Phoneme-in-Context Modeling for Dragon's Continuous Speech Recognizer

Paul Bamberg and Laurence Gillick
Dragon Systems, Inc.
90 Bridge St., Newton MA 02158

Abstract

For large-vocabulary continuous speech recognition, the goal of training is to model phonemes with enough precision so that from the models one could reconstruct a sequence of acoustic parameters that accurately represents the spectral characteristics of any naturally-occurring sentence, including all coarticulation effects that arise either between phonemes in a word or across word boundaries. The aim at Dragon Systems is to collect and process enough training data to accomplish this goal for all of natural spoken English rather than for any one restricted task.

The basic unit that must be trained is the "phoneme in context" (PIC), a sequence of three phonemes accompanied by a code for prepausal lengthening. At present, syllable and word boundaries are ignored in defining PICs.

More than 16,000 training tokens, half isolated words and half short phrases, were phonemically labeled by a semi-automatic procedure using hidden Markov models. To model a phoneme in a specific context, a weighted average is constructed from training data involving the desired context and acoustically similar contexts.

For use in HMM continuous-speech recognition, each PIC is converted to a Markov model that is a concatenation of one to six node models. No phoneme, in all its contexts, requires more than 64 distinct nodes, and the total number of node models ("phonemic segments") required to construct all PICs is only slightly more than 2000. As a result, the entire set of PICs can be adapted to a new speaker on the basis of a couple of thousand isolated words or a few hundred sentences of connected speech.

The advantage of this approach to training is that it is not task-specific. From a single training database, Dragon Systems has constructed models for use in a 30,000-word isolated-word recognizer, for connected digits, and for two different thousand-word continuous-speech tasks.

1. Introduction

The nature of the training process for a speech-recognition system changes radically once the size of the vocabulary becomes larger than the number of words for which a user is willing to provide training tokens. Below this threshold, it is reasonable to make an independent model for each word in the vocabulary. Such a model, based on data from that word and no others, can in principle capture all the acoustic-phonetic subtleties of the word, even though the phonetic spelling of the word is not even used in constructing the model.

For continuous speech recognition, the quantity of data required for complete training grows much more rapidly than vocabulary. In the simple case of a recognizer for three-digit strings, for example, each digit should at a minimum be trained in initial, medial, and final position, while for optimum performance all digit triples should be included in the training data.

The approach to training at Dragon Systems has been to regard the recognition task as all of natural English, whether isolated words or connected speech. We have developed a training database from which we have constructed recognition models for a 30,000 word isolated-word recognizer and for two different 1000-word connected speech tasks. All these recognition models are based on the same set of "phonemes in context."

2. Phonemes in Context

A speaker of English, given a phonemic spelling of an unfamiliar word from a dictionary, can pronounce the word recognizably or recognize the word when it is spoken. On the other hand, it is impossible to put together an "alphabet" of recorded phonemes which, when concatenated, will sound like natural English words. Speakers of English apply a host of duration and coarticulation rules when combining phonemes into words and sentences, and they employ the same rules in recognizing spoken language. It comes as a surprise to most speakers, for example, to discover that the vowels in "will" and "kick", which are identical according to dictionary pronunciations, are as different in their spectral characteristics as the vowels in "not" and "nut", or that the vowel in "size" has more than twice the duration of the same vowel in "seismograph".

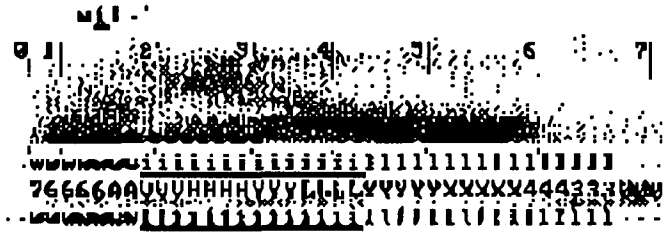

Figure 1. "Will"

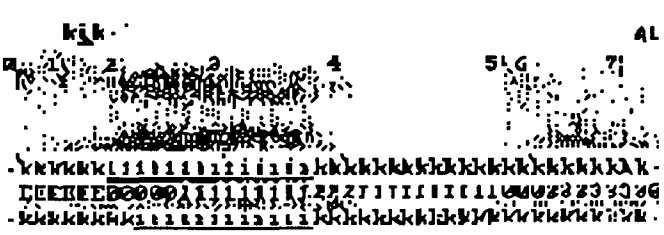

Figure 2. "Kick"

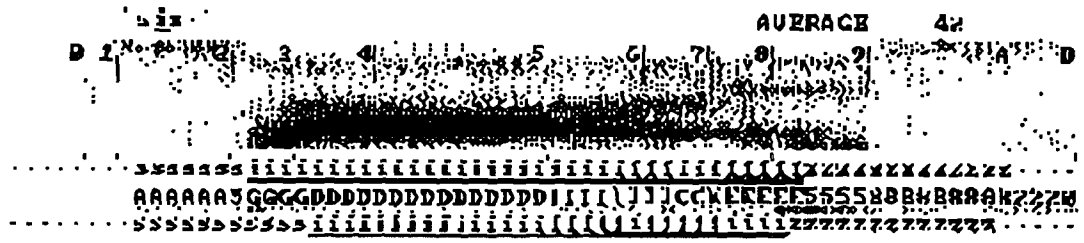

Figure 3. "Size" (prepausal lengthening of "i")

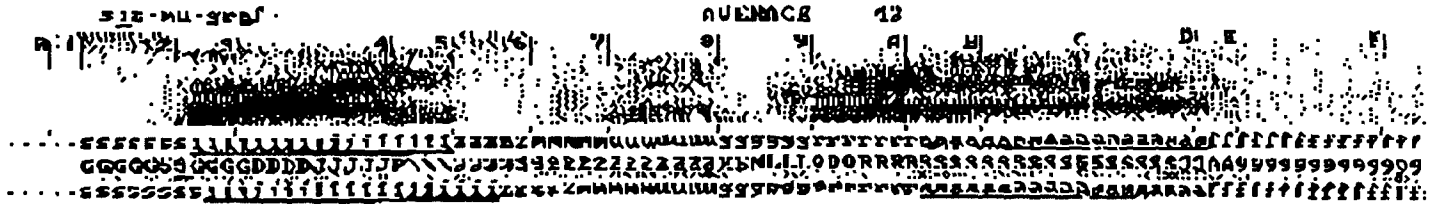

Figure 4. "Seismograph" (no prepausal lengthening of "i")

In the Dragon Systems family of speech recognizers, the fundamental unit of speech to be trained is the "phoneme in context" (PIC)[3]. Ultimately the defining property of a PIC is that by concatenating a sequence of PICs for an utterance one can construct an accurate simulated spectrum for the utterance. In the present implementation, a PIC is taken as completely specified by a phoneme accompanied by a preceding phoneme (or silence), a succeeding phoneme (or silence), and a duration code that indicates the degree of prepausal lengthening. To restrict the proliferation of PICs, syllable boundaries, even word boundaries, are currently ignored

The set of phonemes is taken from *The Random House® Unabridged Dictionary*. The stress of each syllable is regarded as a property of the vowel or syllabic consonant in that syllable. Excluding pronunciations which are explicitly marked as foreign, there are 17 vowels, each with three possible stress levels, plus 26 consonants and syllabic consonants.

A duration code of 3 indicates absence of prepausal lengthening. This will always be the case except in the last two syllables of an utterance.

A duration code of 6 indicates prepausal lengthening to approximately twice the normal duration. This occurs for the vowel in the final syllable of an utterance and for any consonant that follows that vowel, unless the vowel is followed by one of the unvoiced consonants k, p, t, th or ch. For example, in the word "harmed" every PIC except the one for the initial 'h' will have a duration code of 6.

A duration code of 4 indicates prepausal lengthening by a factor of approximately 4/3. This occurs in two cases:

- In the final syllable when the vowel is followed by k, p, t, ch, or th: for example, in both PICS of "at" and in the last three PICS of "bench".

- For consonants that precede the vowel in the final syllable: for example, the 's' in "beside".

PICs contain almost enough information to predict the acoustic realization of a phoneme. For example, the PIC for 't' is different in the word "mighty" (where the 't' is usually realized as a flap) and in the phrase "my tea" (where the 't' is clearly aspirated). This distinction is made, even though syllable and word boundaries, are ignored, because the stress of the following vowel is part of the context. Similarly, PICs capture the information that the final 't' in "create" (preceded by a stressed vowel) is more strongly released that in "probate" (preceded by an unstressed vowel), that the 's' in "horseshoe" is realized as an "sh", that the 'n' in "San Francisco" or "NPR" is realized almost like an 'm', and that the 'n' in "month" or "in the" is the dental allophone of 'n'.

3. Selection of PICs for Training

For isolated-word recognition, one could in principle enumerate all PICs by processing phonetic spellings for all the words in an unabridged dictionary. For the 25,000 words in the DragonDictate recognizer, there are approximately 30,000 PICs. A subset of 8,000 words can be chosen that includes all but about 1,000 of these PICs, most of them occurring in only a single word. Increasing the vocabulary size to 64,000 words would increase the number of PICS only slightly, to about 32,000.

For connected speech the goal of including all possible PICs is unachievable because of the wide variety of PICs that can arise through coarticulation across word boundaries. For example, the sentence "Act proud when you're dubbed Gareth" contains the PICs "ktp" and "bdg", neither of which occurs in any common English word. A further complication is that each PIC in a final syllable can occur in a sentence either with or without prepausal lengthening.

For the sort of connected-speech task which can be carried out in close to real time on today's microcomputers, the majority of PICs already arise only as a result of coarticulation across word boundaries. The 1023 pronunciations for the 842 words in the mammography vocabulary that is used for research at Dragon Systems include 2681 PICs. A set of 3000 sentences using this vocabulary includes only 1929 of these PICs, plus another 4610 that are not present in the isolated words. A different set of 3000 sentences, reserved for testing, includes yet another 1326 new PICs. Among the 121 PICs, not present

in isolated words, that occur 100 or more times in the sentences are the vowel in the suffix "ation" without prepausal lengthening, the dental "n" of "in the" and "on the", and the "zs" combination of "is seen".

The Dragon Systems training set currently includes about 8000 isolated words and about 8000 short phrases, each limited in duration to about 2.4 seconds. Although the total number of words in the training set is no greater than in the 6000 mammography sentences, the training set includes 37,423 distinct PICs. It is still far from complete. For example, a a set of 800 phrases drawn from a Hemingway short story and a newspaper article on parallel processing includes slightly more than 1000 PICs that were not in the training set (most, however, occurred only once).

The problem of finding the smallest training vocabulary that includes a given set of PICs is probably NP-complete. Still, it is easy to find a reasonably good approximation to the solution of the problem. In 6000 isolated words one can include about 22,000 different PICs. Beyond this point it becomes difficult to find words that include more than one or two new PICs, but short phrases of diverse text which contain three or more new PICs are still easy to find. By using such phrases to enlarge the training vocabulary, we hope to acquire training data for 50,000 PICs within the next year.

4. Modeling PICs by Phonemic Segments

A "vocabulary" of 50,000 independent PICs would be no more manageable than a vocabulary of 50,000 independent isolated words, but PICs are not independent. Most of the PICs for a stop consonant, for example, involve an identical segment of silence, for example, while all PICs for the sibilant "s" are characterized by the absence of low-frequency energy. One can hope, therefore, to represent the thousand or so PICs that represent the same phoneme in various contexts in terms of a much smaller number of "phonemic segments". For phonemes that exhibit a great deal of allophonic variation, such as "t", "k", and schwa, as many as 64 different segment models may be required, while for phonemes like "s" and "sh" that are little influenced by context, as few as ten may suffice. For the complete set of 77 phonemes used in English, slightly more than 2000 segment models suffice. In [4], an approach to modeling allphonic models using a small number of distributions was described. Similarly, in [5], an alternate way of performing parameter tying across distinct triphones using a triphone clustering procedure was described.

A phonemic segment can be characterized in two alternative ways. At the simpler level, it can be regarded as a fragment of the sort of acoustic data that would be generated by the "front end" of a speech-recognition system. In the case of the current Dragon recognizer, this is nothing more than a simulated spectrum based on an amplitude parameter and several spectral parameters. At a more sophisticated level, a phonemic segment includes enough information to generate a probability distribution for use in hidden Markov modeling. For the current Dragon recognizer, this requires calculation of the absolute deviation from the mean, as well as the mean for each acoustic parameter. The same distinction between what will be called a "spectral model" and what will be called a "Markov model" applies also to continuous parameters that have no direct spectral interpretation (cepstral parameters, for example), or to discrete parameters. In the following discussion, the term "spectrum" should be interpreted to mean any sequence of parameters that results from processing a speech waveform, while "Markov model" should be interpreted as a random process capable of generating such sequences.

One may think of a PIC as a probabilistic model for a portion of a speech spectrogram corresponding to a single phoneme. The problem of representing this PIC as a sequence of phonemic segments is solved by hidden Markov modeling. The sequence may be from one to six segments in length, and the same segment may occur in more than one position in the sequence. There is no constraint on the order of segments within the sequence.Thus the model for a phoneme with n segments is represented by the diagram below.

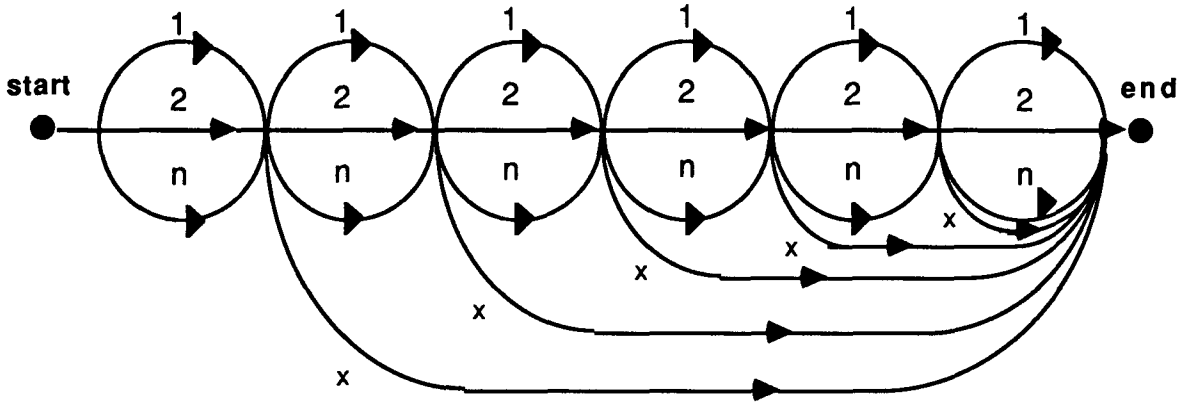

Figure 5. A Markov Model for a Single PIC

The arcs labeled 1, 2, ...n correspond to one or more frames of acoustic data corresponding to the single segment 1, 2, ...n. The arcs labeled x permit a given phoneme to have a sequence of fewer than six phonemes associated with it. These null arcs are assigned slightly higher transition probabilities than the arcs associated with phonemic segments.

Thus a PIC may be represented very compactly as a sequence of one to six pairs, each pair consisting of a phonemic segment and a duration.This sequence may be regarded as the best piecewise-constant approximation to the spectrogram.

For speaker adaptation, the phonemic segment is the basic unit. It is assumed that the representation of a PIC in terms of segments is valid for all speakers, so that adapting the small number of segments for a phoneme will have the effect of adapting the much larger number of PICs. Segment durations within a PIC can also be adapted, but only by acoustic data involving that particular PIC.

5. Labeling Training Data

To build a spectral model for a PIC, one must find one or more spectrograms that involve that PIC, then extract from these spectrograms the data for the phoneme in the desired PIC. Thus phonemically labeled training data are required.

Given a complete set of hidden Markov models representing PICs, the labeling problem could easily be solved by dynamic programming and traceback. This approach is the correct one to use for implementing adaptation, but it is inappropriate for training, since the labeled training data would be required in order to produce the PIC models in the first place. To do semiautomatic labeling with an incomplete set of phonemic segments and with no PIC models, a simpler scheme must be used, one which deals gracefully with the situation where PIC models have not yet been created and where some portions of spectrograms cannot yet be labeled.

The full Markov model for a word is a sequence of models for the phonemes of the word, starting and ending with silence. Silence is modeled, like any other phoneme, by a set of segments. Between the phoneme models are "transition nodes" with fixed transition probabilities that are chosen to be slightly lower than the typical probability for the best phoneme segment. Thus the model for "at" might be represented as follows:

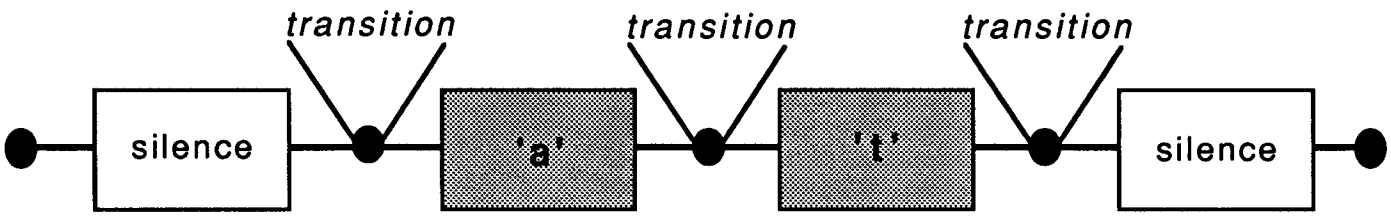

Figure 6.A Markov Model for "at".

Each box represents a phoneme model of one to six states, as described above.

Once the best path has been found by dynamic programming, traceback at the phoneme level assigns a start time and end time to each phoneme. If a complete set of phonemic segments has been constructed, the start time for each phoneme coincides with the end time for its predecessor phoneme. To the extent that there are acoustic segments that are not yet well modeled by any phonemic segment, the data that correspond to this segment will be assigned to an interphoneme transition.

The phoneme-level traceback is recorded within each training token. This makes it possible, without repeating the dynamic programming, to identify the portion of a given training token that correspond to a specified phoneme—an important step in locating training data for a specific PIC.

Traceback can also be performed at a lower level in order to determine the sequence of phonemic segments that corresponds to an individual PIC. The data thus assigned to a segment may then be used as training data for that segment to improve the estimates of the means and variances for the acoustic parameters of that segment.

The net effect of dynamic programming followed by traceback at the word level and at the phoneme level is to assign to each "frame" of acoustic data of the word a phoneme segment label, subject to the following constraints:

- Phonemes appear in the order specified by the pronunciation of the word.
- For each phoneme, there are no more than five transitions from one segment to another.
- Transition frames with no segment assignment may occur only between phonemes.

The process of labeling the training data is not completely automatic, but it becomes more and more nearly so as the set of phonemic segments increases in size. In practice, phonemic segments are initialized "by hand". On a spectral display of a training token, a sequence of frames is selected. The means and variances for the acoustic parameters of those frames provide the initial estimates for the segment parameters. Even in the absence of any previously labeled segments, it is a straightforward matter to initialize a set of segments that will provide a correct phonemic labeling of a single token, and these segments in turn prove useful in labeling other tokens. As more and more tokens are labeled in this manner, a set of segments develops that suffices to label a greater and greater fraction of new tokens, until eventually any new token can be labeled without the need for interphoneme transitions.

As new segments are created during the labeling process, occasionally the limit of 64 segments for a phoneme is reached. Whenever this occurs, the two segments that are most similar are automatically combined into a single segment.

Once a thousand or so training tokens have been labeled, transition segments that are more than about thirty milliseconds long become difficult to find. At this point the

best strategy is to label all the training tokens automatically, then to search for the longest transition segments and to use them to create new phonemic segments. This process can be iterated until no transition segments remain.

To make use of duration constraints in labeling, an alternative version of the dynamic programming is used which closely resembles the one used by Dragon's small-vocabulary recognition and training algorithm. To each phoneme in the word, an expected duration in milliseconds is assigned. To the extent that the actual duration of the speech assigned to that phoneme is less than or greater than the expected duration, a duration penalty is added to the dynamic programming score. The traceback is then determined both by acoustic match and by duration constraints. While a clear-cut phoneme boundary such as one before or after an 's' will be little affected by duration constraints, a boundary that is associated with almost no acoustic feature (between two stops, for example) will be assigned primarily on the basis of durations.

In order to estimate durations, the hypothesis is made that changing the left or right context of a phoneme has little effect on the duration of that phoneme except in the case where the context is silence. As stated above, the duration of the final 'l' in "all" ought to be the same as the duration of the final 'l' in "wheel", "bell", or other words where there is a clear formant transition into the "l". As another example, the 'p' and 't' in "opted" should each have a duration close to that of a single intervocalic stop.

For each PIC, an expected duration is determined by averaging together four quantities:

- the duration of the phoneme in the precise context specified by the PIC (which may occur only once in the training vocabulary).
- the duration of the phoneme with the specified left context and an arbitrary right context.
- the duration of the phoneme with the specified right context and an arbitrary left context.
- the duration of the phoneme with both left and right context arbitrary.

In no case, however, is a silence context substituted for a non-silence context or vice versa.

The semiautomatic labeling process described above has been under development for more than a year, with results that appear more and more satisfactory as the new phonemic segments are identified and duration estimates are improved. By using a set of about 2000 segments and imposing duration constraints on the dynamic programming, it is possible to achieve automatic phonemic labeling that agrees with hand labeling in almost every case and that is probably more consistent than hand labeling with regard to such difficult, arbitrary decisions as placing boundaries between adjacent front vowels or between glides and vowels. Most labels that a human labeler might question can be located by looking just at the small fraction of words for which the actual and expected duration of a phoneme differ significantly.

By exploring situations in which the expected durations of phonemes in correctly labeled words are systematically in error, it is possible to discover new duration rules which can be incorporated into more refined characterization of PICs. Each such rule, though, leads to an increase in the total number of PICs that must be trained.

6. Building Models for PICs

Given a sufficiently large quantity of training data, one can create an excellent model for a PIC by averaging together all examples of that PIC in the training vocabulary. For example, a model can be built for the phoneme "sh" in the context "ation" by averaging together the data labeled as "sh" in words such as "nation", "creation", and "situation". Unfortunately, the assumption of a large quantity of training data for each PIC is unrealistic. There are, for example, about 1500 contexts in the DragonDictate 25,000 word vocabulary, and many contexts in connected speech, for which even the current training set of 16,000 items provides no examples. For thousands of other PICs there is only a single example in the training set. Thus, in modeling a PIC, it is important to employ training data from closely related PICs.

In most cases the left context of a phoneme influences primarily the first half of the phoneme, while the right context influences primarily the second half. Furthermore, there are groups of phonemes which give rise to almost identical coarticulation effects: different stress levels of the same vowel, for example.

The general strategy for building a model for a phoneme in a given context is to compute a weighted average of all the data in the training vocabulary for the given phoneme in the desired context or any similar context. The weight assigned to a context depends upon how well it matches the desired context.

Weights are assigned separately for the left context and the right context, and two models are constructed. The first of these, where a high weight implies that the left context is very close to the desired left context (although the right context may be wrong) is used for the first half of the model. The second model, where a high weight implies that the right context is correct, is used for the second half of the model.

Each phoneme is assigned both to a "left context group" and to a "right context group". The phonemes in left context group should all produce similar coarticulation effects at the start of a phoneme, while those in the same right context group should produce similar effects at the end of a phoneme.

To build a model for a PIC, all examples of contexts similar to the desired PIC are extracted from the training vocabulary. Each context is assigned a "left weight" and a "right weight" according to the degree of match between the desired context in the PIC and the actual context in the training item.

From the data a weighted average of the durations is now computed. Tokens for which the duration is close to the average are doubled in weight, while those that are far from the average duration are halved in weight.

Finally all the examples of the desired phoneme are averaged together using a linear alignment algorithm which normalizes all examples so that they have the same length, then averages together acoustic parameters at intervals of 10 milliseconds. This procedure is carried out twice, once with left weights, once with right weights. The first half of the

"left model" and the second half of the "right model" are concatenated to form the final spectral model for the PIC.

Models for initial and final silence in each context are created by averaging the initial silence from training words that begin with the desired phoneme and by averaging the final silence from words that end with the desired phoneme.

Consider, for example, the comparatively unusual PIC "lak" (secondary stress on vowel, no prepausal lengthening). No word in the training set contains this PIC, although "Cadillacs" has the same PIC with prepausal lengthening. The "left" model, built from "implants", "overlap shadows", "eggplant", "Cadillacs", and "mainland gale", captures well the second formant transition between the "l" and the vowel. The "right" model captures the spectrum of the vowel before "k". The concatenated model has both features well modeled.

These spectral models for PICs are not yet hidden Markov models, since they include only the means of acoustic parameters, but not the variances. They also have no direct connection with phonemic segments. The final step in the training process is to convert them to adaptable Markov models that are based on phonemic segments.

Converting a spectral model for a PIC to a Markov model for that PIC employs the same algorithm that is used for labeling training data. Dynamic programming is used to determine the sequence of phonemic segments that has the greatest likelihood of generating the spectral model for the PIC. These phonemic segments become the nodes of the Markov model for the PIC. Concatenating the parameter means for the nodes, with each node given the duration determined by the dynamic programming, produces the optimal piecewise-constant approximation to the spectral model for the PIC.

The variances in the parameters for each phonemic segment correctly reflect the fact that each segment appears in many different PICs. Because training tokens are already averages of three utterances, the variances underestimate the variation in parameters from one utterance to another. To compensate for this, the variances in the phonemic segment models that are used for recognition are made somewhat larger than the estimates that arise from training.

Because the large number of PIC models are all constructed from about 2000 phonemic segments, they adapt quickly to a new speaker. The strategy for adaptation is simply to treat each utterance as if it were new training data. By dynamic programming the utterance is segmented into PICs, which are in turn subdivided in phonemic segments. The acoustic data assigned to each segment are used to reestimate the means and variance for that segment. For the mammography task, a set of 500 sentences to be used for adaptation has been developed that includes more than 90% of the PICs used by the recognizer. Since most phonemic segments occur in many different PICs, these 500 sentences provide diverse training data for almost all segments, sufficient to provide good estimates of their parameter means and variances for a new speaker. Estimates of segment durations for each PIC are also improved as a result of adaptation, although for this purpose the 500 sentences provide much less data.

To achieve real-time recognition of connected speech, a rapid-match algorithm is used to reduce the number of words for which full dynamic programming is carried out[1]. This algorithm requires models which incorporate accurate duration information and which capture coarticulation effects averaged over all possible contexts for a word. The training for the rapid-match model for a word makes use of a concatenation of spectral models for the PICs of the word, with a "generic speech" left context used for the first phoneme and a "generic speech" right context used for the last phoneme of the word.

7.Recognition Performance

The training strategy described here is intended to yield a set of PICs that will serve for any isolated-word or connected-speech recognition task in English. Testing has been carried out on four tasks, as follows.

1. The DragonDictate isolated-word recognition system uses 25,000 word models based on PICs and phonemic segments, built from the same database of training utterances that is used for connected speech. Recognition performance for two diverse texts, a short story by Hemingway and a newspaper article on parallel processing, was 83% correct on the first 500 words. After adaptation on 1500 words, performance rose to 89% correct for the speaker who recorded the training database. For two other speakers, performance without adaptation was dismal (45% for a male speaker, 18% for a female speaker), but it rose after adaptation on 2500 words to 87% for the male speaker and 85% for the female.

2. For connected digit recognition, the error rate on five-digit strings was less than half a percent for each of three different speakers after adaptation. Less than 0.2% of the training database consists of digit strings.

3. For the mammography task used in testing the real-time implementation of continuous-speech recognition[2] (842 words, 1023 distinct pronunciations), recognition was tested on a set of 1000 sentences which had not been used either in selecting training utterances or in determining which PICs should be modeled. Several hundred of the PICs in this test data did not occur in any of the "practice" sentences that had been for training; these PICs were modeled only by generic PICs in which an average was taken over all left and right contexts. About 15% of the training database consists of short phrases extracted from the 3000 practice sentences. On this task, whose perplexity is about 66, 96.6% of words were recognized correctly. Performance was slightly better on the "practice" sentences that had been used to construct the set of PICs to be modeled, sentences for which no generic PICs were required. Preliminary results indicate that after several hundred sentences of adaptation, performance close to this level can be achieved for other speakers.

4. As a test of performance on a connected-speech task which was not so heavily used in constructing the training database, recognition was carried out on the 600 training sentences of the Resource Management task using the word-pair grammar. This task has a perplexity of about 60, comparable to that of the mammography task. PICs were built from the same training database as described above, in which about 5% of the tokens are phrases based on the resource management vocabulary. Recognition performance was 97.3% correct on a per-word basis. For this task, as for the mammography "practice" sentences, all PICs had been modeled, so that no generic PICs were required.

References

[1] L. Gillick and R. Roth, "A Rapid Match Algorithm for Continuous Speech Recognition", *Proceedings of DARPA Speech and Natural Language Workshop*, June 1990 Hidden Valley, Pennsylvania.

[2] P. Bamberg et al., "The Dragon Continuous-Speech Recognition System: A Real-Time Implementation", *Proceedings of DARPA Speech and Natural Language Workshop*, June 1990 Hidden Valley, Pennsylvania.

[3] R. Schwartz et al., "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1985

[4] Bahl et al., "Large Vocabulary Natural Language Continuous Speech Recognition", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May, 1989

[5] K.F.Lee et al., "The Sphinx Speech Recognition System", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May, 1989