Proceedings of Speech Research Symposium XII, June, 1992.

Multiple-State Context-Dependent Phonetic Modeling with MLPs

Michael Cohen*, Horacio Franco*, Nelson Morgan,**

David Rumelhart*, and Victor Abrash***

* Speech Research Program, SRI International, Menlo Park, CA 94025

** Intl. Computer Science Inst., 1947 Center Street, Suite 600, Berkeley, CA 94704

*** Stanford University, Dept. of Psychology, Stanford, CA 94305

Abstract

Earlier hybrid multilayer perceptron (MLP)/hidden Markov model (HMM) continuous speech recognition systems have not modeled context-dependent phonetic effects, sequences of distributions for phonetic models, or gender-based speech consistencies. In this paper we present a new MLP architecture and training procedure for modeling context-dependent phonetic classes with a sequence of distributions. A new training procedure that "smooths" networks with different degrees of context-dependence is proposed in order to obtain a robust estimate of the context-dependent probabilities. We have used this new architecture to model generalized biphone phonetic contexts. Tests with the speaker-independent DARPA Resource Management database have shown average reductions in word error rates of 20% in both the word-pair grammar and no-grammar cases, compared with our earlier context-independent MLP/HMM hybrid.

1. Introduction

Previous work by Morgan, Bourlard, et al. [1, 2] has shown both theoretically and practically that multilayer perceptrons (MLPs) can be successfully used in a hidden Markov model (HMM) continuous speech recognition system for estimation of the state dependent observation probabilities. Motivations for this approach include relaxation of the restrictive independence assumptions in the computation of the observation likelihoods of traditional HMMs and a reduction in the number of parameters needed for detailed phonetic modeling as a result of increased sharing of model parameters between phonetic classes. Recently, this approach was applied to a state-of-the-art speech recognition system (the SRI-DECIPHER system [3]) in which an MLP provided estimates of context-independent posterior probabilities of phone classes, which were then converted to HMM context-independent state observation likelihoods using Bayes' rule [4]. The best recognition performance on the DARPA Resource Management database with the DECIPHER system has been achieved using a hybrid MLP/HMM that employs a weighted mixture of state observation likelihoods provided by the MLP and the HMM. Table 1 compares the performance of the best pure HMM system with that of the hybrid on three different test sets from the speaker-independent DARPA Resource Management database, in both the (word-pair) grammar and no-grammar cases. Performance improved using the hybrid approach in all six tests, with the reduction in word error ranging from 15% to 30%.

	Grammar		No Grammar	
	HMM	Hybrid	HMM	Hybrid
Feb89	4.6	3.9	21.2	17.2
Oct89	5.1	4.1	23.2	18.9
Feb91	4.6	3.3	19.8	16.9

Table 1: Best pure HMM vs. hybrid MLP/HMM (% word error).

Tests comparing the best DECIPHER pure HMM system with one that replaces the HMM state observation likelihoods with those supplied by an MLP (rather than mixing the HMM and MLP likelihoods) showed better performance by the pure HMM system (which had almost a factor of 40 more parameters than the hybrid). This seems to be due to the simplicity of the MLP used, which did not model context-dependent phonetic effects, multiple sequential distributions for phones, or gender-based speech consistencies. In this paper, we present a new MLP architecture and training method that models context-dependent phonetic effects and allows the modeling of phonetic classes with a sequence of distributions, corresponding to a sequence of HMM states, while training discriminatively at the phonetic level rather than at the subphonetic (HMM-state) level. Ongoing work on the modeling of gender-based speech consistencies with MLPs is described elsewhere [5, 6].

Context-dependent modeling: Experience with HMM technology has shown that using context-dependent phonetic models improves recognition accuracy significantly [7, 8]. This is so because acoustic correlates of coarticulatory effects are explicitly modeled, producing sharper and less overlapping probability density functions for the different phone classes. Context-dependent HMMs use different probability distributions for every phone in every different relevant context. This practice causes problems that are due to the reduced amount of data available to train phones in highly specific contexts, resulting in models that are not robust and generalize poorly. The solution to this problem used by many HMM systems is to train models at many different levels of context-specificity, including biphone (conditioned only on the phone immediately to the left or right), generalized biphone (conditioned on the broad class of the phone to the left or right), triphone (conditioned on the phone to the left and the right), generalized triphone, and word specific phone. Models conditioned by more specific contexts are linearly smoothed with more general models. The "deleted interpolation" algorithm [9] provides linear weighting coefficients for the observation probabilities with different degrees of context dependence by maximizing the likelihood of the different models over new, unseen data. This approach cannot be directly extended to MLP-based systems because the "smoothing" of MLP weights makes no sense. It would be possible to use this approach to average the probabilities from different MLPs, however, since the MLP training algorithm is a discriminant procedure, it would be desirable to use a discriminant procedure to "smooth" the MLP probabilities together.

An earlier approach to context-dependent phonetic modeling with MLPs was proposed by Bourlard et al. [10]. It is based on factoring the context-dependent likelihood and uses a set of binary inputs to the network to specify context classes. The number of parameters and the computational load using this approach are not much greater than those for the original context-independent net.

The context-dependent modeling approach we present here uses a different factoring of the desired context-dependent likelihoods, a network architecture that shares the input-to-hidden layer among the context-dependent classes to reduce the number of parameters, and a training procedure that "smooths" networks with different degrees of context-dependence in order to achieve robustness in probability estimates.

Multidistribution modeling: Experience with HMM-based systems has shown the importance of modeling phonetic units with a sequence of distributions rather than a single distribution. This allows the model to capture some of the dynamics of phonetic segments. The SRI-DECIPHER system models most phones with a sequence of three HMM states. Our initial hybrid system used only a single MLP output unit for each HMM phone model. When this was first extended to three MLP output units for each phone, corresponding to the three states of the HMM phone model, the word recognition error rate increased by almost 30%. A similar result was found in experiments performed at ICSI (personal communication). This seemed to be due to the discriminative nature of the MLP training algorithm. The appropriate level to train discrimination is likely to be at the level of the phone (or higher) rather than at the subphonetic HMM-state level, to which these outputs units correspond. The new architecture presented here accomplishes this by training separate output layers for each of the three HMM states, resulting in a network trained to discriminate at the phone level, while allowing three distributions to model each phone.

2. Hybrid MLP/HMM

The baseline MLP/HMM DECIPHER hybrid (described by Renals et al. [4]) substitutes (scaled) probability estimates computed with MLPs for the tied mixture HMM state-dependent observation probability densities. The topology of the HMM system is kept unchanged.

The hybrid system is bootstrapped from the basic HMM DECIPHER system [3] already trained using the forward-backward maximum-likelihood method. Forced Viterbi alignments for every training sentence provide phone labels, among 69 phone classes, for every frame of speech.

A feed-forward MLP is trained with stochastic gradient descent using these labeled data. A minimum relative entropy between posterior target distribution and posterior output distribution is used. The target distribution is defined as 1 for the index corresponding to the phone class label and 0 for the other classes. With this target distribution, assuming enough parameters in the MLP, enough training data, and that the training does not get stuck in a local minimum, the MLP outputs will approximate the posterior class probabilities $p(q_j|Y_t)$, where q_j corresponds to the *jth* phone class and Y_t is the acoustic vector at time t [1]. Frame classification on an independent cross-validation set is used to control the learning rate and to decide when to stop training as in [4]. The initial learning rate is kept constant until cross-validation performance increases less than 0.5%, after which it is reduced as $\frac{1}{2^n}$ until performance increases no further.

The network architecture consists of an input layer of 234 units, spanning 9 frames of cepstra, delta cepstra, energy, and delta energy features that are normalized to have zero mean and unit variance, a 1000-unit hidden layer, and an output layer with 69 units, one per phone class. Both hidden and output layers consist of sigmoidal units.

During recognition, the posterior class probabilities are converted to (scaled) phone class conditioned observation likelihoods using Bayes' rule.

3. Context-Dependent Multidistribution Hybrid MLP/HMM

The context-independent hybrid MLP/HMM described above has been extended to model context-dependent phonetic classes using a Bayesian factoring of scaled context-dependent posterior phone probabilities computed with an MLP architecture as shown in Figure 1. A separate output layer (consisting of 69 output units corresponding to 69 context-dependent phonetic classes) is trained for each context. Two approaches are used to control the number of parameters: error-based smoothing of context-dependent and independent parameters, and sharing of input-to-hidden weights among all context classes. The context-dependent MLP can be viewed as a set of MLPs, one for each context, which have the same input-to-hidden weights. Separate sets of output layers are used to model context effects in different states of HMM phone models. As a result, since the training proceeds as if each output layer was part of an independent net, the system learns discrimination between the different phonetic classes within an output layer, but does not learn discrimination between the different states of the same phonetic class (which are represented in different output layers).

3.1. Context-dependent factoring

In the HMM framework, every state is associated with a specific phone class and context. During the Viterbi [11] recognition search, $p(Y_t|q_j,c_k)$ (the probability of acoustic vector Y_t given the phone class q_j in the context class c_k) is required for each state. Since MLPs can compute Bayesian posterior probabilities, we compute the required HMM probabilities using

$$p(Y_t|q_j,c_k) = \frac{P(q_j|Y_t,c_k)p(Y_t|c_k)}{P(q_j|c_k)} \tag{1}$$

where $p(Y_t|c_k)$ can be factored again as

$$p(Y_t|c_k) = \frac{P(c_k|Y_t)p(Y_t)}{P(c_k)} \quad . \tag{2}$$

The factor $p(q_j|Y_{t,}c_k)$ is the posterior probability of phone class q_j given the input vector Y_t and the context class c_k. To compute this factor, we use a direct interpretation of the definition of conditional probability, considering the conditioning on c_k in (1) as restricting the set of input vectors only to those produced in the context c_k. If M is the number of context classes, this implementation uses a set of M MLPs (all sharing the same input-to-hidden layer) similar to those used in the context-independent case except that each MLP is trained using only input-output examples obtained from the corresponding context c_k.

Every context-specific net performs a simpler classification than in the context-independent case because in a given context the acoustic correlates of different phones have much less overlap in their class boundaries.

$p(c_k | Y_t)$ can be computed using a standard MLP whose outputs correspond to the context classes. $p(q_j | c_k)$ and $p(c_k)$ are estimated by counting over the training examples. Finally, $p(Y_t)$ is common to all states for any given time frame, and can therefore be discarded in the Viterbi computation, since it will not change the optimal state sequence used to get the recognized string.

3.2. Context-dependent training and smoothing

In order to achieve robust training of context-specific nets, we use the following method:

Initially, a context-independent MLP is trained as in [4] to estimate the context-independent posterior probabilities over the N phone classes. After the context-independent training converges, the resulting weights are used to initialize the weights of the context-specific nets. The context-dependent training proceeds by presenting each training example (the acoustic vector with an associated phone label and context label) only to the appropriate context-specific network. Otherwise, the training procedure is similar to that for the context-independent net, using stochastic gradient descent and a relative entropy training criterion. The overall classification performance evaluated on an independent cross-validation set is used to determine the learning rate as in [4]. Training stops when overall cross-validation performance does not improve any further. In this phase, we are actually training a set of M independent nets, each one trained on a nonoverlaping subset of the original training data.

Every context-specific net would asymptotically converge to the context conditioned posteriors $p(q_j | Y_t, c_k)$ given enough training data and training iterations. Due to the initialization, the net starts estimating $p(q_j | Y_t)$, and from that point it follows a trajectory in weight space (see Figure 2), incrementally moving away from the context-independent parameters as long as classification performance on the cross-validation set improves. As a result, the net retains useful information from the context-independent initial conditions. In this way, we perform a type of nonlinear smoothing between the pure context-independent parameters and the pure context-dependent parameters.

Furthermore, whereas the mixing coefficients are determined with a maximum likelihood approach in the deleted interpolation method, in the method proposed here, the cross-validation classification error is the criterion that determines how much context-dependent learning is effective for discrimination. Thus, the degree of smoothing is based on the point where cross-validation classification error attains a local minimum. Since we start from a good point in the parameter space, training time may be reduced below that needed in the case of random initialization.

This same approach can be extended to handle a hierarchy of context-dependent models that go from very broad context classes to highly specific ones by defining a hierarchy of context classes in which every context class at one level is included in a broader class at the higher level. Every context-specific MLP at a given level in the hierarchy is initialized with the weights of a previously trained context-specific MLP at the higher level in the hierarchy whose associated context class includes that for the MLP being initialized (see Figure 3).

In order to reduce the number of independent parameters to train, we propose an architecture in which all the context-specific nets share the input-to-hidden layer (see Figure 1). Consequently, the hidden layer representation of the acoustic features is shared by all context classes. The different sets of hidden-to-output weights are expected to capture the different acoustic boundaries between phone classes in different contexts.

As a further simplification to speed up training, given that the input-to-hidden weights are the hardest and most time-consuming to train and are already trained in the context-independent training phase, we keep them fixed during the context-dependent training phase. In doing so, we are assuming that the features provided by the hidden layer are useful for context-dependent classification. Thus, the only new parameters to train for every context-specific net are the hidden-to-output weights.

3.3. Multiple states for phone models

In the SRI-DECIPHER system, which the hybrid system extends, two or three states are used to model each phonetic class. In addition to capturing some of the dynamics of phonetic segments, multiple-state phone models allow more precise modeling of context effects because the initial portion of a phone segment is influenced more by the previous phone while the final portion of a phone segment is influenced more by the following phone. In the present approach, we have defined different context classes for different HMM phone states. In the current implementation, two different sets of context classes are used: generalized left-biphone dependent for the first state and generalized right-biphone dependent for the last state of every phone model. For the three state models, the middle state is treated as context-independent (see Figure 4). This approach has the combined advantage of allowing both previous and following context to be modeled for individual HMM phones although only generalized biphone classes are trained and allowing training of the different states of HMM phonetic models in separate output layers so that discrimination is not learned between the different states of the same phone, but only between phonetic classes.

3.4. Recognition

During recognition, as in training, first states of HMM phones are associated with the context-specific MLP output unit according to the context class to which the predecessor phone belongs. Last states of HMM phones are associated with the context-specific output unit according to the context class to which the following phone belongs. Middle states of 3-state HMM phones are associated with a context-independent layer which was trained only on frames that were aligned to middle HMM phone states.

The smoothed context-dependent posterior probabilities supplied by the MLP have to be converted during recognition to (scaled) state-conditioned observation probabilities using the normalization factors provided by Eqs. (1) and (2). However, because these values are a result of smoothing context-dependent and independent networks, the normalization factors should be a combination of those corresponding to the context-dependent and context-independent cases. We use the following heuristics for converting the smoothed posteriors $p^s(q_j|Y_{t,}c_k)$ to smoothed (scaled) observation probabilities $p^s(Y_t|q_{j,}c_k)$:

$$p^s(Y_t|q_j,c_k) = p^s(q_j|Y_t,c_k)\left(\alpha_j^k \frac{1}{p(q_j)} + (1-\alpha_j^k)\frac{p(c_k|Y_t)}{p(q_j|c_k)p(c_k)}\right) \qquad (3)$$

where

$$\alpha_j^k = \frac{N_{ci}(j)}{N_{ci}(j)+b[N_{cd}(j,k)]} \qquad . \qquad (4)$$

$N_{ci}(j)$ is the number of training examples for phone class j for the context-independent net. $N_{cd}(j,k)$ is the number of training examples for the context-dependent net for phone class j and for the context class k. Constant b is optimized on a development set for minimum word-recognition error.

4. Evaluation

Training and recognition experiments with the MLP/HMM hybrid were conducted using the speaker-independent, continuous-speech, DARPA Resource Management database. The vocabulary size is 998 words. Tests were run both with a word-pair (perplexity 60) grammar, and with no grammar. The training set consisted of 3510 sentences (approximately one million frames), and the cross-validation set consisted of 480 sentences. The 3990 sentences making up these two sets comprise the standard DARPA speaker-independent training set for this task. A set of eight left and eight right generalized biphone phonetic context classes were chosen, based principally on place of articulation and acoustic characteristics. These are shown in tables 2 and 3. The differences between Tables 2 and 3 account for the different contextual effects of preceding vs. following diphthongs.

silence	4	q hh hv -
labials	9	b bcl m em w f p pcl v
alvpallat	17	d dcl t tcl s z dh dx th n en sh zh ch jh l el
velars	6	g k gcl kcl ng eng
r	4	r er er+1 axr
round	10	ao+1 ow+1 uw+1 ao ow uw uh uh+1 aw aw+1
unround-low	7	aa+1 ae+1 ah+1 aa ae ah ax
unround-high	12	ih ey iy eh eh+1 ey+1 ih+1 iy+1 y oy+1 ay+1 ay

Table 2: Generalized left biphones.

silence	4	q hh hv -
labials	9	b bcl m em w f p pcl v
alvpallat	17	d dcl t tcl s z dh dx th n en sh zh ch jh l el
velars	6	g k gcl kcl ng eng
r	4	r er er+1 axr
round	9	ao+1 ow+1 uw+1 ao ow uw uh uh+1 oy+1
unround-low	11	aa+1 ae+1 ah+1 aa ae ah ax aw aw+1 ay+1 ay
unround-high	9	ih ey iy eh eh+1 ey+1 ih+1 iy+1 y

Table 3: Generalized right biphones.

For every speech frame, a 12th-order mel cepstrum was computed and 26 coefficients were produced: log energy, 12 cepstral coefficients, and their smoothed derivatives. A 9-frame window of 234 input values was presented as the input vector Y_t to the input layer. The phone class label associated with the central frame defined the target output class. The context class to which the previous or following phone belongs (depending on which phone state the frame was aligned with) determined the context class. Two additional networks were trained to provide the probabilities of the right and left context classes. These networks have 234 input units (receiving the same input vector Y_t as the context-dependent and context-independent networks), 512 hidden units, and 8 output units, one for each context class.

The final cross-validation error for the context-dependent net was 21.4% vs. 30.6% obtained with the context-independent network. Expecting this degree of improvement in actual recognition may be overoptimistic because the context is assumed to be known for this cross-validation error evaluation. Nevertheless, it suggests that the context-dependent architecture is capable of much more detailed modeling of the acoustic variability of the speech signal.

Computational load for context-dependent training was approximately the same as for context-independent training because, although the context-dependent net is significantly larger than the context-independent net, only the corresponding context-specific output layer is updated for each frame presentation. During recognition, the computational load for the context-dependent hybrid was more than four times that of the context-independent hybrid. This is a result of the fact that, because of the huge number of hypotheses that are explored in the Viterbi search, forward propagation is computed for every context-specific output layer for every frame.

Table 4 presents word recognition error for three different speaker-independent Resource Management test sets (February 89, October 89, and February 91) for both the (word-pair) grammar and no-grammar cases. In all six tests, we see reductions in word error rates using the new architecture, with an average reduction in word error of 20%.

	Grammar		No Grammar	
	CI-MLP	CD-MLP	CI-MLP	CD-MLP
Feb89	5.4	4.7	24.9	19.4
Oct89	7.6	5.7	27.0	20.8
Feb91	6.2	5.0	25.0	20.5

Table 4: Context-independent vs. context-dependent MLP (% word error).

5. Discussion

5.1. Refining the hybrid MLP/HMM

Two principal problems must be resolved in order to model context-dependent phonetic classes with a sequence of distributions in a hybrid MLP/HMM system:

[1] Context-dependent phonetic modeling requires many more parameters than context-independent phonetic modeling. Many contexts will be poorly represented in training databases of realistic size, which can lead to models that lack robustness.

[2] A straightforward extension of our original MLP to the modeling of multiple distributions for phonetic classes leads to discriminative training between subphonetic classes (corresponding to HMM states). It is likely that the appropriate level for discriminative training is at the phonetic level (or higher).

Parameter explosion: In standard HMMs, most of the parameters in the system are in the distributions associated with the individual states. MLPs use representations that are more distributed in nature, allowing more sharing of representational resources and better allocation of representational resources based on training. We exploit this characteristic of MLPs in the approach described here by sharing the input-to-hidden layer weights among all context classes. This sharing substantially reduces the number of parameters to train and the amount of computation required during both training and recognition. The context-dependent MLP has 17 times as many output units (classes) as the context-independent MLP, but has only a factor of 4.6 times as many parameters. In addition, we do not adjust the input-to-hidden weights during the context-dependent phase of training, assuming that the features provided by the hidden layer activations are relatively low level and are appropriate for context-dependent as well as context-independent modeling. This procedure substantially reduces context-dependent training time. The large decrease in cross-validation error going from context-independent to context-dependent MLPs suggests that the features learned by the hidden layer during the context-independent training phase, combined with the extra modeling power of the context specific hidden-to-output layers, were adequate to capture the more detailed context-specific phone classes.

The other approach used to increase the robustness of the context-dependent models is error-based smoothing, in which the context-dependent net is initialized with context-independent weights, and training proceeds until a minimum in cross-validation performance is reached. This results in a network that uses a combination of context-independent and context-dependent information in order to maintain performance for poorly represented contexts by relying more heavily on context-independent training for those cases. This approach serves the same purpose as the deleted interpolation algorithm in the training of standard context-dependent HMMs, but uses cross-validation error as the criterion to determine the amount of context-dependent learning that is effective for discrimination.

Subphonetic discrimination: A direct extension of our original hybrid approach to handle a sequence of distributions for phonetic classes (corresponding to multiple HMM states), based on increasing the size of the output layer to handle the additional classes, resulted in a degradation in recognition accuracy. The decline seemed to be due to the fact that the MLP was being trained to discriminate subphonetic classes. As a result, the MLP was attempting to discriminate into separate classes acoustic vectors that corresponded to the same phone and in many cases were very similar, but were aligned with different HMM states. There were likely to have been many cases in which almost identical acoustic vectors were labeled as a positive example in one instance and a negative example in another for the same output class. The use of separate output layers corresponding both to different states of an HMM phonetic model and to different contexts allows the training of multiple distributions for phonetic models, without training discriminatively between the states of a single HMM phonetic model. In addition, this approach avoids discriminative training between output units associated with the same phone in different contexts.

5.2. Comparison with earlier systems

The results shown in Table 4 suggest that these techniques were successful in improving performance of our original hybrid MLP/HMM system, which was based on a simpler MLP. Comparing the results in Tables 1 and 4 shows that our new hybrid (in which we did not mix MLP and HMM supplied probabilities) does not perform as well as our best case which mixes MLP and HMM probabilities, or as well as our best pure HMM system. The MLP described here is still far simpler than our best pure HMM, with approximately a factor of eight fewer parameters, modeling of only generalized biphone phonetic contexts (the HMM models a full hierarchy of phonetic contexts including generalized biphone, specific biphone, generalized triphone, specific triphone, and word-specific phone), and no modeling of gender-specific speech consistencies. In the future, we will extend the current approach to the modeling of finer phonetic contexts and incorporate models of gender-specific speech consistencies. The new MLP developed here, and the planned extensions, can be used in a system that mixes MLP and HMM supplied probabilities, although we expect that the planned extensions to our current MLP model will lead to a system that performs better than our current best system (which mixes MLP and HMM probabilities) without the need for such mixing, therefore resulting in a simpler system.

6. Conclusions

The limitations of earlier MLP/HMM hybrids include the lack of modeling of context-dependent phonetic effects, sequences of distributions for phonetic models, and gender-based speech consistencies. We have presented a new MLP architecture and training procedure for modeling context-dependent phonetic classes with a sequence of distributions. Tests using the DARPA Resource Management database have shown improvements in recognition performance using this new approach, modeling only generalized biphone context categories. These results suggest that sharing input-to-hidden weights between context categories (and not retraining them during the context-dependent training phase) results in a hidden layer representation which is adequate for context-dependent as well as context-independent modeling, error-based smoothing of context-independent and context-dependent weights is effective for training a robust model, and using separate output layers and hidden-to-output weights corresponding to different context classes of different states of HMM phone models is adequate to capture acoustic effects which change throughout the production of individual phonetic segments.

Acknowledgments

Support for this research project was provided by DARPA Contract MDA904-90-C-5253.

References

[1] H. Bourlard, and N. Morgan, ‘‘Merging Multilayer Perceptrons and Hidden Markov Models: Some Experiments in Continuous Speech Recognition,’’ in E. Gelenbe, Ed., Neural Networks: Advances and Applications, Amsterdam, North Holland Press, 1990.

[2] N. Morgan and H. Bourlard, ‘‘Continuous Speech Recognition Using Multilayer Perceptrons with Hidden Markov Models,’’ ICASSP 90, pp. 413-416, Alburquerque, New Mexico, 1990.

[3] H. Murveit, M. Cohen, P. Price, G. Baldwin, M. Weintraub, and J. Bernstein, ‘‘SRI’s DECIPHER System,’’ DARPA Speech and Natural Language Workshop, February 1989.

[4] S. Renals, N. Morgan, M. Cohen, H. Franco, ‘‘Connectionist Probability Estimation in the DECIPHER Speech Recognition System,’’ ICASSP 92, Vol. 1, pp. 601-604, San Francisco, 1992.

[5] V. Abrash, H. Franco, M. Cohen, N. Morgan, and Y. Konig, ‘‘Connectionist Gender Adaptation in a Hybrid Neural Network/Hidden Markov Model Speech Recognition System,’’ in press, proceedings ICSLP, Banff, Canada, 1992.

[6] Y. Konig, N. Morgan, and C. Chandra, ‘‘GDNN: A Gender-Dependent Neural Network for Continuous Speech Recognition,’’ International Computer Science Institute Technical Report TR-91-071, December 1991.

[7] 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,’’ ICASSP 85, 1205-1208, 1985.

[8] Kai-Fu Lee, ‘‘Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition,’’ IEEE Trans. on Acoust., Speech and Signal Proc., Vol 38, No. 4, April 1990.

[9] F. Jelinek and R. L. Mercer, ‘‘Interpolated estimation of markov source parameters from sparse data,’’ in Pattern Recognition in Practice, E. S. Gelsema and L. N Kanal, Eds. Amsterdam: North-Holland, 1980, pp. 381-397.

[10] H. Bourlard, N. Morgan, C. Wooters, S Renals, ‘‘CDNN: A Context Dependent Neural Network for Continuous Speech Recognition,’’ ICASSP 92, Vol. 2, pp. 349-352, San Francisco, 1992.

[11] S. E. Levinson, L. R. Rabiner, and M.M. Sondhi, ‘‘An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition,’’ Bell Syst. Tech. Journal 62, 1035-1074, 1983.