorrect segment boundaries between stops and sonorants, between fricatives and sonorants, between affricates and sonorants and between breath noise and sonorants. (Note that stops, fricatives, affricates and breath have low energy, while sonorants have high energy.) Segment boundaries between low energy and high energy phonemes can be located accurately by forcing a more precise alignment of energy contours. One example of incorrect segment boundaries can be seen in the top spectrogram of Fig. 4 which shows Viterbi segmentation without any duration or energy constraints. In this example, the boundaries between /h/ and /ɪ/, and between /ɪ/ and /z/ in *his* are not right. In this instance, we can use the frame energy (C_0) to correct the segmentation errors. The segment boundaries

TABLE III. Comparison of average values of [log likelihood (correct word) − log likelihood (top choice incorrect word)] before and after duration constraints

Speaker	Log likelihood (correct) − log likelihood (incorrect)	
	No duration constraints	Duration constraints
LM	14·6	18·0
ML	20·1	29·8
NM	20·5	31·2

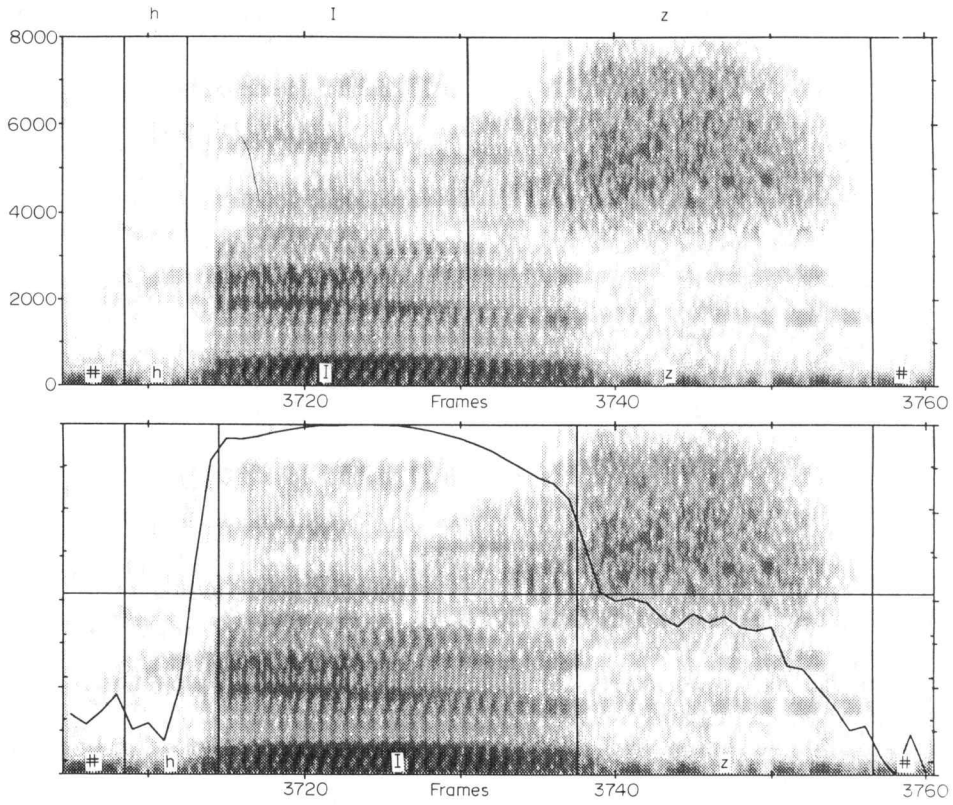

Figure 4. Example of improved segmentation in the training set using duration and energy constraints. In this example, the word spoken is *his*. The top spectrogram shows segmentation into phonemes without energy constraints. Notice that part of /h/ is mapped into /ɪ/, while a part of /ɪ/ is mapped into /z/. Bottom spectrogram shows segment boundaries after energy and duration constraints. Notice that the above segmentation errors have been corrected. The bottom spectrogram also shows the energy contour.

after consideration of energy thresholds are shown in the bottom spectrogram of Fig. 4. The energy contour is superimposed on the spectrogram. The energy increases rapidly from /h/ to /ɪ/, and it drops rapidly during the transition from /ɪ/ to /z/.

Because sonorants have similar energy contours we cannot use energy to correct segment boundaries between sonorants. However, we can use minimum duration constraints to correct segment boundaries between sonorants. In other words, the energy and duration constraints are complementary. The duration constraints are imposed in the same way as in recognition as outlined in Section 2. An example of duration constraints in training can be seen in Fig. 5, where we are able to refine the segment boundary between /ɛ/ and [l̩]. The top spectrogram shows the phoneme boundaries when no duration constraints are imposed, while the bottom spectrogram shows phoneme boundaries after duration constraints are imposed. Note that the start of the /ɛ/ has moved by one frame to satisfy the energy constraints, while the end has moved more than 40 ms from the start of /ɛ/ segment. This is because the optimal path with duration for /ɛ/ of 40 ms or more results in duration of /ɛ/ to be 80 ms. In Fig. 6, also, the segment

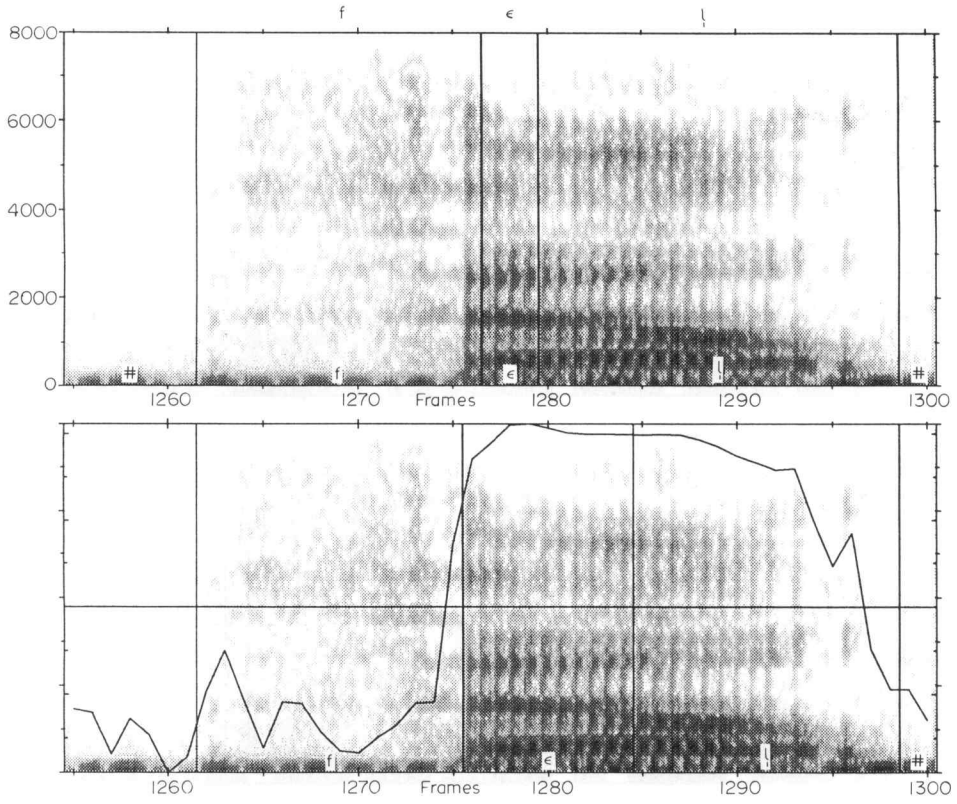

Figure 5. Example of improved segmentation in the training set using duration and energy constraints. The word spoken is *fell*. In this instance, the durations play a major role in correcting the segment boundary between /ɛ/ and [l̩]. Notice that, after duration constraints are applied, the resulting duration of /ɛ/ is actually longer than the allowed minimum.

boundary between /ɛ/ and [r̩] has been improved by imposing duration constraints. Here, also, before imposing duration constraints, /ɛ/ has a duration less than the minimum (top spectrogram in Fig. 6). However, after application of duration constraints, the duration for /ɛ/ is much longer than the minimum.

We have observed that the C_0 thresholds required to achieve the best segment boundaries are speaker-dependent. The thresholds are optimized iteratively from the training data by using a set of initial threshold assignments, looking at the resulting segmentation,[7] and then selecting new threshold values to improve the segmentation. At this point, we are not looking at automating the process for optimizing the thresholds. This is a pilot study to see if imposing reasonable energy constraints would improve recognition accuracy. The thresholds used for speakers DS and AM are compared in Table IV.

Let us discuss some of the thresholds given in Table IV in more detail. In all vowels (except /ɪ/ and /ə/), the lowest energy in any frame inside a vowel can drop as low as

[7] We only look at words whose likelihoods have dropped significantly.

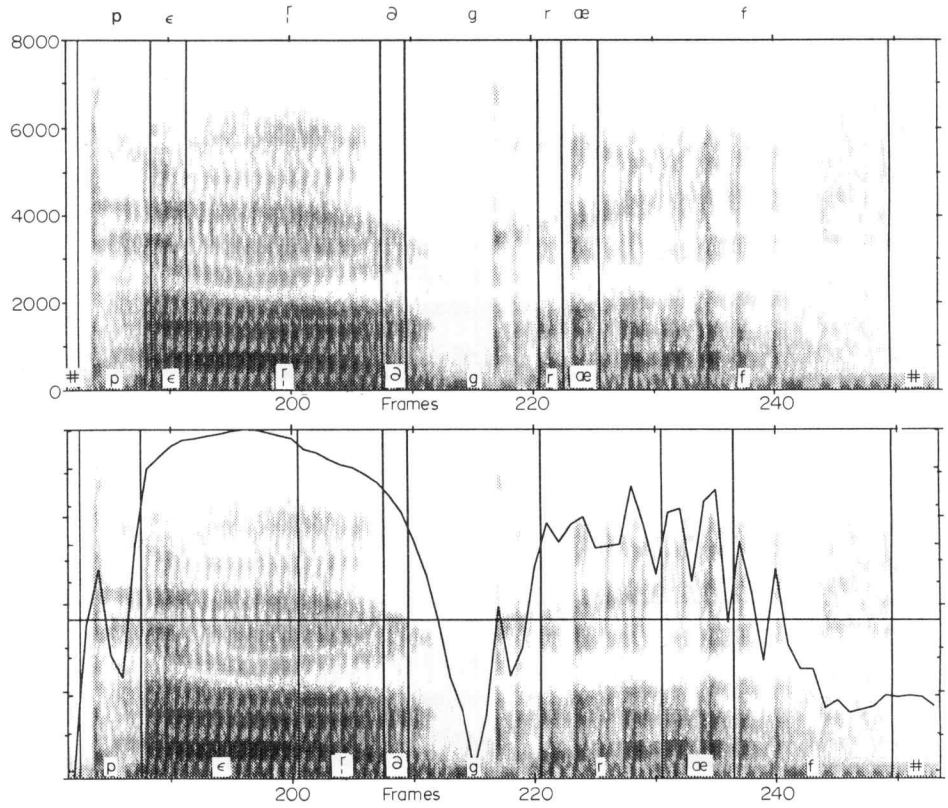

Figure 6. Example of improved segmentation in the training set using duration and energy constraints. The word spoken is *paragraph*. This is an example where the vocal fry in the last syllable /græf/ reduces the effectiveness of energy constraints. Only duration constraints are effective in improving the segment boundaries.

15 dB below the *peak energy*, where the peak energy in the word is that of the frame with the highest energy. The lowest energy in /ɪ/ and /ə/ can be as low as 20 dB below the peak energy for speaker DS, and as low as 25 dB below the peak energy for speaker AM. Note that, as the threshold constraints get weaker, the effectiveness of thresholds to correct segmentation errors is reduced. In fact, the energy threshold is only marginally effective in correcting segmentation errors between the final vowel and a non-sonorant phoneme for speaker AM. This is primarily due to the strong vocal fry observed in the final syllable for this speaker. One example of the vocal fry can be seen in the syllable /græf/ of *paragraph* as shown in Fig. 6. Due to the vocal fry, the energy levels in the last vowel can drop as low as 35 dB below the peak energy level. Also, as is evident from Table IV, the energy constraints are only marginally effective in placing a segment boundary between a flap[8] and a sonorant segment.

The energy thresholds are applied during Viterbi segmentation of words into phonemes in the training set. Application of energy thresholds is very similar to the

[8] Any /t d/ occurring in intervocalic position where the following vowel is unstressed.

TABLE IV. C_0 constraints for phonemes for speakers DS and AM. Note that all the constraints are relative to the peak energy in the word

Energy minima		
Phonemes	Min. relative to peak (DS) (dB)	Min. relative to peak (AM) (dB)
/aj aw ɔj ɑ ɔ i e ɛ æ ɑr ʌ u ʊ o ɔr/	−15	−15
/ɪ ə/	−20	−25
Final vowel	−15	−35
/j w l r/	−30	−35
Energy maxima		
Phonemes	Max. relative to peak (DS) (dB)	Max. relative to peak (AM) (dB)
/n m ŋ/	0	0
Initial and final breath	−15	−15
t-flats and d-flaps	−1	−1
Initial and medial stops, fricatives and affricates	−5	−5
Final stops, fricatives and affricates (threshold at the beginning of the phoneme)	−10	−10
Final stops, fricatives and affricates (threshold everywhere except the beginning)	−5	−5

application of duration constraints as explained in Section 2. In applying energy constraints, a transition in the phoneme model can only be taken if the C_0 value for the observed frame is within the specified constraints for the corresponding phoneme model. Such a restriction gives us Viterbi segmentation of the word into phonemes with the specified constraints on C_0 (see Table IV for C_0 constraints). The new phoneme models are trained using this segmented training data.

We have applied energy thresholds during training only on two speakers (DS and AM). Recognition results for the two speakers using the new phoneme models are shown in Table V. The recognition accuracy for both the speakers has improved significantly. The acoustic recognition errors have been reduced by 13%, while the recognition errors after the language model have been reduced by 23%. For the speakers DS and AM, the combined effect of duration constraints in recognition, and duration and energy constraints in training is a 16% reduction in acoustic recognition errors and a 39% reduction in errors after the language model.

5. Conclusions

We have used minimum duration constraints and energy thresholds for phonemes to

TABLE V. Recognition results for two speakers with and without energy (C_0) and duration constraints during training. During recognition, duration constraints are used

	Acoustic recog.				Errors after lang. model	
	Search errors		Recog. errors			
Speaker	No C_0 (%)	C_0 (%)	No C_0 (%)	C_0 (%)	No C_0 (%)	C_0 (%)
DS	3·1	3·0	21·9	18·8	10·4	7·5
AM	1·8	2·2	31·0	27·3	12·2	10·0
Avg	2·5	2·6	26·5	23·1	11·3	8·8

increase the recognition accuracy of our large vocabulary recognizer. Minimum duration constraints force the phoneme models to map to acoustic segments longer than the duration minima for the phonemes. Such minimum duration constraints result in significant lowering of likelihoods of many incorrect word choices, improving the accuracy of both acoustic recognition and recognition with the language model. The number of word recognition errors with the language model is reduced from 620 to 481 (9·2 to 7·2%). The minimum phoneme duration constraints applied are the same across all nine speakers tested.

Energy thresholds improve segment boundaries between phonemes in the training set. Training the phoneme models with improved segmentation for phonemes results in increased acoustic recognition accuracy and in enhanced recognition accuracy after the language model. Not all segment boundaries between phonemes can be improved by using energy thresholds. Only the segment boundaries between stops and vocalic segments, between fricatives and vocalic segments, between affricates and vocalic segments and between breath and vocalic segments can be improved. Also, the energy thresholds used are speaker-dependent and cannot be applied blindly across all speakers. Vocal fry significantly reduces the effectiveness of the energy thresholds. The energy thresholds are tighter for the speaker with very little vocal fry than for the speaker exhibiting a lot of vocal fry. The combined effect of minimum duration constraints in recognition and minimum duration and energy constraints in training is to reduce the word error rate by 40% (from 14·3 to 8·8%) with the language model. To achieve speaker-independent thresholds, one would have to examine additional data from numerous speakers.

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

References

Bush, M. A. & Kopec, G. E. (1987). Network-based connected digit recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP 35**, 1401–1413.

Deng, L., Lennig, M., Gupta, V., Kenny, P., Seitz, F. P. & Mermelstein, P. (1990). The acoustic recognition component of the INRS-Telecom 86,000-word speech recognizer. In *Proceedings of the 1990 IEEE International Conference on Acoustics, Speech and Signal Processing*. Albuquerque, New Mexico, pp. 741–744.

Gupta, V., Lennig, M. & Mermelstein, P. (1988). Fast search strategy in a large vocabulary word recognizer. *Journal of the Acoustical Society of America*, **84**, 2007–2017.

Gupta, V., Lennig, M. & Mermelstein, P. (1992). A language model for very large vocabulary speech recognition. *Computer Speech and Language*, **6**, 331–344.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, **64**, 532–556.

Kopec, G. E. & Bush, M. A. (1985). Network-based isolated digit recognition using vector quantization. *IEEE Transactions, Acoustics, Speech and Signal Processing* **ASSP 33**, 850–867.

Levinson, S. E., Rabiner, L. R. & Sondhi, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell Systems Technical Journal*, **62**, 1035–1074.

Nilsson, N. J. (1980). *Principles of Artificial Intelligence*. Tioga Publishing Co., Palo Alto, California.

Seitz, F., Gupta, V., Lennig, M., Kenny, P., Deng, L. & Mermelstein, P. (1990). A dictionary for a very large vocabulary word recognition system. *Computer Speech and Language*, **4**, 193–202.

Soong, F. K. (1989). A phonetically labeled acoustic segment (PLAS) approach to speech analysis–synthesis. *Proceedings of the 1989 IEEE International Conference on Acoustics, Speech and Signal Processing*. Glasgow, pp. 584–587.