HIGH-ACCURACY LARGE-VOCABULARY SPEECH RECOGNITION USING MIXTURE TYING AND CONSISTENCY MODELING

Vassilios Digalakis and Hy Murveit

SRI International
Speech Technology and Research Laboratory
333 Ravenswood Ave., Menlo Park, CA 94025-3493

ABSTRACT

Improved acoustic modeling can significantly decrease the error rate in large-vocabulary speech recognition. Our approach to the problem is twofold. We first propose a scheme that optimizes the degree of mixture tying for a given amount of training data and computational resources. Experimental results on the Wall Street Journal (WSJ) Corpus show that this new form of output distribution achieves a 25% reduction in error rate over typical tied-mixture systems. We then show that an additional improvement can be achieved by modeling local time correlation with linear discriminant features.

1. INTRODUCTION

To improve the acoustic-modeling component of SRI's DECIPHER™ speech recognition system, our research has focused on two main directions. The first is to decrease the degree of mixture tying in the mixture observation densities, since continuous-density hidden Markov models (HMMs) have recently been shown to outperform discrete-density and tied-mixture HMMs [16]. The second is the removal of the simplifying output independence assumption commonly used in HMMs.

Tied mixtures (TM) achieve robust estimation and efficient computation of the density likelihoods. However, the typical mixture size used in TM systems is small and does not provide a good representation of the acoustic space. Increasing the number of the mixture components (the codebook size) is not a feasible solution, since the mixture-weight distributions become too sparse. In large-vocabulary problems, where a large number of basic HMMs is used and each has only a few observations in the training data, sparse mixture-weight distributions cannot be estimated robustly and are expensive to store. To solve this problem, we follow the approach of simultaneously reducing the codebook size and increasing the number of different sets of mixture components (or codebooks). This procedure reduces the degree of tying, and the two changes can be balanced so that the total number of component densities in the system is effectively increased. The mapping from HMM states to codebooks can be determined using clustering techniques. Since our algorithm transforms a "less" continuous, or tied-mixture system, to a "more" continuous one, it has enabled us to investigate a number of traditional differences between tied-mixture and fully continuous HMMs, including codebook size and modeling of the speech features using multiple vs. single observation streams.

Our second main research direction is focused on removing the simplifying assumption used in HMMs that speech features from different frames are statistically independent given the underlying state sequence. In this paper we will deal with the modeling of the local temporal dependencies, that is, ones that span the duration of a phonetic segment. We will show through the use of recognition experiments and information theoretic criteria that achieving decorrelation of the speech features is not a sufficient condition for the improvement in recognition performance. To achieve the latter, it is necessary to improve the discrimination power of the output distributions through the use of new information. Local correlation modeling has recently been incorporated in our system through the use of linear discriminant features, and has reduced the word error rate by 7% on the Wall Street Journal (WSJ) corpus.

The remainder of the paper is organized as follows: in Section 2 we present the general form of mixture observation distributions used in HMMs, we discuss variations of this form that have appeared in the literature, and present an algorithm that enables us to adjust the mixture tying for optimum recognition performance. In Section 3 we deal with the problem of local time-correlation modeling: we comment on the potential improvement in recognition performance by incorporating conditional distributions, and describe the type of local consistency modeling currently used in our system. In Section 4 we present experimental results on the WSJ Corpus. These results are mainly a by-product of the system development for the November 1993 ARPA evaluation [16]. Finally, we conclude in Section 5.

2. GENONIC MIXTURES

A typical mixture observation distribution in an HMM-based speech recognizer has the form

$$p(x_t \mid s) = \sum_{q \in Q(s)} p(q|s) f(x_t|q) \qquad (1)$$

where s represents the HMM state, x_t the observed feature at frame t, and $Q(s)$ the set of mixture-component densities used in state s. We will use the term *codebook* to denote the set $Q(s)$. The stream of continuous vector observations can be modeled directly using Gaussians or other types of densities in the place

of $f(x_t \mid q)$, and HMMs with this form of observation distributions are known as continuous HMMs [19].

Various forms of tying have appeared in the literature. When tying is not used, the sets of component densities are different for different HMM states—that is, $Q(s) \neq Q(s')$ if $s \neq s'$. We will refer to HMMs that use no sharing of mixture components as *fully continuous* HMMs. The other extreme is when all HMM states share the same set of mixture components—that is, $Q(s) = Q$ is independent of the state s. HMMs with this degree of sharing were proposed in [8], [2] under the names *Semi-Continuous and Tied-Mixture* (TM) HMMs. Tied-mixture distributions have also been used with segment-based models, and a good review is given in [11]. Intermediate degrees of tying have also been examined. In phone-based tying, described in [17], [13], only HMM states that belong to allophones of the same phone share the same mixture components—that is, $Q(s) = Q(s')$ if s and s' are states of context-dependent HMMs with the same center phone. We will use the term *phonetically tied* to describe this kind of tying. Of course, for context-independent models, phonetically tied and fully continuous HMMs are equivalent. However, phonetically tied mixtures (PTM) did not significantly improve recognition performance in previous work.

The continuum between fully continuous and tied-mixture HMMs can be sampled at any other point. The choice of phonetically tied mixtures, although linguistically motivated, is somewhat arbitrary and may not achieve the optimum trade-off between resolution and trainability. We have recently introduced an algorithm [4] that allows as to select the degree of tying that attains optimum recognition performance for the given computational resources. This algorithm follows a bootstrap approach from a system that has a higher degree of tying (i.e., a TM or a PTM system), and progressively unties the mixtures using three steps: clustering, splitting and pruning, and reestimation.

2.1. Clustering

The HMM states of all allophones of a phone are clustered following an agglomerative procedure. The clustering is based on the weighted-by-counts entropy of the mixture-weight distributions [12]. The clustering procedure partitions the set of HMM states S into disjoint sets of states

$$S = S_1 \cup S_2 \cup \ldots \cup S_n \qquad (2)$$

The same codebooks will be used for all HMM states belonging to a particular cluster S_i.

2.2. Splitting and Pruning

After determination of the sets of HMM states that will share the same codebook, seed codebooks for each set of states that will be used by the next reestimation phase are constructed. These seed codebooks can be constructed by either one or a combination of two procedures:

- Identifying the most likely subset of mixture components of the boot system for each cluster of HMM states S_i and using these subsets $Q(S_i) \subset Q(S)$ as seed codebooks for the next phase
- Copying the original codebook multiple times (one for each cluster of states) and performing one iteration of the Baum-Welch algorithm over the training data with the new tying scheme; the number of component densities in each codebook can then be reduced using clustering [10]

2.3. Reestimation

The parameters are reestimated using the Baum-Welch algorithm. This step allows the codebooks to deviate from the initial values and achieve a better approximation of the distributions.

We will refer to the Gaussian codebooks as *genones* and to the HMMs with arbitrary tying of Gaussian mixtures as *genonic* HMMs. Clustering of either phone or subphone units in HMMs has also been used in [18], [12], [1], [9]. Mixture-weight clustering of different HMM states can reduce the number of free parameters in the system and, potentially, improve recognition performance because of the more robust estimation. It cannot, however, improve the resolution with which the acoustic space is represented, since the total number of component densities in the system remains the same. In our approach, we use clustering to identify sets of subphonetic regions that will share mixture components. The later steps of the algorithm, where the original set of mixture components is split into multiple overlapping genones and each one is reestimated using data from the states belonging to the corresponding cluster, effectively increase the number of distinct densities in the system and provide the desired detail in the resolution.

Reestimation of the parameters can be achieved using the standard Baum-Welch reestimation formulae for HMMs with Gaussian mixture observation densities, since tying does not alter their form, as pointed out in [21]. During recognition, and to reduce the large amount of computation involved in evaluating Gaussian likelihoods, we can use the fast computational techniques described in [15].

In place of the component densities $f(x_t \mid q)$ we use exponentially weighted Gaussian distributions:

$$p(x_t|s) = \sum_{q \in Q(s)} p(q|s) \left[N(x_t;\mu_q, \Sigma_q)\right]^{\alpha} \qquad (3)$$

where the exponent $\alpha \leq 1$ is used to reduce the dynamic range of the Gaussian scores (that would, otherwise, dominate the mixture probabilities $p(q \mid s)$) and also to provide a smoothing effect at the tails of the Gaussians.

3. TIME CORRELATION MODELING

For a given HMM state sequence, the observed features at nearby frames are highly correlated. Modeling time correlation can significantly improve speech recognition performance for two reasons. First, dynamic information is very important [6], and explicit time-correlation modeling can potentially outperform more traditional and simplistic approaches like the incorporation of cepstral derivatives as additional feature streams.

Second, sources of variability—such as microphone, vocal tract shape, speaker dialect, and speech rate—will not dominate the likelihood computation during Viterbi decoding by being rescored at every frame. We will call techniques that model such temporal dependencies *consistency modeling*.

The output-independence assumption is not necessary for the development of the HMM recognition (Viterbi) and training (Baum-Welch) algorithms. Both of these algorithms can be modified to cover the case when the features depend not only on the current HMM state, but also on features at previous frames [20]. However, with the exception of the work reported in [3] that was based on segment models, explicit time-correlation modeling has not improved the performance of HMM-based speech recognizers.

To investigate these results, we conducted a pilot study to estimate the potential improvement in recognition performance when using explicit correlation modeling over more traditional methods like time-derivative information. We used information-theoretic criteria and measured the amount of mutual information between the current HMM state and the cepstral coefficients at a previous "history" frame. The mutual information was always conditioned on the identity of the left phone, and was measured under three different conditions:

- *I(h,s)*—mutual information between the current HMM state *s* and a cepstral coefficient *h* at the history frame; a single, left-context-dependent Gaussian distribution for the cepstral coefficient at the history frame was hypothesized,
- *I(h,s/c)*—conditional mutual information between the current HMM state *s* and a cepstral coefficient *h* at the history frame when the corresponding cepstral coefficient *c* of the current frame is given; a left-context-dependent, joint Gaussian distribution for the cepstral coefficients at the current and the history frames was hypothesized,
- *I(h,s/c,d)*—same as above, but conditioned on both the cepstral coefficient *c* and its corresponding derivative *d* at the current frame.

The results are summarized in Table 1 for history frames with lags of 1, 2, 4 and a variable one. In the latter case, we condition the mutual information on features extracted at the last frame t_0 of the previous HMM state, as located by a forced Viterbi alignment. We can see from this table that in the unconditional case, the cepstral coefficients at frames closer to the current one provide more information about the identity of the current phone. However, the amount of additional information that these coefficients provide when the knowledge of the current cepstra and their derivatives is taken into account is smaller. The additional information in this case is larger for lags greater than 1, and is maximum for the variable lag.

These measurements predict that the previous frame's observation is not the optimal frame to use when conditioning a state's output distribution. To verify this, and to actually evaluate recognition performance, we incorporated time-correlation modeling in an HMM system with genonic mixtures. Specifically, we generalized the Gaussian mixtures to mixtures of conditional Gaussians, with the current cepstral coefficient x_t conditioned on the corresponding cepstral coefficient x_{t_0} of the history frame t_0:

Lag t_0	0	1	2	4	Variable
I(*h*, *s*)	0.28	0.27	0.25	0.19	0.25
I(*h*, *s* | *c*)	0	0.13	0.15	0.15	0.21
I(*h*, *s* | *c*, *d*)	0	0.11	0.14	0.13	0.20

Table 1. Mutual information (in bits) between HMM state *s* at time *t* and cepstral coefficient *h* at time t-t_0 for various lags; included is the conditional mutual information when the corresponding cepstral coefficient and its derivative at time *t* are given

$$p\left(x_t \mid s, x_{t-t_0}\right) = \sum_{q \in Q(s)} p(q|s) f\left(x_t | q, x_{t_0}\right) \qquad (4)$$

We either replaced the original unconditional distributions of the cepstral coefficients and their derivatives with the conditional Gaussian distributions, or we used them in parallel as additional observation streams. The results on the 5,000-word recognition task of the WSJ0 corpus are summarized in Table 2 for fixed-lag history frames. We can see that the recognition results are in perfect agreement with the behavior predicted by the mutual-information study. The improvements in recognition performance over the system that does not use conditional distributions are actually proportional to the measured amount of conditional mutual information at the various history frames. However, these improvements are small and statistically insignificant, and indicate that the derivative features effectively model the local dynamics.

Delay	Word Error—Conditional only (%)	Word Error—Both (%)	I(*h*, *s* | *c*, *d*)
0	10.32	-	0
1	10.98	10.19	0.11
2	10.50	9.65	0.14
4	10.32	9.83	0.13

Table 2. Recognition rates on 5,000-word WSJ corpus with conditional distributions either replacing the unconditional ones or used in parallel

Instead of using conditional Gaussian distributions, one can alternatively choose to use features obtained with linear discriminants. Local time correlation can be modeled by estimating the transformations over multiple consecutive frames [5],[7]. This approach has the additional advantage that it is computationally less expensive, since the discriminant transformations can be computed in the recognizer front end and only once at each frame. However, as we will see in the following section, linear discriminants gave only moderate improvements in recognition performance, and this is consistent with the conditional Gaussian results of this section. From the conditional information measurements that we have presented, we can see that in order to provide additional information to the recognizer we must condition the output distributions not only on a previous history frame, but also on the start time of the current subphonetic segment, and this is an area that we are currently investigating.

4. EXPERIMENTAL RESULTS

We used the algorithms described in this paper on the 5,000- and 64,000-word recognition tasks of the WSJ corpus. We used the progressive-search framework [14] for fast experimentation. With this approach, an initial fast recognition pass creates word lattices for all sentences in the development set. These word lattices are used to constrain the search space in all subsequent experiments. In our development we used both the WSJ0 5,000 word and the WSJ1 64,000 word portions of the database, and the baseline bigram and trigram language models provided by Lincoln Laboratory.

4.1. Degree of Mixture Tying

To determine the effect of mixture tying on the recognition performance, we evaluated a number of different systems on both WSJ0 and WSJ1. Table 3 compares the performance and the number of free parameters of tied mixtures, phonetically tied mixtures, and genonic mixtures on a development set that consists of 18 male speakers and 360 sentences of the 5,000-word WSJ0 task. The training data for this experiment included 3,500 sentences from 42 speakers. We can see that systems with a smaller degree of tying outperform the conventional tied mixtures by 25%, and at the same time have a smaller number of free parameters because of the reduction in the codebook size.

System	Number of Genones	Gaussians per genone	Total Parameters (thousands)	Word Error (%)
TM	1	256	5,126	14.1
PTM	40	100	2,096	11.6
Genones	495	48	1,530	10.6

Table 3. Comparison of various degrees of tying on 5,000-word WSJ development set

The difference in recognition performance between PTM and genonic HMMs with smaller tying is, however, much more dramatic in the WSJ1 portion of the database. The training data consisted of 37,000 sentences from 280 speakers, and gender-dependent models were built. The male subset of the 20,000-word November 1992 evaluation set was used, with a bigram language model. Table 4 compares various degrees of tying by varying the number of genones used in the system. We can see that, because of the larger amount of available training data, the improvement in performance of genonic systems over PTM systems is much larger (20%) than in our 5,000-word experiments. Moreover, the best performance is achieved for a larger number of genones—1,700 instead of the 495 used in the 5,000-word experiments. These results are depicted in Figure 1.

	PTM	Genonic HMMs			
Number of Genones	40	760	1250	1700	2400
Word error rate (%)	14.7	12.3	11.8	11.4	12.0

Table 4. Recognition performance on the male subset of 20,000-word WSJ November 1992 ARPA evaluation set for various numbers of codebooks using a bigram language model.

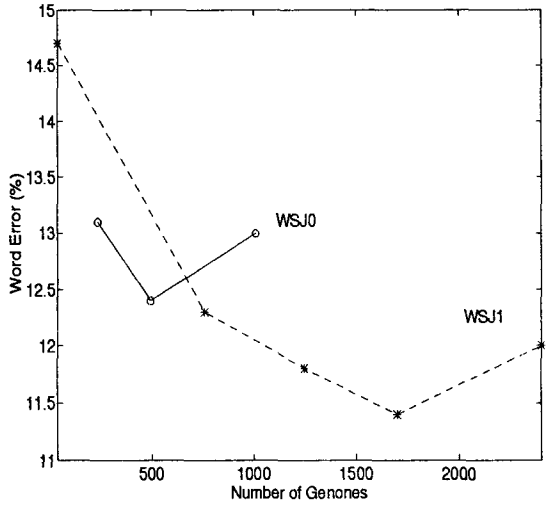

Figure 1: Recognition performance for different degrees of tying on the 5,000-word WSJ0 and 20,000-word WSJ1 tasks of the WSJ corpus

In Table 5 we explore the additional degree of freedom that genonic HMMs have over fully continuous HMMs, namely that states mapped to the same genone can have different mixture weights. We can see that tying the mixture weights in addition to the Gaussians introduces a significant degradation in recognition performance. This degradation increases when the features are modeled using multiple observation streams (see following section) and as the amount of training data and the number of genones decrease.

	Number of Genones	Number of Streams	Word Error (%) Tied	Untied
5K WSJ0	495	6	9.7	7.7
20K WSJ1	1,700	1	12.2	11.4

Table 5. Comparison of state-specific vs. genone-specific mixture weights for different recognition tasks

4.2. Multiple vs. Single Observation Streams

Another traditional difference between fully continuous and tied mixture systems is the independence assumption of the latter when modeling multiple speech features. Tied mixture systems typically model static and dynamic spectral and energy features as conditionally independent observation streams given the HMM state, because tied mixture systems provide a very coarse representation of the acoustic space. It is, therefore, necessary to "quantize" each feature separately and artificially increase the resolution by modeling the features as independent: the number of "bins" of the augmented feature is equal to the product of the number of "bins" of all individual features. The disadvantage is, of course, the independence assumption. When, however, the degree of tying is smaller, the finer representation of the acoustic space makes it unnecessary to artificially improve the resolution accuracy by modeling the features as independent. Hence, for systems that are loosely tied we can remove the feature-independence assumption. This claim is verified experimentally in Table 6. The first row shows the recognition performance of a system that models the six static and dynamic spectral and energy features used in DECIPHER™ as independent observation streams. The second row shows the performance of a system that models the six features in a single stream. We can see that the performance of the two systems is similar.

System	Sub (%)	Del (%)	Ins (%)	Word Error (%)
6 streams	9.0	0.8	2.5	12.3
1 stream	8.7	0.8	2.3	11.8

Table 6. Comparison of modeling using 6 versus 1 observation streams for 6 underlying features on the male subset of 20,000-word WSJ November 1992 evaluation set with a bigram language model

4.3. Linear Discriminant Features

To capture local time correlation we used a linear discriminant feature extracted using a transformation of the features within a window around the current frame. The discriminant transformation was obtained using linear discriminant analysis with classes defined as the HMM state of the context-independent phone. The state index that was assigned to the frame was determined using the maximum *a-posteriori* criterion and the forward-backward algorithm.

We found that the performance of the linear discriminant feature was similar to that of the original features. However, we found that an improvement in performance can be obtained if the discriminant features are used in parallel with the original features. A genonic HMM system with 1,700 genones and linear discriminants as an additional feature was evaluated on the 20,000-word open-vocabulary November 1993 ARPA evaluation set. It achieved word-error rates of 16.5% and 14.5% with the standard bigram and trigram language models, respectively. These results, however, were contaminated by the presence of a large DC offset in most of the waveforms of the phase 1 WSJ1 corpus. We later removed the DC offset from the waveforms, and reestimated the models using the exact procedure followed during the development of the system used in the November 1993 evaluation. From Table 6, we can see that the linear discriminant feature reduced

System	Bigram LM	Trigram LM
1,700 Genones	20.5	17.0
+ Linear Discriminants	19.1	15.8

Table 7. Word error rates (%) on the 20,000-word open-vocabulary male development set of the WSJ1 corpus with and without linear discriminant transformations

the error rate on the WSJ1 20,000-word open-vocabulary male development set by approximately 7% using either a bigram or a trigram language model. Table 4 presents the results of the system with linear discriminants on various test and development sets.

	Test set		
Grammar	Nov92	WSJ1 Dev	Nov93
Bigram	11.2	16.6	16.2
Trigram	9.3	13.6	13.6

Table 8. Word error rates on the November 1992 evaluation, the WSJ1 development, and the November 1993 evaluation sets using 20,000-word open-vocabulary bigram and trigram language models

5. CONCLUSIONS

New acoustic modeling techniques significantly decrease the error rate in large-vocabulary continuous speech recognition. The genonic HMMs balance the trade-off between resolution and trainability, and achieve the degree of tying that is best suited to the available training data and computational resources. For example, one can decrease the computational load by decreasing the number of genones (i.e., increasing the degree of tying) with a small penalty in recognition performance [15]. Our results on the various test sets represent state-of-the-art recognition performance on the 20,000-word open-vocabulary WSJ task.

ACKNOWLEDGMENTS

We gratefully acknowledge support for this work from ARPA through Office of Naval Research Contract N00014-92-C-0154. The Government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Government funding agencies.

REFERENCES

1. L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo and M. A. Picheny, "Context Dependent Modeling of Phones in Continuous Speech Using Decision Trees,"

DARPA Workshop on Speech and Natural Language, pp. 264-269, February 1991.

2. J. R. Bellegarda and D. Nahamoo, "Tied Mixture Continuous Parameter Modeling for Speech Recognition," *IEEE Trans. ASSP*, Vol. 38(12), pp. 2033-2045, Dec. 1990.
3. V. Digalakis, J. R. Rohlicek and M. Ostendorf, "ML Estimation of a Stochastic Linear System with the EM Algorithm and its Application to Speech Recognition," *IEEE Trans. Speech and Audio Processing*, October 1993.
4. V. Digalakis and H. Murveit, "Genones: Optimizing the Degree of Tying in a Large Vocabulary HMM-based Speech Recognizer," to appear in *Proc. ICASSP*, 1994.
5. G. R. Doddington, "Phonetically Sensitive Discriminants for Improved Speech Recognition," *Proceedings ICASSP-89*, pp. 556-559.
6. S. Furui, "On the Role of Spectral Transition for Speech Perception," *Journal of the Acoustical Society of America*, vol. 80(4), pp. 1016-1025, October 1986.
7. R. Haeb-Umbach and H. Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition," *Proc. ICASSP*, pp. I-13 - I-16, March 1992.
8. X. D. Huang and M. A. Jack, "Performance Comparison Between Semi-continuous and Discrete Hidden Markov Models," *IEE Electronics Letters*, Vol. 24 no. 3, pp. 149-150.
9. M.-Y. Hwang and X. D. Huang, "Subphonetic Modeling with Markov States - Senone," *Proc. ICASSP*, pp. I-33-36, March 1992.
10. A. Kannan, M. Ostendorf and J. R. Rohlicek, "Maximum Likelihood Clustering of Gaussians for Speech Recognition," in *IEEE Trans. Speech and Audio Processing*, to appear July 1994.
11. O. Kimball and M. Ostendorf, "On the Use of Tied-Mixture Distributions," *Proc. ARPA HLT Workshop*, March 1993.
12. K. F. Lee, "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition," *IEEE Trans. ASSP*, pp. 599-609, April 1990.
13. C. Lee, L. Rabiner, R. Pieraccini and J. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition," *Computer Speech and Language*, April. 1990, pp. 127-165.
14. H. Murveit, J. Butzberger, V. Digalakis and M. Weintraub, "Large Vocabulary Dictation using SRI's DECIPHER™ Speech Recognition System: Progressive Search Techniques," *Proc. ICASSP*, pp. II-319 - II-322, April 1993.
15. H. Murveit, P. Monaco, V. Digalakis and J. Butzberger, "Techniques to Achieve an Accurate Real-Time Large-Vocabulary Speech Recognition System," this proceedings.
16. D. Pallet, J. G. Fiscus, W. M. Fisher and J, S. Garofolo, "1993 Benchmark Tests for the ARPA Spoken Language Program," this proceedings.
17. D. B. Paul, "The Lincoln Robust Continuous Speech Recognizer," *Proc. ICASSP*, pp. 449-452, May 1989.
18. D. B. Paul and E. A. Martin, "Speaker Stress-resistant Continuous Speech Recognition," *Proc. ICASSP*, pp. 283-286, April 1988.
19. L. R. Rabiner, B. H. Juang, S. E. Levinson and M. M. Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities," *Bell Systems Tech. Journal*, Vol. 64(6), pp. 1211-34, 1985.
20. Wellekens, C., "Explicit Time Correlation in Hidden Markov Models for Speech Recognition," *Proc. ICASSP-87*.
21. S. J. Young, "The General Use of Tying in Phoneme-Based HMM Speech Recognizers," *Proc. ICASSP*, pp. I-569 - I-572, March 1992.