Improvements in the Stochastic Segment Model for Phoneme Recognition

V. Digalakis† M. Ostendorf† J. R. Rohlicek‡

†Boston University
‡BBN Systems and Technologies Corp.

ABSTRACT

The heart of a speech recognition system is the acoustic model of sub-word units (e.g., phonemes). In this work we discuss refinements of the stochastic segment model, an alternative to hidden Markov models for representation of the acoustic variability of phonemes. We concentrate on mechanisms for better modelling time correlation of features across an entire segment. Results are presented for speaker-independent phoneme classification in continuous speech based on the TIMIT database.

INTRODUCTION

Although hidden Markov models (HMMs) are currently one of the most successful approaches to acoustic modelling for continuous speech recognition, their performance is limited in part because of the assumption that observation features at different times are conditionally independent given the underlying state sequence and because the Markov assumption on the state sequence may not adequately model time structure. An alternative model, the stochastic segment model (SSM), was proposed to overcome some of these deficiencies [Roucos and Dunham 1987, Ostendorf and Roucos 1989, Roucos *et al* 1988].

An observed segment of speech (e.g., a phomeme) is represented by a sequence of q-dimensional feature vectors $Y = [y_1 \; y_2 \; \ldots \; y_k]^T$, where the length k is variable and T denotes block transposition. The stochastic segment model for Y has two components [Roucos *et al* 1988]: 1) a time transformation T_k to model the variable-length observed segment, Y, in terms of a fixed-length unobserved sequence, $X = [x_1 \; x_2 \; \ldots \; x_m]^T$, as $Y = T_k X$, and 2) a probabilistic representation of the unobserved feature sequence X. The conditional density of the observed segment Y given phoneme α and observed length k is:

$$p(Y|\alpha, k) = \int_{X\,:\,Y = T_k X} p(X|\alpha) dX.$$

Assuming the observed length is less than or equal to the length of X, $k \leq m$, T_k is a time-warping transformation which obtains Y by selecting a subset of elements of X and the density $p(Y|\alpha, k)$ is a qk-dimensional marginal distribution of $p(X|\alpha)$. In practice, we can accomodate observations of length $k > m$ by either tying distributions of X (so m is effectively larger) or discarding some of the observations in Y. In this work, as in previous work, the time transformation, T_k, is chosen to map each observed frame y_i to the nearest model sample x_j according to a linear time-warping criterion. The distribution $p(X|\alpha)$ for the segment X given the phoneme α is then modelled using an mq-dimensional multi-variate Gaussian distribution.

Algorithms for automatic recognition and training of the stochastic segment model are similar to those for hidden Markov modelling. The maximum *a posteriori* probability rule is used for classification of segments when the phoneme segmentation is known:

$$\max_{\alpha} p(Y|\alpha, k) p(k|\alpha) p(\alpha),$$

where $p(k|\alpha)$ is the probability that phoneme α has length k. A Viterbi search over all possible segmentations is used for recognition with unknown segmentations. The models are trained from known segmentations using maximum likelihood parameter estimation. When segmentations are unknown, an iterative algorithm based on automatic segmentation and maximum likelihood parameter estimation exists for which increasing the probability of the observations with each iteration is guaranteed.

Initial results with segment-based models have been encouraging. A stochastic segment model has previously been used for speaker-dependent phoneme and word recognition, demonstrating that a segment model outperformed a discrete hidden Markov model when both models were context-independent [Ostendorf and Roucos 1989]. Other segment-based models have also showed encouraging results in speaker-independent applications [Bush and Kopec 1987, Bocchieri and Doddington 1986, Makino and Kido 1986, Zue *et al* 1989].

Although the previous results using the stochastic segment model were encouraging, there were several limitations. First, the comparison to HMMs did not clearly show the advantages of the segment model since the SSM and the HMM used disparate feature distributions: the segment model used ten continuous distributions and the HMM used three discrete distributions for each phoneme. Second, the flexibility of the segment model was not fully exploited because time samples within a segment were assumed independent due to training data limitations in these speaker-dependent applications. Finally, results showed that the context-dependent HMM triphone models [Schwartz *et al* 1985] outperformed the context-independent segment models. Again due to training data limitations, context-dependent segment models were not effective.

In this work we address the issues of 1) time correlation modelling and 2) meaningful comparisons of the SSM with HMMs in a speaker-independent phoneme classification task. In the next section, we describe refinements to the SSM which improve the time correlation modelling capability, including time-dependent parameter reduction and assumption of a Markov time correlation structure. Then experimental results using the TIMIT database are described. These results include comparisons of HMM and SSM, as well as the effects of modelling time correlation. Although the HMM performance is similar to segment model performance when time-sample independence is assumed for the segment model, we demonstrate that the refinements improve performance of the stochastic segment model so that it outperforms the HMM for phoneme classification.

TIME CORRELATION MODELLING

Preliminary experiments using full, mq-dimensional covariance structure for $p(\mathrm{X}|\alpha)$ did not improve performance over the block-diagonal structure used when time samples are assumed to be uncorrelated. We believe that this was due to insufficient training data since a full covariance model has roughly m times as many free parameters as a block diagonal one and since the particular task on which the SSM was tested had a very limited amount of training data. Our efforts have focused on two parallel approaches of handling the training data problem: devising an effective parameter reduction method and constraining the structure of the covariance matrix to further reduce the number of parameters to estimate.

PARAMETER REDUCTION

A first step towards the incorporation of time correlation in the SSM is parameter reduction; an obvious candidate is the method of linear discriminants [Wilks 1962]. Our intuition suggested that sample-dependent reduction would outperform a single transformation for reduction. In fact, contrary to other results [Brown 1987], the single tranformation yielded poor performance (see Section 3).

Linear discriminant parameter reduction for the SSM was implemented using sample-dependent transformations as follows. The speech segment $\mathrm{X} = [x_1\ x_2\ \ldots\ x_m]^T$ is substituted by a sequence of linear transformations of the m original cepstral vectors

$$\tilde{\mathrm{X}} = [\tilde{x}_1\ \tilde{x}_2\ \ldots\ \tilde{x}_m]^T = [R^{(1)}x_1\ R^{(2)}x_2\ \ldots\ R^{(m)}x_m]^T$$

The transformation matrices, $R^{(i)}$, are obtained by solving m generalized eigenvalue problems,

$$S_B^{(i)}w = \lambda S_W^{(i)}w, \quad i = 1, 2, \ldots, m$$

The classes used at each time instant from 1 to m are the phonemes, but the between- and within-class scatter matrices ($S_B^{(i)}$ and $S_W^{(i)}$, respectively) are computed using only the observations for that specific sample.

MARKOVIAN ASSUMPTION

Structure

In order to obtain the advantages of parameterization of time correlation without the m-fold increase in the number of parameters in the full-covariance case, we consider a constrained structure for the covariance matrix. More specifically, we assume that the density of the unobserved segment is that of a non-homogeneous Markov process

$$p(\mathrm{X}) = p_1(x_1)p_2(x_2|x_1)p_3(x_3|x_2)\ldots p_m(x_m|x_{m-1})$$

Under this hypothesis, the number of parameters that have to be estimated increases by less than a factor of 3 over the block-diagonal case (see Table 1). Furthermore, by introducing the Markov restriction to the covariance matrix, we shall also see in the following section that we can simplify the reestimation formulas for the Estimate-Maximize (EM) algorithm.

Block Diagonal	$\frac{md^2}{2} + \frac{3md}{2}$
Markovian	$\frac{(3m-2)d^2}{2} + \frac{3md}{2}$
Full Covariance	$\frac{m^2d^2}{2} + \frac{3md}{2}$

Table 1: Number of parameters per phone model

Parameter Estimation

As mentioned earlier, we assume that the observed segment Y is a "downsampled" version of the underlying fixed-length "hidden" segment, X. We used two different approaches for the parameter estimation problem.

Linear Time Upsampling. The first, *linear time upsampling*, interpolates an unobserved sample x_i of the underlying sequence X, by mapping that point to an observed frame, y_j, with a linear-in-time warping transformation of the observed length k to the fixed length m. The disadvantage of this method is that linear time upsampling introduces a correlation problem when models with non-independent frames are assumed, and in [Roucos *et al* 1988] better results were reported when the parameter estimates were obtained with the EM algorithm. However, in the case of frame dependent transformations, a missing observation is not interpolated by an adjacent one, but by a different transformation of that observation. This partially eliminates the correlation problem.

Estimate-Maximize Algorithm for the Markovian Case. A second approach for the parameter estimation problem is to use the *Estimate-Maximize (EM) algorithm* to obtain a maximum likelihood solution under the assumptions given in Section 2.2.1. As defined in Section 1, X represents the sequence of the incomplete data and Y that of the observed data. In this case, the observed length k is mapped through a linear time warping transformation to the fixed length m, and each observation y_j is assigned to the closest in time x_i. Under the assumption that m is always greater than k, there are certain elements of X that have no elements of Y assigned to them, and we refer to them as "missing". Let μ_i^r and C_{ij}^r denote the estimates at the r-th iteration of the mean vector of the i-th sample and the cross-covariance between the i-th and j-th samples respectively. Then the steps of the EM algorithm are:

1. Estimate step: Estimate the following complete data statistics for each frame or appropriate combination of frames and each observation:

$$E^r(x_i|\mathrm{Y}) = \begin{cases} x_i, & \text{if } x_i \text{ is observed;} \\ E^r(x_i|\mathrm{Y}), & \text{if } x_i \text{ is missing.} \end{cases}$$

$$E^r(x_i x_j'|\mathrm{Y}) = \begin{cases} x_i x_j', & \text{if both are observed;} \\ x_i E^r(x_j'|\mathrm{Y}), & \text{if } x_j \text{ is missing;} \\ E^r(x_i x_j'|\mathrm{Y}), & \text{if both are missing.} \end{cases}$$

where $j = i$ or $j = i+1$, $'$ denotes transposition and $E^r(\cdot)$ is the expectation operator using the r-th iteration density estimates. Under the assumption that the observations form a Markov chain, we have that

$$E^r(x_i|\mathrm{Y}) = E^r(x_i|x_k, x_l), \quad k < i < l$$

$$E^r(x_i x_j'|\mathrm{Y}) = E^r(x_i x_j'|x_k, x_l), \quad k < i, j < l$$

where k and l are the immediately last and next non-missing elements of X to i and j. Let μ_i be the mean of x_i and C_{ij} be the covariance of x_i and x_j. Assuming Gaussian densities, the conditional expectations become

$$E^r(x_i|x_k, x_l) = \mu_i^r + C_{ik}^r V_{kk}(x_k - \mu_k^r) + C_{ik}^r V_{kl}(x_l - \mu_l^r) +$$
$$+ C_{il}^r V_{kl}'(x_k - \mu_k^r) + C_{il}^r V_{ll}(x_l - \mu_l^r)$$

and

$$E^r(x_i x_j'|x_k, x_l) = C_{ij}^r - \{C_{ik}^r V_{kk} C_{jk}^{r\prime} + C_{ik}^r V_{kl} C_{jl}^{r\prime} +$$
$$+ C_{il}^r V_{kl}' C_{jk}^{r\prime} + C_{il}^r V_{ll} C_{jl}^{r\prime}\} + E^r(x_i|x_k, x_l) E^r(x_j'|x_k, x_l)$$

where

$$\begin{bmatrix} V_{kk} & V_{kl} \\ V_{kl}' & V_{ll} \end{bmatrix} = \begin{bmatrix} C_{kk}^r & C_{kl}^r \\ C_{kl}^{r\prime} & C_{ll}^r \end{bmatrix}^{-1}$$

and the V_{kk}, V_{kl}, V_{ll} matrices are obtained from the matrix inversion by partitioning lemma.

2. Maximize step: The $(r+1)$-th estimates are:

$$\mu_i^{r+1} = \frac{1}{|T|} \sum_{X \in T} E^r(x_i|\mathrm{Y})$$

$$C_{ij}^{r+1} = \frac{1}{|T|} \sum_{X \in T} E^r(x_i x_j'|\mathrm{Y}) - \mu_i^{r+1}(\mu_j^{r+1})'$$

where $\mathcal{T}$ is the set of all observations for a certain phoneme. In order to simplify the reestimation formulas, we also investigated a "forward prediction" type approximation, where the expectations of the r-th step are conditioned only on the last observed sample instead of both the last and the next:

$$E^r(x_i|x_k) = \mu_i^r + C_{ik}^r {C_{kk}^r}^{-1}(x_k - \mu_k^r)$$

$$E^r(x_i x_j'|x_k) = C_{ij}^r - C_{ik}^r {C_{kk}^r}^{-1} C_{jk}^{r\prime} +$$
$$+E^r(x_i|x_k)E^r(x_j'|x_k)$$

EXPERIMENTAL RESULTS

In this section we present experimental results for speaker-independent phoneme recognition. We performed experiments on the TIMIT database [Lamel *et al* 1986] for segment models and hidden Markov models using known phonetic segmentations. Mel-warped cepstra and their derivatives, together with the derivative of log power, are used for recognition. We used 61 phonetic models. However, in counting errors, different phones representing the same English phoneme can be substituted for each other without causing an error. The set of 39 English phonemes that we used is the same as the ones that the CMU SPHINX and MIT SUMMIT systems reported phonetic recognition results on [Lee and Hon 1988, Zue *et al* 1989].

The portion of the database that we have available consists of 420 speakers and 10 sentences per speaker, two of these, the "sa" sentences, are the same across all speakers and were not used in either recognition or training because they would lead to optimistic results. We designated 219 male and 98 female speakers for training (a total of 2536 training sentences) and a second set of 71 male and 32 female different speakers for testing (a total of 824 test sentences with 31,990 phonemes). We deliberately selected a large number of testing speakers to increase the confidence of performance estimates. The best-case results, reported at the end of this section, were obtained using all of the available training sentences (from both male and female speakers) and testing over the entire test set. Most of the other results for algorithm development and comparisons were obtained by training over the male speakers only and testing on the Western and N.Y. dialect male speakers (a total of 219 training and 17 test speakers), a subset that gives us good estimates of the overall performance as we can see from the global results.

SSM w/o duration	67.0%
SSM with duration	68.6%
HMM	68.7%

Table 2: HMM and Segment Comparison (m = 5, 10 cepstra and deriv., no time correlation).

The TIMIT database has also been used by other researchers for the evaluation of the phonetic accuracy of their speech recognizers. Lee and Hon, 1988, reported a phonetic accuracy for the SPHINX system of from 58.8% with 12% insertions when context-independent (61) phone models were used. Zue *et al*, 1989 obtained a 70% classification performance on the same database for unknown speakers.

HMM/segment comparison. We first evaluated the relative performance of the SSM to a continuous Gaussian distribution HMM [Bahl *et al* 1981]. In this experiment, the features were ten mel-warped cepstra and their derivatives. Both the SSM and the HMM had the same number of distributions (SSM of length m = 5 and no time correlation versus a 7-state, 5 distributions HMM), and the recognition algorithm for the HMM was a full search along the observed length. For the SSM, we obtained results for two cases: with and without using the duration information $p(k|\alpha)$. Note that, for a fair comparison, the segment model should include duration information, which was not incorporated in earlier versions of the segment model. With the duration information, both the SSM and the HMM gave similar performance (68.6% and 68.7%, see Table 2). The superiority of the SSM becomes clearer after the time correlation and the parameter reduction methods are incorporated, even though the SSM with time correlation suffered from limited training.

Parameter Reduction. We compared the single and multiple transformation reduction methods on a single-speaker task, using 14 mel-warped cepstra (but not their derivatives) as original features. We evaluated the recognition performance of the SSM for 1) different numbers of the original cepstral coefficients (from 4 up to 14), 2) different number of linear discriminants obtained using a single transformation and 3) different numbers of linear

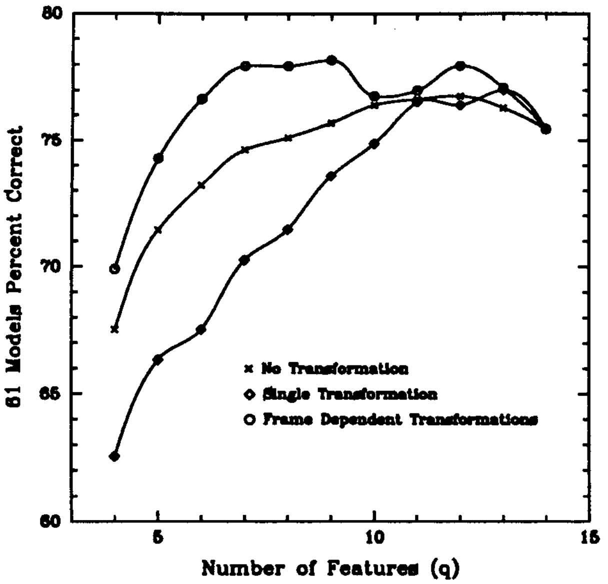

Figure 1: Evaluation of parameter reduction methods on a single-speaker task.

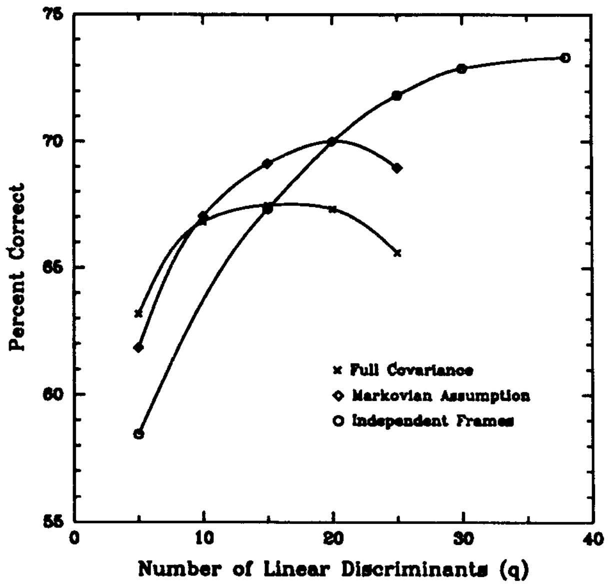

Figure 2: Performance of different covariance structure for a segment model length $m = 5$ using linear upsampling parameter estimation.

discriminants obtained from frame dependent transformations (see Figure 1). The frame dependent features gave the best performance of 78.2% when the features were reduced to 9 due to training problems, whereas the single transformation features gave actually lower performance than the original features. This can be explained by the fact that there was a small between-class scatter for the single transformation, relative to the sample-dependent transformations. The eigenvalue spread for the single transformation was only 6.2 (ratio of largest to smallest eigenvalue) whereas in the case of multiple transformations this ratio ranged from 178.7 to 318.3. A larger ratio occurs at the middle frames since the effect of adjacent phonemes is smaller at the middle of a certain phoneme and is easier to discriminate. The recognition performance for the single speaker reported here was measured on a set of 61 phonemes counting misrecognized allophones as errors.

Time Correlation. We performed a series of experiments on three different types of covariance matrices for the SSM. The length of the SSM in this case was $m = 5$. In Figure 2, we have plotted the phonetic accuracy versus the number of features q for 1) a full covariance, 2) a Markov structure and 3) for a block diagonal covariance (independent frames). When the number of features is small and there is enough data to estimate the parameters, the full covariance model outperforms the other two. It should be noted though that the performance of the Markovian model is close to that of the full covariance even for a small number of features. Hence, the Markov hypothesis represents well the structure of the covariance matrix, and as the number of features increases the Markov model outperforms the full covariance model since it is more easily trainable. In addition, we expect that the curves for those two models would be further separated for a bigger segment length m, since the number of parameters for the full model is quadratic in m, whereas that of the Markov model is linear. (For $m = 5$ the full model has almost twice as many parameters as the Markov model).

As the number of parameters increases, the independence assumption gives the best recognition performance. However, with more training data the models that use time correlation will outperform the model which does not. Furthermore, we were able to duplicate the best case results using a Markov model for the first features and a second independent distribution for the "less significant" features. (In this case, the correlation

Method	5-Cepstra 5-DCepstra	10 LD
Lin.Time Resampling	45.7%	67.3%
EM Algorithm	59.0%	62.8%
Forward Approxim.	58.7%	60.5%

Table 3: Parameter Estimation Algorithms ($m = 8$).

between the first and the last features is lost, but the time correlation between successive frames compensates for this).

Parameter Estimation Algorithms. We evaluated the different methods of parameter estimation that we presented in Section 2. In this set of experiments, we only used 10 features (either 5 mel-warped cepstra and their derivatives, or 10 linear discriminants) due to limitations in the available computer time. A segment of length $m = 8$ rather than the usual $m = 5$ was used, in order to obtain a better understanding of the interpolation potential of each algorithm. The comparative results are summarized in Table 3. When cepstra and their derivatives are used, the EM algorithm clearly gives better results than the linear time upsampling method. In addition, the "forward prediction" approximation gave us similar recognition performance to the one obtained when the full reestimation formulas were used. However, the situation is inverted when the features are linear discriminants, and we refer the reader to Section 2 for an explanation of those results.

Global Results. The best case system – based on independent samples, $m = 5$, and 38 linear discriminants – was evaluated using the entire data set. The classifier was trained on the whole training set of all male and female speakers, and tested on 824 sentences from 103 speakers. As it can be seen in Table 4, where we present the results by region and sex, the phoneme classification rate does not have large variations among different regions, indicating the robustness of our classifier. The somewhat higher numbers on the male speakers can be attributed to the fact that approximately 70% of our training set consisted of sentences spoken by male speakers and the classifier was biased in this sense. The results were also consistent among different speakers. The recognition rates for all speakers ranged from 59.9% to 80.7%, with the median speakers at 72.7% for the male test speakers, 71.9% for the female speakers and 72.3% for the whole test set. Approximately 80% of all the test speakers (82 out of 103) had a recognition performance over 69%, and only 8% of the speakers gave performance below 65%, including some "problematic" speakers.

Region	Males	Females	M + F
New England	72.4%	70.1%	71.4%
Northern	73.5%	71.7%	73.1%
North Midland	70.8%	70.0%	70.6%
South Midland	72.4%	70.6%	71.8%
Southern	73.2%	70.7%	72.2%
N.Y.City	69.2%	70.6%	69.7%
Western	72.8%	75.9%	73.6%
Army Brat	73.8%	73.3%	73.7%
Total	72.3%	71.4%	72.1%

Table 4: Phonetic Accuracy by Region and Sex

Our best case result of 72% correct classification can be compared to the SUMMIT 70% classification performance on the TIMIT data for unknown speakers [Zue *et al* 1989]. Although these results are based on known segmentations, past work in segment modelling for speaker-dependent phoneme recognition showed that recognition with unknown segmentations yields a small loss in recognition performance with a cost of 10% phoneme insertion [Ostendorf and Roucos 1989]. With this small loss in performance, the segment models can still be expected to outperform HMM phoneme recognition performance of 59% on this task [Lee and Hon 1988].

CONCLUSIONS

In summary, we have shown that with sufficient training data, it is possible to model detailed time correlation in a segment-based model which can outperform HMMs in context-independent phoneme classification tasks. It remains to be shown that this result also holds for phoneme recognition, when phoneme segmentation boundaries are not known. In addition, the result should be extended to new tasks, where automatic training will be required.

There are several directions to further develop the SSM. Since context-dependent models have been shown to give dramatic improvements in HMM word recognition, it is important to demonstrate similar results for segment models. This will require research in robust parameter estimation techniques. In addition, research on the variable-to-fixed length transformation is also important. Although a constrained transformation is probably an advantage of the segment model, it is not clear that linear time warping is the best transformation for all phonemes, and it may be useful to develop a mechanism for estimating transformations.

Acknowledgement

This work was jointly funded by NSF and DARPA under NSF grant number IRI-8902124.

References

[Bahl *et al* 1981] L.R. Bahl, R. Bakis, P.S. Cohen, A. Cole, F. Jelinek, B.L. Lewis, and R.L. Mercer, "Continuous parameter acoustic processing for recognition of a natural speech corpus," In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1149–1152, Atlanta, GA, April 1981.

[Bocchieri and Doddington 1986] E. Bocchieri and G. Doddington, "Frame-specific statistical features for speaker-independent speech recognition," *IEEE Trans. Acoust., Speech and Signal Proc.*, ASSP-34(4):755–764, August 1986.

[Brown 1987] P. F. Brown, *The Acoustic-Modeling Probelm in Automatic Speech Recognition*, PhD thesis, Carnegie-Mellon University, May 1987. IBM Technical Report RC12750.

[Bush and Kopec 1987] M.A. Bush and G.E. Kopec, "Network-based connected digit recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-35(10):1401–1413, October 1987.

[Lamel *et al* 1986] L.F. Lamel, R. H. Kassel and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," in *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, pp. 100-109, Feb. 1986.

[Lee and Hon 1988] K.-F. Lee and H.-W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," CMU Technical Report No. CMU-CS-88-121.

[Makino and Kido 1986] S. Makino and K. Kido, "Recognition of Phonemes Using Time-Spectrum pattern," *Speech Communication*, Vol. 5, No. 2, June 1986, pp. 225-238.

[Ostendorf and Roucos 1989] M. Ostendorf and S. Roucos, "A stochastic segment model for phoneme-based continuous speech recognition," to appear, *IEEE Trans. Acoustic Speech and Signal Processing*, December 1989.

[Roucos and Dunham 1987] S. Roucos and M. Ostendorf Dunham, "A stochastic segment model for phoneme-based continuous speech recognition," In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 73–89, Dallas, TX, April 1987. Paper No. 3.3.

[Roucos *et al* 1988] S. Roucos, M. Ostendorf, H. Gish, and A. Derr, "Stochastic segment modeling using the estimate-maximize algorithm," In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 127–130, New York, New York, April 1988.

[Schwartz *et al* 1985] R.M. Schwartz, Y.L. Chow, O.A. Kimball, S. Roucos, M. Krasner, and J. Makhoul, "Context-dependent modeling for acoustic-phonetic recognition of continuous speech," In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1205–1208, Tampa, FL, March 1985. Paper No. 31.3.

[Wilks 1962] S.S. Wilks, *Mathematical Statistics*, John Wiley & Sons, 1962.

[Zue *et al* 1989] V. Zue, J. Glass, M. Phillips and S. Seneff, "Acoustic Segmentation and Phonetic Classification in the Summit System," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 389–392, Glasgow, Scotland, May 1989.