# SPEAKER INDEPENDENT PHONETIC TRANSCRIPTION OF FLUENT SPEECH FOR LARGE VOCABULARY SPEECH RECOGNITION

*S. E. Levinson*
*M. Y. Liberman*
*A. Ljolje*
*and*
*L. G. Miller*
AT&T Bell Laboratories
Murray Hill, New Jersey 07974

## ABSTRACT

Speaker independent phonetic transcription of fluent speech is performed using an ergodic continuously variable duration hidden Markov model (CVDHMM) to represent the acoustic, phonetic and phonotactic structure of speech. An important property of the model is that each of its fifty-one states is uniquely identified with a single phonetic unit. Thus, for any spoken utterance, a phonetic transcription is obtained from a dynamic programming (DP) procedure for finding the state sequence of maximum likelihood. A model has been constructed based on 4020 sentences from the TIMIT database. When tested on 180 different sentences from this database, phonetic accuracy was observed to be 56% with 9% insertions. A speaker dependent version of the model was also constructed. The transcription algorithm was then combined with lexical access and parsing routines to form a complete recognition system. When tested on sentences from the DARPA resource management task spoken over the local switched telephone network, phonetic accuracy of 64% with 8% insertions and word accuracy of 87% with 3% insertions was measured. This system is presently operating in an on-line mode over the local switched telephone network in less than ten times real time on an Alliant FX-80.

## INTRODUCTION

Though rarely explicitly stated, a fundamental assumption on which many speech recognition systems are implicitly based is that speech is literate. That is, it is a code for communication having a small number of discrete phonetic symbols in its alphabet. These symbols are, however, merely mental constructs and, as such, are not directly accessible but are, instead, observable only in their highly variable acoustic manifestation. It is also well-known but equally seldom expressed that a hidden Markov model comprises a finite set of discrete inaccessible states observable only via a set of random processes, one associated with each hidden state. When these two simple ideas are juxtaposed, it seems to us inescapable that the most natural representation of speech by a hidden Markov model is one in which the hypothetical phonetic symbols are identified with the hidden states of the Markov chain and the variability of the measurable acoustic signal is captured by the observable, state-dependent random processes.

The mathematical details of just such a model are given in [6]. Its application to a small-vocabulary continuous speech recognition system and a large-vocabulary isolated word recognition system are described in [7] and [8], respectively. Here we present a brief overview of the use of this approach in large vocabulary continuous speech recognition and some preliminary results of two experiments performed with it on the TIMIT [4] and DARPA [9] databases.

## THE MODEL

We have constructed two models, a 51 state model on which the speaker-independent phonetic transcription results are based, and a 43 state model on which the speaker-dependent recognition of sentences from the DARPA resource management task are founded. The 51 states in the first model correspond to 51 of the phonetic symbols used in the standard transcriptions of the TIMIT sentences. The 43 states of the second model are associated with the 43 symbols used in the pronunciation guide of the Collins English dictionary [1]. The phonetic units are listed in figure 1. Flap and closure units are not included in the 43 state model.

| STATE | TIMIT | COLLINS | EXPLANATION |
|---|---|---|---|
| 1 | h# pau | + | silence |
| 2 | eh | e | bet |
| 3 | ao | > | bought |
| 4 | aa | @ | cot |
| 5 | uw ux | U | boot |
| 6 | er | R | bird |
| 7 | ay | I | bite |
| 8 | ey | A | bait |
| 9 | aw | W | now |
| 10 | ax | & | schwa |
| 11 | ih | i | bit |
| 12 | ae | a | bat |
| 13 | ah | ^ | butt |
| 14 | uh | u | book |
| 15 | oy | Y | boy |
| 16 | iy | E | beat |
| 17 | ow | O | boat |
| 18 | axr | # | diner |
| 19 | l | l | led |
| 20 | el | ~ | bottle |
| 21 | r | r | red |
| 22 | w | w | wet |
| 23 | y | y | yet |
| 24 | hh hv | h | hay |
| 25 | s | s | sister |
| 26 | sh | S | shoe |
| 27 | z | z | zoo |

| | | | |
|---|---|---|---|
| 28 | zh | Z | measure |
| 29 | ch | C | church |
| 30 | jh | J | judge |
| 31 | th | T | thief |
| 32 | dh | D | they |
| 33 | f | f | food |
| 34 | v | v | verve |
| 35 | m em | m | mom |
| 36 | n en | n | nun |
| 37 | ng eng | N | sing |
| 38 | nx | [ | nasal flap |
| 39 | p | p | pop |
| 40 | t | t | tot |
| 41 | k | k | kick |
| 42 | pcl | 1 | p closure |
| 43 | tcl | 2 | t closure |
| 44 | kcl | 3 | k closure |
| 45 | dx | ] | alveolar flap |
| 46 | b | b | bob |
| 47 | d | d | dad |
| 48 | g | g | gag |
| 49 | bcl | 4 | b closure |
| 50 | dcl | 5 | d closure |
| 51 | gcl | 6 | g closure |

Figure 1: Phonetic Units and Symbols

Both models are of the same form, CVDHMM, as described in the reference cited earlier. The state transition matrices define ergodic Markov chains and weakly capture the phonotactic structure of English. The acoustic measurements are represented by 26-dimensional Gaussian density functions. The first twelve coordinates are LPC based cepstra; the second twelve, delta-cepstra [2], and the last two, log energy and its time-derivative, respectively. The temporal structure of the acoustic signal is reflected in the durational densities which are of the two-parameter gamma family. Because of the presence of the durational densities, self-transitions are forbidden.

## PARAMETER ESTIMATION

The parameters for both the 51 and 43 state models were estimated in the same way although on different training data. In both cases, the state transition matrix was computed from bigram statistics extracted from the Collins dictionary. No attempt was made to count bigrams resulting from word junctures. Also, in both cases, the respective databases were segmented by hand and labeled with respect to the appropriate phonetic alphabet. Acoustic observations were sorted into sets corresponding to the phonetic symbols. The necessary parameters, spectral means and covariances and durational means and standard deviations, were then calculated for each set independently. No parameter optimization was applied to these estimates.

The 51 state speaker-independent model was trained on 4200 sentences of TIMIT data. Ten different sentences were selected from each of 402 different speakers. The 43 state speaker-dependent model was trained on one reading of the 450 sentences in the TIMIT phonetically balanced list by a single male speaker. These utterances were recorded over the local switched telephone network with a conventional telephone handset.

At this writing, we have yet to train a speaker-independent model using the DARPA training material. Although we expect to do so, we are concerned about its utility since the phonetic contexts in this database are rather restrictive compared with those of the TIMIT sentences.

## PHONETIC TRANSCRIPTION

Phonetic transcription is accomplished by means of a DP technique for finding the state sequence that maximizes the joint likelihood of state, duration and observation sequences. The details of this algorithm are given in [7]. Note that this procedure makes no use of lexical or syntactic structure. The algorithm runs in approximately twice real time on an Alliant FX-80.

## EXPERIMENTAL RESULTS ON TRANSCRIPTION

The transcription algorithm was tested on 180 sentences from the TIMIT database. Neither the sentences nor the speakers were used in the training. Transcription accuracy was determined by computing the Levenshtein distance between the derived transcription and the standard transcription supplied with the database. By this measure, the 51 state model yielded a phonetic recognition rate of 56% with a 9% insertion rate. The 43 state model resulted in a 64% recognition rate with an 8% insertion rate on 48 sentences from the DARPA task collected from the male speaker on whose speech the model had been trained.

The reader should bear in mind that these are the very first experiments performed with this system. We fully expect that the performance will improve greatly as a result of refinements we are presently making to the model. These include accounting for coarticulation, making the durational densities more faithful and using parameter reestimation techniques.

## THE SPEECH RECOGNITION SYSTEM

The phonetic transcription algorithm described above is the first stage of a complete speech recognition system. The architecture of the system is unchanged from that described in [8] but the details of the lexical access procedure and the parser are utterly different from those given in the reference.

The lexical access procedure is simply that of computing the likelihood of every word in the lexicon over every sub-interval of the observation sequence. We define the likelihood of a word on an observation sub-sequence to be the joint likelihood of the standard phonetic transcription for that word as given in the lexicon and the phonetic transcription of that subsequence provided by the transcription algorithm. Because the standard transcription need not have the same length as the one computed for an arbitrary observation sub-sequence, the calculation is carried out by means of a DP algorithm. Note that this procedure is synchronized at the segment rate, not the frame rate.

The parser takes as input, the word lattice constructed by the lexical access procedure and finds the well-formed sentence of maximum likelihood. Here, well-formed means with respect to the strict DARPA resource management task grammar. This is a finite state grammar having 4767 states, 60433 state transitions, 90 final states and a maximum entropy of 4.4 bits/word. The parser itself is yet another DP algorithm. The search it effects is not pruned in any way.

The system has been tested in an on-line mode over the switched local telephone network. Under these conditions, we obtained an 87% correct word recognition rate and a 3% insertion rate. On an Alliant FX-80, a sentence is recognized in less than ten times real time. A sample of the recognizer output is shown in figure 2.

PHONETIC TRANSCRIPTION: h@riRriEZUzpl>grUDENw^ndTWz&ndSEdED&nd>rTtUsiZ &kObS&n

DURATIONS:
5 5 7 4 8 8 7 5 10 17 9 12 4 6 13 8 8 7 6 9 7 6 6 7 4
9 19 10 3 6 3 11 17 5 12 4 4 5 3 9 6 7 6 7 14 5 8 3 10 12
3 13 7 5

LOG LIKELIHOOD = 0.23880190715663E+04

| POSITION | BEGIN | END | STATE | LOG LIKELIHOOD | WORD |
|---|---|---|---|---|---|
| 1 | 49 | 53 | 19 | 0.2250650E+02 | ocean |
| 2 | 42 | 48 | 394 | 0.2147619E+02 | pacific |
| 3 | 37 | 41 | 344 | 0.1887782E+02 | north |
| 4 | 35 | 36 | 265 | 0.1787559E+02 | the |
| 5 | 34 | 34 | 378 | 0.1733334E+02 | in |
| 6 | 30 | 33 | 299 | 0.1590989E+02 | feet |
| 7 | 24 | 29 | 926 | 0.1379514E+02 | thousand |
| 8 | 21 | 23 | 838 | 0.1118440E+02 | one |
| 9 | 18 | 20 | 758 | 0.1093698E+02 | than |
| 10 | 13 | 17 | 691 | 0.9208550E+01 | longer |
| 11 | 9 | 12 | 623 | 0.6723166E+01 | ships |
| 12 | 6 | 8 | 557 | 0.4362227E+01 | any |
| 13 | 3 | 5 | 513 | 0.3019794E+01 | there |
| 14 | 1 | 2 | 491 | 0.1371470E+01 | are |

RECOGNIZED SENTENCE: are there any ships longer than one thousand feet in the north pacific ocean

LOG LIKELIHOOD = 0.22506502151489E+02

RECOGNITION TIME = 49.78 CPU-SECONDS

Figure 2: Sample of Sentence Recognition Results

## CONCLUSION

We have presented some very early results of experiments on phonetic transcription and recognition of fluent speech based on a novel use of a hidden Markov model. While our error rates are substantially higher than those achieved by more conventional systems [5,3,10], we believe that by improving the acoustic/phonetic model - the only adjustable part of the system - results comparable to those obtained by other investigators can be realized.

**References**


[1] Hanks, P., ed., *Collins Dictionary of the English Language*, Collins, London, 1972.

[2] Juang, B. H., Rabiner, L. R. and Wilpon, J. G., "On the use of Bandpass Liftering in Speech Recognition", IEEE Trans. Acoust. Speech and Signal Processing, ASSP-35 (7), pp. 947-954, July, 1987.

[3] Kubala, F. et al., "Continuous Speech Recognition Results of the BYBLOS System on the DARPA 1000-Word Resource Management Database", Proc. ICASSP-88, New York, NY, pp. 291-294, April, 1988.

[4] Lamel, L. F., Kassel, R. H. and Seneff, S., "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus", Proc. DARPA Speech Recognition Workshop, Palo Alto, CA, pp. 100-109, Feb., 1986.

[5] Lee, K. F. and Hon, H. W., "Large Vocabulary Speaker-Independent Speech Recognition System using HMM", Proc. ICASSP-88, New York, NY, pp. 123 126, April, 1988.

[6] Levinson, S. E., "Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition", Computer Speech and Language 1 (1), pp. 29-45, 1986.

[7] Levinson, S. E., "Continuous Speech Recognition by means of Acoustic-Phonetic Classification Obtained from a Hidden Markov Model", Proc. ICASSP-87, Dallas, TX, pp. 93-96, April, 1987.

[8] Levinson, S. E., Ljolje, A. and Miller, L. G., "Large Vocabulary Speech Recognition using a Hidden Markov Model for Acoustic Phonetic Classification", Proc. ICASSP-88, New York, NY, pp. 505-508, April, 1988.

[9] Price, P., Fisher, W., Bernstein, J. and Pallett, D., "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition", Proc. ICASSP-88, New York, NY, pp. 651-654, April, 1988.

[10] Pieraccini, R., Lee, C. H., Rabiner, L. R. and Wilpon, J. G., "Some Preliminary Results on Speaker Independent Recognition of the DARPA Resource Management Task", in this proceedings.