# Continuous Speech Recognition from a Phonetic Transcription

*S. E. Levinson*
*A. Ljolje*
*L. G. Miller*

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

## 1. Introduction

A long-standing and widely accepted linguistic theory of speech recognition holds that natural spoken messages are understood on the basis of an intermediate representation of the acoustic signal in terms of a small number of phonetic symbols. The traditional linguistic theory is very attractive for several reasons. First, it provides a natural way to partition the process of communication by spoken language into distinct acoustic, phonetic, lexical and syntactic sub-processes. Second, it provides for a reduction in bandwidth at each successive stage of the process. And, finally, it seems to be reflected in the development of written language. It is thus not surprising that this seminal idea formed the basis for several early speech recognition machines [1,2,3,4].

In this report we offer what we believe to be the simplest and most direct expression of the linguistic theory in a working speech recognition system. The present system is the culmination of a succession of experiments conducted over the past three years. The method of acoustic phonetic mapping is described in [5], and results of its application to speaker-dependent recognition of fluently spoken digit strings are given in [6]. Next, a new method of lexical access was devised and applied to the problem of speaker-dependent recognition of isolated words from a large vocabulary [7] and sentences composed of them [8]. Attention was then turned to speaker-independent phonetic transcription [9, 10] which was then used in an early account of speaker independent recognition of fluent speech from the 991 word DARPA [11] resource management task [12].

In its present form, our speech recognition system uses a particular kind of hidden Markov model in conjunction with an appropriate dynamic programming algorithm to accomplish the acoustic-to-phonetic mapping. This part is not constrained by lexical or syntactic considerations and is thus vocabulary and task independent. Word recognition is then easily treated as a classical string-to-string editing problem which is solved by a two-level dynamic programming algorithm, the lower level of which performs lexical access while the upper level performs the parsing function.

Our account of the present speech recognition system is given in the following order. We first give an overview of the system at the block diagram level. This is followed by a detailed description of each of the component blocks, the acoustic phonetic model, the phonetic decoder and, finally, the lexical access and parsing techniques which, because they are so closely coupled, are treated as a unit. This is followed by an account of our experimental results and an interpretation of them.

To summarize our results, on the DARPA resource management task with the perplexity 9 grammar, we attained 88% correct word recognition with 3% insertions yielding a word accuracy of 85%. Phonetic transcription accuracy was assessed by resynthesizing directly from the phonetic transcription. In a few informal listening tests, we judged the word intelligibility rate to be approximately 75%.

The word accuracy of our system is not as good as that obtained on exactly the same data by several other conventional systems [13, 14, 15, 16]. However, we believe that a few correctable shortcomings of the existing system are responsible for the disparity. We hope to make the necessary changes in the near future.

## 2. The System

Acoustic signal processing is an autocorrelation based linear predictive analysis. The LPC's are transformed into cepstral coefficients at a centisecond frame rate. The phonetic decoding module is a dynamic programming algorithm applied to a 47-state ergodic semi-Markov model. There are two very important points to be made regarding this stage of processing. First, no lexical or syntactic information of any kind is available to the phonetic decoder. Second, once the decoding is accomplished, the acoustic signal is discarded. All that remains is its phonetic transcription and the duration, in centiseconds, of each phonetic unit in that transcription.

The lexical access and parsing functions are conceptually separate but are combined here in a two-level dynamic programming algorithm. The lower level is the lexical part while the upper level accomplishes the grammatical analysis. The two are intricately coupled. The DP algorithm simply performs a string-to-string editing in which the error-ridden phonetic transcription is mapped into sentences of conventional orthography. The lexicon used simply gives the phonetic transcription of each vocabulary word pronounced in citation form. The grammar is a strict right linear grammar with no null productions.

The entire system is implemented in FORTRAN-77 and runs on an Alliant FX-80. Because the phonetic decoding and lexical access stages have a high degree of intrinsic parallelism, we can exploit the architecture of the FX-80 to full advantage resulting in an execution time of 15 times real time for a typical sentence.

We have applied this system to the DARPA Naval Resource Management Task [11] which allows one to inquire about and display in various ways, the status of a 180 ship fleet. The vocabulary is 992 words including silence and the grammar imposes a highly stylized word order syntax resulting in a entropy of about 4.4 bits/word.

We now turn our attention to the individual components of this system.

## 3. Signal Processing

The speech was sampled at 8 kHz and was analyzed using a sliding 30 ms. window at a 100 Hz frame rate. The spectrum, $S(\omega, t)$, was represented using 12 cepstral coefficients, where the approximate relationship between the spectral magnitude and the resulting cepstral coefficients is defined as

$$\log |S(\omega, t)| \approx 2 \sum_{m=1}^{12} C_m(t) \cos(\omega mt) + C_0(t) . \quad (1)$$

The cepstral coefficients were computed from autocorrelation coefficients via LPC's [17] and they were liftered using the bandpass lifter [18]

$$\hat{C}_m = (1 + 6 \sin(\pi m/12))\, C_m \quad 1 \le m \le 12 . \quad (2)$$

Twelve additional parameters were obtained by evaluating the differential cepstral coefficients, $\Delta\hat{C}_m$, which contain important information about the temporal rate of change of the cepstrum, and are given in [19] as

$$\Delta \hat{C}_m(t) = \frac{\sum_{k=-2}^{2} k\hat{C}_m(t+k)}{\sum_{k=-2}^{2} k^2} \approx \frac{\partial \hat{C}_m}{\partial t} . \quad (3)$$

The combined cepstral and delta cepstral vectors form a set of 24-parameter observation vectors, $\mathbf{O}_t$, which were used in all the experiments described below.

## 4. The Acoustic-Phonetic Model

It is generally accepted that speech is an acoustic manifestation of an underlying phonetic code having a relatively few symbols. The code is, however, a purely mental representation of the spoken language and, as such, is not directly observable. Since the hidden Markov model comprises an unobservable Markov chain and a set of random processes that can be directly measured, it seems most natural to represent speech as a hidden Markov chain in which the hidden states correspond to the putative unobservable phonetic symbols and the state-dependent random processes account for the variability of the observable acoustic manifestation of the corresponding phonetic symbol.

The model that we use to represent the acoustic-phonetic structure of the English language is the continuously variable duration hidden Markov model (CVDHMM) [5]. The states of the model, $\{q_i\}_{i=1}^{n}$, represent the hidden phonetic units. The phonotactic structure of the language is modelled, to a first order approximation, by the state transition matrix, $a_{ij}$, which defines the probability of occurrence of state (phoneme) $q_j$ at time $t+\tau$ conditioned on state (phoneme) $q_i$ at time $t$, where $\tau$ is the duration of phoneme $i$. The information about the temporal structure of the hidden units is contained in the set of durational densities $\{d_{ij}(t)\}_{i,j=1}^{n}$. The acoustic correlates of the phonemes are the observations, denoted $\mathbf{O}_t$, and their distributions, which are defined by a set of observation densities $\{b_{ij}(\mathbf{O}_t)\}_{i,j=1}^{n}$.

The durational densities are 3-parameter gamma distributions

$$d_{ij}(\tau) = \frac{\eta_{ij}^{\nu_{ij}}}{\Gamma(\nu_{ij})} (\tau - \tau_{\min}(i,j))^{\nu_{ij}-1} e^{-\eta_{ij}(\tau-\tau_{\min}(i,j))} \quad (4)$$

where $\Gamma(x)$ is the ordinary gamma function. The observation densities are multivariate Gaussian distributions. Note that they are both indexed by state transition rather than initial state. This affords a rudimentary ability to account for coarticulatory phenomena.

The complete model thus consists of the set of $n$ states (phonemes), the state transition probabilities, $a_{ij}$, $1 \le i,j \le n$; the observation means, $\boldsymbol{\mu}_{ij}$, $1 \le i,j \le n$; the observation covariances, $\mathbf{U}_{ij}$, $1 \le i,j \le n$; and the durational parameters, $\nu_{ij}$ and $\eta_{ij}$, $1 \le i,j \le n$, where the mean duration associated with state transition $i$ to $j$ is $\frac{\nu_{ij}}{\eta_{ij}}$ and the variance of that duration is $\frac{\nu_{ij}}{\eta_{ij}^2}$.

With $n = 47$ phonetic units, the model has 191,000 parameters in all.

## 5. Phonetic Decoding

Since we identify each phonetic unit with a unique state of the CVDHMM as described above, phonetic transcription reduces to the task of finding the most likely state sequence of

the model corresponding to the sequence of acoustic vectors, $\mathbf{O} = \mathbf{O}_1\ \mathbf{O}_2\ \ldots\ \mathbf{O}_t\ \ldots\ \mathbf{O}_T$.

We do so by finding the state and duration sequences whose joint likelihood with $\mathbf{O}$ is maximum. The required optimization is accomplished using a modified Viterbi [20] algorithm. Let $\alpha_t(i)$ denote the maximum likelihood of $\mathbf{O}_1\ \mathbf{O}_2\ \ldots\ \mathbf{O}_t$ over all state and duration sequences terminating in state $i$. This quantity can be evaluated recursively according to

$$\alpha_t(j) = \max_{1 \le i \le n} \left\{ \max_{\tau_{\min}(i,j) \le \tau \le \tau_{\max}} \left\{ \alpha_{t-\tau}^{(i)}\ a_{ij}\ d_{ij}(\tau) \prod_{\theta=0}^{\tau-1} b_{ij}\ (\mathbf{O}_{t-\theta}) \right\} \right\} \tag{5}$$

for $1 \le j \le n$, $1 \le t \le T$ where $\tau_{\min}(i,j)$ is the minimum duration for which $d_{ij}(\tau)$ is defined and $\tau_{\max}$ is the maximum allowable duration for any phonetic unit.

If, at each stage of the recursion on $t$ and $j$, the values of $i$ and $\tau$ that maximize (5) are retained, then one can trace back through the $\alpha_t(j)$ array to obtain the best state and duration sequences

$$\begin{aligned} \hat{\mathbf{q}} &= \hat{q}_1\ \hat{q}_2\ \ldots\ \hat{q}_N \\ \hat{\mathbf{d}} &= \hat{d}_1\ \hat{d}_2\ \ldots\ \hat{d}_N \end{aligned} \tag{6}$$

## 6. Lexical Access and Parsing

The function of the lexical access and parsing algorithms is to find that sentence, $W$, which is well-formed with respect to the task grammar, $G$, and best matches, in some sense, the phonetic transcription, $\hat{\mathbf{q}}$. The lexical access part of the process is that of matching words to subsequences of $\hat{\mathbf{q}}$, while parsing is the part that joins the lexical hypotheses together according to grammatical rules. The two components are conceptually separate and sequential as indicated in Figure 1. However, in order to achieve an efficient implementation, the two are interleaved in a two-level dynamic programming algorithm and hence are treated together in this section.

Lexical access is effected by the lower level of the two-level DP algorithm and consists in matching standard transcriptions of lexical items to various subsequences of $\hat{\mathbf{q}}$. In particular we seek the word, $v$, whose standard transcription $\mathbf{q} = q_1\ q_2\ \ldots\ q_K$ is closest, in a well defined sense, to parts of $\hat{\mathbf{q}}$, say $\hat{q}_{t+1}\ \hat{q}_{t+2}\ \ldots\ \hat{q}_{t+L}$. The well-known solution to this problem [29] is a search over the lattice shown in Figure 3 in which the desired interval of $\hat{\mathbf{q}}$ is placed on the horizontal axis and the correct transcription, $\mathbf{q}$, of some word, $v$, is lined up along the vertical access. The lattice point $(k,l)$ signifies the alignment of $\hat{\mathbf{q}}$ and $\mathbf{q}$ such that $\hat{q}_{t+l}$ coincides with $q_k$.

Let $S_{jkl}$ be the cost of substituting $\hat{q}_{t+l}$ for $q_k$ given that the previous state is $q_j$; $D_{kl}$, the cost of deleting $q_l$ from $\mathbf{q}$ given that the previous state is $q_k$; and $I_{kl}$ the cost of inserting $\hat{q}_{t+1}$ in $\hat{\mathbf{q}}$ when $\hat{q}_{t+l-1} = q_k$. Let us denote by $C_{KL}(v)$ the cost of matching the word, $v$, to $\hat{q}_{t+1}, \ldots, \hat{q}_{t+L}$ where $v$ has the phonetic spelling $q_1, q_2, \ldots, q_K$. Then the lattice is evaluated according to

$$C_{kl}(v) = \min \{C_{kl-1}(v) + I_{kl}, C_{k-1\,l}(v) + D_{k-1\,k}, C_{k-1\,l-1}(v) + S_{k-1\,kl}\} \tag{7}$$

for $1 \le k \le K$ and $1 \le l \le L$. The relation (7) is based upon the symmetric local constraints [21]. The boundary values needed to perform the recursion indicated in (7) are

$$\begin{aligned} C_{00}(v) &= 0 \\ C_{k0}(v) &= \sum_{j=1}^{k} D_{1j}\,\bar{\tau}_{1j} \\ C_{0l}(v) &= \sum_{j=1}^{l} I_{1j}\,\hat{d}_j \end{aligned} \tag{8}$$

for $1 \le k \le K$, $1 \le l \le L$ and $\forall\, v$. In (8), $\bar{\tau}_{ij}$ is the average duration of $q_j$ when preceded by $q_i$ and $\hat{d}_j$ is the duration of $\hat{q}_{t+j}$ as computed by (5).

One could evaluate (7) and (8) based on the Levenshtein metric [22] in which case we would set

$$\begin{aligned} S_{jkl} &= \begin{cases} 0 & \text{if } \hat{q}_{t+l} = q_k \\ 1 & \text{otherwise} \end{cases} && \forall\, j,k,l \\ D_{kl} &= 1 && \forall\, k,l \\ I_{kl} &= 1 && \forall\, k,l \end{aligned} \tag{9}$$

However, the acoustic-phonetic model tells us a great deal about the relative similarities of the phonetic units so we can be more precise than simply using (9) allows.

The dissimilarity between two phonetic units is naturally expressed as the distance between their respective acoustic distributions integrated over their estimated durations. If we adopt the rhobar metric [23] between $b_{jk}(\mathbf{x})$ and $b_{jl}(\mathbf{x})$ then we have

$$S_{jkl} = |\boldsymbol{\mu}_{jk} - \boldsymbol{\mu}_{jl}|\,\hat{d}_{t+l} \quad \forall j \tag{10a}$$

We use a simple heuristic for the costs of insertion and deletion. We treat them both as substitutions with silence, which is represented for convenience by $q_1$. Thus

$$\begin{aligned} D_{kl} &= S_{k1\,l} \\ I_{kl} &= S_{kl1} \end{aligned} \tag{10b}$$

The lexical hypotheses evaluated by the lower level of the DP algorithm, (7), are combined to form sentences by the upper level in accordance with the finite state diagram of the task grammar. The form of the finite state diagram is shown in Figure 4. The state set, $Q$, contains 4767 states connected by 60,433 transitions. There are 90 final states. This grammar was produced from the original specification of the task by a grammar compiler [24]. The language generated by this grammar has a maximum entropy of 4.4 bits/word. The states $r$ and $s$ are completely separate from and not to be confused with the states of the acoustic/phonetic model. The state transition from $r$ to $s$ given word $v$ is denoted by $\delta(r,v) = s$.

Let $R(s,k)$ be the minimum accumulated cost of any phrase of $k$ words starting in state 1 and ending in state $s$. The cumulative cost function obeys the recursion

$$R(s,k) = \min_{P} \left\{ \min_{l} \{R\,(r,\, k-l) + C_{k-l,\,k}(v)\} \right\} \tag{11}$$

$\forall s \in Q$ and $1 \le k \le N$. In (11),

$$P = \{\delta(r,v) = s \quad \text{for any} \quad v\} \tag{12}$$

and the global constraints on expansion and compression of words are given by

$$k - |v|\varepsilon_1 \le l \le k - |v|\varepsilon_2 \tag{13}$$

where $\varepsilon_1 = \frac{5}{2}$ and $\varepsilon_2 = \frac{1}{4}$. Note that the incremental costs $C_{l-k,k}(v)$ are supplied by the lower level from (7). Because the outer minimization of (11) is over the set $P$ as defined in (12), the operation is parallel in $s$.

While computing $R$ from (11), we retain the values of $r$, $v$ and $l$ that minimize each $R(s,k)$. When $R$ is completely evaluated, we trace back through it beginning at the least $R(s,N)$ for which $s$ is a final state. This allows the recovery of the best sentence, $\hat{W}$, and its parse in the form of a state sequence.

## 7. Experimental Results

All the tests described below, except for one informal listening test, were conducted on standard DARPA data that has been filtered and downsampled to a 4 kHz bandwidth. The training set consists of 3,267 sentences spoken by 109 different speakers. This comprises about 4 hrs. of speech. Two test sets each consist of 300 sentences spoken by 10 speakers. The third test set comprises 54 sentences spoken by one of us (SEL) recorded using equipment similar to that used for the DARPA data. All four data sets are completely independent.

The acoustic/phonetic model was trained as follows. The training data was segmented in terms of the 47 phonetic symbols by means of the segmental $k$-means algorithm [25]. All frames so assigned to each phonetic unit were collected and sample statistics for the spectral means and covariances, $\mu_{ij}$ and $\mathbf{U}_{ij}$ and the durational means and variances, $m_{ij}$ and $\sigma_{ij}$, were computed for $1 \le i,j \le 47$. If fewer than 500 samples were available for a particular value of $i$, then the samples for all values of $i$ and fixed $j$ were pooled and only a single statistic was computed and used for all values of $i$. The durational means and variances were then converted to parameters appropriate to the gamma distribution $\nu_{ij}$ and $\eta_{ij}$ according to $m_{ij} = \nu_{ij}/\eta_{ij}$ and $\sigma_{ij} = \nu_{ij}/\eta_{ij}^2$.

The transition matrix was computed from the lexicon. All adjacent pairs of words allowed by the grammar were formed and all occurrences of phonetic units and bigrams were counted. These were then converted to transition probabilities from

$$a_{ij} = \frac{\mathcal{N}(i,j)}{\mathcal{N}(i)} \tag{14}$$

where $\mathcal{N}(i,j)$ is the total number of occurrences of the bigram $q_i\, q_j$ and $\mathcal{N}(i)$ is the total number of occurrences of the unit $q_i$.

Word recognition results are summarized in Table I. All results are for the perplexity 9 grammar.

| Data | # words | % ins. | % del. | % subs. | % correct | % word accuracy | # sents | % sentence accuracy |
|---|---|---|---|---|---|---|---|---|
| train109 | 1838 | 2.5 | 2.1 | 6.3 | 91.6 | 89.1 | 218 | 57.3 |
| feb89 | 2561 | 2.6 | 3.8 | 10.3 | 85.9 | 83.3 | 300 | 40 |
| oct89 | 2684 | 2.3 | 4.1 | 7.9 | 88 | 85.7 | 300 | 44 |
| sel1 | 457 | 0.9 | 0.4 | 2.2 | 97.4 | 96.5 | 54 | 75.9 |

**Table I. Recognition Results**

Data set train109 is a subset of the training data formed by taking two sentences at random from each of the training set speakers. This set was used for algorithm development. The three independent test sets were run only once. Recognition requires about 15 times real time on an 8 CE Alliant FX-80.

Rather than try to measure the accuracy of the phonetic transcription directly, we tried to get an impression of its quality by listening to speech resynthesized from it. For this purpose we use the PRONOUNCE module of tts [26] with $\hat{\mathbf{q}}$, $\hat{\mathbf{d}}$, and a pitch contour computed by the harmonic sieve method [27]. The average data rate for these quantities is approximately 100 bps pointing to the possible utility of the phonetic decoder as a very-low-bit-rate vocoder.

Our informal test was made on six sentences recorded by one of us (SEL). An audio tape was made of the resynthesis and played for several listeners from whose responses we judged that about 75% of the 91 words were intelligible. The speech recognition system gave an 96% word accuracy on these sentences. We have also recorded, decoded and resynthesized several Harvard phonetically balanced sentences with nearly identical results. This is significant since these sentences have no vocabulary in common with the DARPA task.

## 8. Interpretation of the Results

The results listed in Table I are approximately the same as those achieved by more conventional systems tested on the same data [13, 14, 15, 16] and the perplexity 60 grammar. Given the difficulty of the task and the early stage of development of this system, however, we consider these results quite respectable. Also, note that the performance on training data is not substantially different from that obtained on new test data indicating a certain robustness of our method. Moreover, almost all of the insertions and deletions are of monosyllabic articles and prepositions which do not change the meaning of the sentence.

It appears that there are two straightforward ways to improve performance. First we need to improve the acoustic/phonetic model. Desirable structural changes would appear to be the incorporation of trigram phonotactics by making the underlying Markov chain second order [28]. This would allow us to associate the spectral distributions with three states rather than two. This should afford a better model of coarticulatory effects. Also, the spectral distributions can be made more faithful by using Gaussian mixtures rather than unimodal multi-variate densities. Fidelity can be further improved by accounting for temporal correlations among observations. Finally, we need to make a global improvement in the model by optimizing it. We have repeatedly tried reestimation techniques but, thus far, they have actually degraded performance. We speculate that applying constraints to the reestimation

formulae by forcing the state sequence to be fixed will ameliorate the results of optimization.

Second, we can improve the lexical access technique by rationalizing the insertion, deletion, substitution metric. One possible alternative is to replace the rhobar distance with error probabilities determined either analytically or empirically. Also, applying phonological rules to the fixed, citation form, pronunciations stored in the lexicon may eliminate some errors.

## 9. Summary

We have described a novel method for speaker independent recognition of fluent speech from a large vocabulary. The system is a clear and simple implementation of well known linguistic theories of speech perception. The two most striking features of the system are that phonetic decoding is accomplished by a simple optimal search algorithm operating on a stochastic model of the acoustic-to-phonetic mapping and that, after phonetic transcription, processing is entirely symbolic and makes no reference to the acoustic signal.

The performance obtained is not competitive with those obtained from traditional techniques but offers several advantages deriving from the fact that phonetic transcription is independent of lexical or syntactic considerations.

The method described here is in its very earliest stage of development. We are optimistic that further experimentation will soon yield performance at least as good as that displayed by conventional methods.

## REFERENCES


1. Lesser, V. R., Fennell, R. D., Erman, L. D. and Reddy, D. R., "Organization of the HEARSAY-II Speech Understanding System," *IEEE Trans. Acoust. Speech and Signal Processing*, ASSP-23, pp. 11-24, 1975.

2. Woods, W. A., "Motivation and Overview of SPEECHLIS: An Experimental Prototype for Speech Understanding Research," *IEEE Trans. Acoust. Speech and Signal Processing*, ASSP-23, pp. 2-10, 1975.

3. Jelinek, F., "Continuous Speech Recognition by Statistical Methods," *Proc. IEEE*, Vol. 64, pp. 532-556, 1976.

4. Mercier, G., Nouhen, A., Quinton, P. and Siroux, J., "The KEAL Speech Understanding System," in *Spoken Language Generation and Understanding*, J. C. Simon, Ed., D. Reidel, Dordrecht, The Netherlands, pp. 525-544, 1979.

5. Levinson, S. E., "Continuously Variable Duration Hidden Markov Models for Speech Analysis," *Computer Speech and Language*, Vol. 1, No. 1, pp. 29-46, March, 1986.

6. 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, Apr., 1987.

7. 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, Apr., 1988.

8. Miller, L. G. and Levinson, S. E., "Syntactic Analysis for Large Vocabulary Speech Recognition Using a Context-Free Covering Grammar," *Proc. ICASSP-88*, New York, NY, pp. 271-274, Apr., 1988.

9. Levinson, S. E., Liberman, M. Y., Ljolje, A. and Miller, L. G., "Speaker Independent Phonetic Transcription of Fluent Speech for Large Vocabulary Speech Recognition," *Proc. ICASSP-89*, Glasgow, Scotland, UK, pp. 441-444, May, 1989.

10. Levinson, S. E., Liberman, M. Y., Ljolje, A. and Miller, L. G., "Speaker Independent Phonetic Transcription of Fluent Speech for Large Vocabulary Speech Recognition," *Proc. DARPA Workshop on Speech and Natural Language*, Philadelphia, PA, pp. 75-80, Feb., 1989.

11. 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.

12. Levinson, S. E. and Ljolje, A., "Continuous Speech Recognition from Phonetic Transcription," *Proc. DARPA Workshop on Speech and Natural Language*, Harwichport, MA, Oct., 1989.

13. Schwartz, R., Barry, C., Chow, Y.-L., Derr, A., Feng, M.-W., Kimball, O., Kubala, F., Makhoul, J. and Vandegrift, J., "The BBN BYBLOS Continuous Speech Recognition System," *Proc. DARPA Speech and Natural Language Workshop*, Philadelphia, PA, pp. 94-99, Feb., 1989.

14. Paul, D. B., "The Lincoln Continuous Speech Recognition System: Recent Development and Results," *Proc. DARPA Speech and Natural Language Workshop,* Philadelphia, PA, pp. 160-166, Feb., 1989.

15. Murveit, H., Cohen, M., Price, P., Baldwin, G., Weintraub, M. and Bernstein, J., "SRI's DECIPHER System," *Proc. DARPA Workshop on Speech and Natural Language Workshop*, Harwichport, MA, Oct., 1989.

16. Lee, C. H., Rabiner, L. R., Pieraccini, R. and Wilpon, J. G., "Acoustic Modeling for Large Vocabulary Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop,* Harwichport, MA, Oct. 1989.

17. Tohkura, Y., "A Weighted Cepstral Distance Measure for Speech Recognition," *Proc. ICASSP-86*, Tokyo, Japan, pp. 761-764, Apr., 1986.

18. 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, No. 7, pp. 947-954, July, 1987.

19. Soong, F. K. and Rosenberg, A. E., "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," *Proc. ICASSP-86*, Tokyo, Japan, pp. 877-880, April, 1986.

20. Viterbi, A. J., "Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm," *IEEE Trans. Information Theory*, Vol. IT-13, pp. 260-269, 1967.

21. Sakoe, H. and Chiba, S., "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Trans. Acoust. Speech and Signal Processing,* ASSP-26, pp. 43-49, Feb., 1978.

22. Levenshtein, V. I., "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," *Sov. Phys.-Dokl.,* Vol. 10, pp. 707-710, 1966.

23. Gray, R. M., *Probability, Random Processes and Ergodic Properties*, Springer-Verlag, New York, 1988, pp. 254 ff.

24. Brown, M. K. and Wilpon, J. G., "A Grammar Compiler for Connected Speech Recognition," submitted to *IEEE Trans. Acoust. Speech and Signal Processing.*

25. Rabiner, L. R., Wilpon, J. G. and Juang, B.-H., "A Segmental *K*-means Training Procedure for Connected Word Recognition," *AT&T Tech. J.*, Vol. 65, No. 3, pp. 21-31, May-June, 1986.

26. Olive, J. P. and Liberman, M. Y., "Text to Speech: An Overview," *J. Acoust. Soc. Am.*, Vol. 78, Supp. 1, p. 56, Fall, 1985.

27. Duifhuis, H., Willems, L. F. and Sluyter, R. J., "Measurement of Pitch in Speech: An Implementation of Goldstein's Theory of Pitch Perception," *J. Acoust. Soc. Am.*, 71, pp. 1568-1580, 1982.

28. Levinson, S. E., "A Method for the Incorporation of a Tri-gram Model of English Phonotactics in a System for Phonetic Transcription of Unrestricted Speech," Unpublished Bell Laboratories Technical Memorandum, 1988.

29. Wagner, R. and Fischer, M., "The String-to-String Correction Problem," *JACM*, Vol. 21, No. 1, pp. 168-173, 1974.