Proceedings of International Conference on Spoken Language Processing, October, 1992.

CONNECTIONIST GENDER ADAPTATION IN A HYBRID NEURAL NETWORK / HIDDEN MARKOV MODEL SPEECH RECOGNITION SYSTEM

Victor Abrash, Horacio Franco, Michael Cohen †
Nelson Morgan, Yochai Konig ‡

† Speech Research Program, SRI International, Menlo Park, CA 94025
‡ Intl. Computer Science Inst., 1947 Center Street, Suite 600, Berkeley, CA 94704

ABSTRACT

An approach to modeling long-term consistencies in a speech signal within the framework of a hybrid Hidden Markov Model (HMM) / Multilayer Perceptron (MLP) speaker-independent continuous-speech recognition system is presented. Several ways to model male and female speech more accurately with separate models are discussed, one of which is investigated in depth. A method which combines gender-independent and -dependent MLP training is demonstrated, improving recognition accuracy while retaining robustness. A series of network architectures (using our training method) for the connectionist estimation of gender-dependent HMM observation probabilities are evaluated in terms of recognition performance and number of additional parameters needed. Experimental evalutation shows a significant improvement in word recognition accuracy over the gender-independent system with a moderate increase in the number of parameters.

1. INTRODUCTION

Bourlard and Morgan [1,8] have demonstrated that multilayer perceptrons (MLPs) can successfully estimate the state-dependent observation probabilities in a hidden Markov model (HMM) based speech recognition system [12]. We have incorporated MLP probability estimation into the SRI-DECIPHER system. The MLP in our hybrid system computes the posterior probabilities of context-independent phone classes, which are then converted to HMM state observation likelihoods using Bayes's rule. The hybrid HMM/MLP system combines discriminative MLP training procedures with the sequential mapping strengths of conventional HMM speech recognizers.

A weakness in the underlying HMM framework is the output-independence assumption, which states that the probability of observing any given acoustic vector depends on only the current HMM state. With the HMM considered as a stochastic speech generator, outputs from frame to frame are assumed independent of each other regardless of the exact generating process (or state). A purely context-independent HMM is a memoryless system, which cannot model short or long-term correlations in its acoustic input. This simplifying assumption, although necessary to reduce the size of the parameter estimation problem, is obviously inaccurate; all portions of any given utterance are produced by the same speaker, using the same vocal tract, in the same acoustic environment, and share a common dialect, vowel space, and speaker gender. Context-dependent HMM's partially overcome these deficiencies using our linguistic knowledge and a richer parameter space to partially model acoustic correlations based on phonetic context. The use of delta parameters as additional speech features in state of the art systems provides a way of modeling short-time correlations explicitly.

The purpose of consistency modeling is to model long-term consistencies in speech using training databases of practical size. Consistent features include the identity of a speaker, speaker vocal tract size and shape, vowel space, dialect, gender, or speech rate or style. For example, the training data could be clustered according to speaker height, which may be correlated with vocal tract length. Separate HMM models could be estimated from each data cluster, with recognizers using each set of models run in parallel. The most probable hypothesis from all available recognizers would then be chosen as the best sentence hypothesis. Alternatively, the HMM parameters could be made a function of height, the value of which would adapt the speech recognition process. Finally, height-dependent models could be combined with or adapted from initial height-independent ones. One natural first choice of consistent feature to model is speaker gender; with only two values, each data cluster has a maximal size. Furthermore, there are obvious differences between male and female speech, and it is possible to reliably estimate speaker gender.

Past experiments with separate modeling of male and female speech in HMMs have shown improved recognition performance [10, 11], given enough training data. In DECIPHER, the usual approach doubles the number of parameters to be estimated by training gender-specific models on male/female partitioned datasets. Very large amounts of training data were necessary to make a significant improvement [10]. Although they attempt to model more self-consistent speech, each model is estimated on only half the amount of training data, making them less robust to novel speakers. A better way (not implemented in DECIPHER) to model speaker gender effects in the HMM framework may be to combine gender-independent (GI) and gender-dependent (GD) models, using the deleted interpolation algorithm [5], which solves the reduced training set size problem. This approach still has the disadvantage of doubling parameter memory needs, as compared to GI systems.

This paper describes our first attempts at consistency modeling within the MLP component of the hybrid HMM/MLP DECIPHER system. We chose to model speaker gender consistency initially. In the future we will be able to compare results with separate male/female modeling in the pure HMM Decipher system.

In the future, we plan to extend our results to other sources of pronunciation consistency, and to combine our gender- and context-dependent [2,3,4] approaches. Since other consistency classes may have more than two natural clusters (speaker-dialect region has eight values in our speech database, for example), a naive data-partitioning approach where we train separate models from disjoint training subsets could result in too little data for robust training. Therefore, a secondary research goal was to minimize the number of additional parameters needed for our consistency-modeling approach. The approaches described here both test a smoothing method to combine gender-independent and gender-dependent parameters, and compares a number of alternative MLP architectures which limit the duplication of parameters to small regions of the entire parameter space, under the assumption that large portions of the acoustic model of English should be the same for males and females.

2. GENDER INDEPENDENT HYBRID HMM/MLP

The baseline hybrid HMM/MLP DECIPHER speech-recognition system [2,12] replaces the tied-mixture HMM state-dependent observation probability densities with (scaled) probability estimates computed by a MLP, keeping the HMM topology unchanged. The MLP architecture is a feed-forward network with 234 inputs, 512 hidden units, and 69 outputs. Input units are linear and all others are sigmoidal. The 234 inputs represent 9 frames of cepstra, delta cepstra, log energy, and delta log energy speech features as described in Section 5 and [9].

The training of the hybrid system is described elsewhere in this volume [2]. To summarize, the MLP is trained using stochastic gradient descent and a relative-entropy error criterion. The 1 of N target distribution is defined as 1 for the output index corresponding to the correct phone class label and 0 otherwise. Assuming we have enough training data, choose an appropriate number of parameters in the MLP, and that training does not get stuck in poorly performing local minima, the MLP will approximate [13] the posterior class probabilities $P(q_j \mid Y_t)$, where q_j corresponds to the j-th phone class and Y_t is the acoustic vector at time *t*. Frame classification performance on an independent cross-validation set is used to control the learning rate and decide when to halt training. Learning rate is controlled as in Cohen [2].

The hybrid system is bootstrapped from the pure HMM DECIPHER system [10]. Target labels for the MLP outputs, representing each of 69 context-independent phone classes, are obtained from the HMM forced-Viterbi alignments.

During recognition, Bayes's rule is used to convert the network outputs to the (scaled) phone-class conditional observation likelihoods required by the HMM,

$$P(Y_t \mid q_j) = \frac{P(q_j \mid Y_t)\, P(Y_t)}{P(q_j)} \qquad (1)$$

where $P(q_j \mid Y_t)$ is the network output. The prior probability distribution, $P(q_j)$, of the phone classes over the MLP training set is obtained by counting over the training target data. $P(Y_t)$, the pdf for the acoustic vector, is the same over all competing HMM paths and can be ignored as it contributes equally to all sentence hypotheses. These MLP observation likelihoods replace those formerly generated by tied Gaussian mixtures in the baseline system.

3. GENDER DEPENDENT HYBRID HMM/MLP

A GD MLP is trained to compute $P(q_j \mid Y_t, gender)$, the posterior probability of a context-independent phone class given the current acoustic vector and the speaker's gender. Network architecture and training are described in Section 4.

To use the network in the hybrid HMM/MLP DECIPHER, the network's output must be converted to observation likelihoods using Bayes's rule,

$$P(Y_t \mid q_j, gender) = \frac{P(q_j \mid Y_t, gender) \times P(Y_t \mid gender)}{P(q_j \mid gender)} \qquad (2a)$$

$$= \frac{P(q_j \mid Y_t, gender)}{P(q_j \mid gender)} \times \frac{P(gender \mid Y_t)\, P(Y_t)}{P(gender)} \qquad (2b)$$

where $P(q_j \mid Y_t, gender)$ is the MLP output and $P(q_j \mid gender)$ is the prior probability of each phone class over the male or female portions of the MLP training set, respectively. All experiments reported here obtain $P(q_j \mid gender)$ by counting the examples of each phone class in the MLP training set.

The factor $P(gender \mid Y_t)$ is the posterior probability of the speaker's gender given the acoustic evidence. $P(gender)$ is the prior probability of male or female utterances, and is simply the proportion of male or female speech frames in the training set. As before, $P(Y_t)$ can be ignored in recognition. We obtain $P(gender \mid Y_t)$ from a second, gender-classification neural network, which estimates speaker sex with greater than 90% accuracy [6,7].

Gender-dependent speech recognition is performed by first combining each of the two gender specific MLPs with the GI HMM to produce male and female subrecognizers, which are run in parallel on unknown speech. The sentence hypothesis with the highest probability from either recognizer is then chosen.

As a way of simplifiying the system, we ran the subrecognizers without the gender classifier $P(gender \mid Y_t)$. In these experiments, however, the male sentence hypothesis was chosen a disproportionate number of times, leading to reduced recognition accuracy. The GD MLPs used to compute the observation likelihoods in Eqn. (2) were trained with twice as much male as female data, perhaps leading to a stronger "male" output activation and hence a higher male sentence likelihood, even when male and female networks were given the same acoustic input.

4. GENDER-DEPENDENT MLPS

We had two goals in this investigation. First, we wanted to increase speech recognition accuracy in our hybrid HMM/MLP system through better modeling of acoustic regularities within speech produced by the same speaker. Although there are some fundamental differences between male and female speech (fundamental frequency, or pitch, being the most obvious) there are many more linguistic similarities that should be exploited for better recognition.

Simultaneously, we wanted to explore the cost-performance trade-off associated with the number of additional parameters needed for gender modeling or adaptation. Our standard HMM technique for modeling gender-consistency performs poorly on this scale, yielding a small performance boost at high cost. In this scenario, the training data are first partitioned into male and female sets, which are used to train separate models. Because of the increased similarity of the speech within the separate training sets, the acoustic models will have sharper distributions producing improved recognition performance. Unfortunately, without smoothing of the GI and GD models (using deleted interpolation for example) the GD models will be at best trained with only half the data available for GI models, so there may not be enough training data to robustly estimate the model parameters. Gender modeling in HMM systems has therefore significantly improved recognition performance only when large amounts of training data are available. Furthermore, this separate modeling technique may take advantage of gender-consistent features of the speech but decrease the usefulness of other consistent features. The decrease would be further exacerbated if separate models were constructed from smaller portions of the data.

In the connectionist domain, we could also choose to partition the training data and construct entirely separate models, with the same advantages and disadvantages described for HMMs. However, due to the distributed nature of MLP-based representations, it is straightforward to design MLP architectures which share part of the parameter space, and only split off a portion of the parameters to model gender-specific characteristics. It is also simple to add extra M/F inputs to the network, making network parameters into a function of speaker gender or other characteristics; we will investigate this approach in detail in a future paper.

In neural network training, the starting point in weight space often determines the final solutions available to the network. In this work, we decided to perform gender dependent training by adjusting our weights only incrementally from a starting point defined by the fully trained gender-independent parameters, keeping much of the information encoded in the original network. Conceptually, this procedure provides some degree of non-linear smoothing between our initial gender-independent and final gender-dependent parameters [4]. Furthermore, this method reduces the training time needed obtain gender-dependent networks (compared with training from random initial weights).

We constructed and trained a series of networks which share different subsets of their weight matrices, ranging from a case which increases the number of parameters in the GD network by 512 over the number in the GI MLP, to one which increases the number of weights by 155,717. This last network has no parameter sharing, thereby doubling the size of the GI network. In practice, the GD MLPs are first initialized with the GI weights, and then trained separately on male and female training subsets with some connection strengths fixed, according to the architectural definition (See figure 2). The labels (G1—G5) assigned to the different network architectures will be used throughout this paper.

G1: Gender-dependent (GD) hidden layer biases: This network had the fewest number of additional weights (512), since we are replacing the original, "always on" hidden bias input with two M/F bias inputs, one of which is always on and the other off. If we interpret the hidden layer units as feature detectors, the different biases adjust the activation of each detector, depending on the gender.

G2: GD output weights and biases: This architecture is based on the assumption that the hidden layer activations represent relatively low level acoustic features, and allows the separate M/F sets of hidden-to-output weights to use those features in different ways. This architecture gives 35,397 more weights than the GI network.

G3: GD input-hidden layer weights and biases: This approach is based on the assumption that during the initial GI training the hidden layer activations learn to represent a useful set of features for phonetic classification. The goal of GD training is to allow the input-to-hidden weights to learn alternative (refined) mappings from the input acoustics to the hidden layer features. This architecture results in 120,320 more weights than the GI network.

G4: GD hidden-output weights, hidden and output biases: This network is a combination of G1 and G2. The motivation for this architecture was similar to G2, but with added flexibility to alter the hidden layer features. This approach results in 35,397 more parameters than the GI MLP.

G5: All weights are gender dependent: This architecture is the extreme case, in which we use two separate networks which do not share any of their parameters. However, this network uses the same approach to smoothing GI and GD weights as G1—G4, gaining full flexibility to refine the speech features learned by the GI network. This architecture doubles the number of parameters (from 155,717 to 311,434) of the GI network.

5. EXPERIMENTAL PROCEDURE

Training and recognition experiments used the speaker-independent, continuous-speech DARPA Resource Management (RM) database, which has a vocabulary size of 998 words. For MLP training, the standard DARPA 3,990 sentence training set was subdivided into 3,510 training (2,430 male, 1,080 female) and 480 (400 male, 80 female) cross-validation sentences, corresponding to 1.17 and 0.16 million speech frames respectively. The testing data consisted of 600 sentences from the Feb89 and Oct89 DARPA test sets (5,245 words). We ran tests with both the standard word-pair (perplexity 60) grammar and with no grammar.

A 12th order Mel cepstrum was computed for each 10-ms speech frame to produce an input speech vector with 26 components: log energy plus 12 cepstral coefficients, and their first derivatives. A nine-frame window of 234 input values was presented as the network input vector Y_t, and the phone class label associated with the central frame was defined as the target class. Means and variances were calculated over the 3,510 sentence MLP training set, and were used to normalize each input acoustic vector to zero mean and unit variance. All GD networks trained from GI initial weights continued to use GI mean and variance vectors.

An initial gender-independent MLP with 512 hidden units was trained. The GD MLP weight matrices were initialized with these GI weights rather than from random values, and then trained with a small learning rate for from one to three additional iterations (depending on cross-validation performance) on the partitioned male and female training sets. The same mean and variance normalization vectors used with the GI training were used in the GD training phase, since the initial weights were optimized with respect to input data preprocessed with these values.

For comparison, two additional networks were trained from random weights on the gender-partitioned data. One of these used the same architecture as G5 (without the GI smoothing), the other was similar but used a reduced 50-ms input window (130 inputs). The input vectors were normalized with means and variances computed from the partitioned data, to allow them to take advantage of the different range of acoustic values present in the male and female speech data.

Recognition results over an independent, 600-sentence test set are shown in Tables 1 and 2. Table 1 shows the word recognition accuracy using an all-word grammar, where any word can follow any other word in the vocabulary. Word error rate is defined as

$$WE\,(\%) = \frac{insertions\ +deletions\ +substitutions}{total\ \ number\ \ of\ \ words} \times 100 \qquad (3)$$

Table 2 shows the word error rate using the standard DARPA perplexity 60 word-pair grammar (which allows only a predefined subset of the vocabulary words to follow any given word).

The first column in each table is obtained by summing the word errors from the highest probability sentence hypothesis from the male and female subrecognizers. The second column shows recognition error when the MLP output is used as defined in equation (2), using the auxiliary gender classification network. The third column shows what the hypothetical word error rate would be if a 100% accurate gender classifier were available.

The recognition score for each grammar case was significantly improved by using our hybrid HMM/MLP over the pure context-independent HMM system (not the state-of-the-art DECIPHER). For each GI-initialized, GD MLP G1—G5 we saw an additional small but consistent improvement compared with the GI MLP.

Using the same number of units but training the "Separate 5/9" networks from scratch on separate on M/F partitioned data (ie, without GI smoothing) yielded worse performance, probably because we no longer had enough training data to estimate the parameters adequately. A smaller network with fewer weights did better, but did not improve on our previous GI results.

The best reduction in word error rate achieved by our gender-dependent networks was 6.78% (with GD) in the no-grammar case, and 4.48% (with G5) using the word-pair grammar (compared to the GI hybrid). The difference between GI and G3 is significant at the 95% level in the all-word grammar case. None of the differences are significant in the grammar case.

Table 1: Word Error Rate (%) with No grammar. Tested on combined Feb89 and Oct89 DARPA test sets (13 male and 7 female speakers, 5245 words).

Network	Without Classifier	With Classifier	Told Correct Gender
Baseline - pure HMM	44.84	N/A	N/A
Baseline - GI MLP	30.07	N/A	N/A
G1	29.82	29.59	29.48
G2	29.59	28.90	28.94
G3	28.37	28.03	27.89
G4	29.38	28.58	28.64
G5	28.35	28.14	28.03
Separate - 9 frames	36.57	35.35	35.20
Separate - 5 frames	29.99	30.07	29.76

Table 2: Word Error Rate (%) with Word-Pair grammar. Same test set as in Table 1.

Network	Without Classifier	With Classifier	Told Correct Gender
Baseline - pure HMM	14.01	N/A	N/A
Baseline - GI MLP	7.82	N/A	N/A
G1	7.68	7.68	7.68
G2	7.70	7.55	7.51
G3	7.70	7.53	7.42
G4	7.70	7.51	7.45
G5	7.49	7.47	7.47
Separate - 9 frames	11.73	15.21	18.21
Separate - 5 frames	8.08	8.16	8.01

6. DISCUSSION

Earlier, we specified three possible ways to introduce gender consistency into our networks:

(1) Train separate models on smaller, gender-partitioned data sets.

(2) Make model parameters a function of both the acoustics and the speaker's gender.

(3) Initializing GD MLP training with GI weights to create robust GD networks.

In this paper, we have presented results for approaches (1) and (3); work on method (2) is in progress, but is not yet ready for presentation.

Training separate GD models from scratch with no GI smoothing gave much worse results than a purely GI MLP, probably because there was not enough training data to fully train two complete sets of weights. We tried recognition with a smaller MLP to reduce the number of free parameters, producing better results, but still not as good as the smoothed GD/GI networks.

GI and GD smoothing improved recognition accuracy in all cases, though only by a small amount. Larger improvements were reported by Konig [6]. However, the experiments reported there started with a simpler recognizer with poorer performance, leaving more room for improvement.

We initialized our male and female MLPs from a good set of gender-independent weights and then trained them separately on partitioned male and female training subsets. We used a cross-validation halting criterion to guarantee good generalization without losing robustness. Mixing GI and GD training with cross-validation guaranteed that GD phone-classification performance never dropped below GI levels, at least for our cross-validation set.

We trained five different MLP architectures with the mixed procedure to explore the modeling power of each network connection layer, and to discover the relationship between the number of free GD parameters and recognition performance; in GD training mode, only a subset of weights were changed, all others remained fixed at their initial GI values. Of the five, only architectures G3 and G5 (see Section 4) resulted in statistically significant gains in word-recognition accuracy (only in the no-grammar case). The common feature of both these MLP architectures was GD input-hidden layer weights. The other networks, G1, G2, and G4, modified only hidden-output connections or biases, and did not produce (statistically) significantly improved accuracy. It seems that the GD adjustment of input-to-hidden weights was necessary to achieve any improvement in recognition performance. The difference in performance between networks G3 and G5 were very small, and not significant.

We speculate that the (hidden layer) features learned by the original, initializing MLP were primarily independent of speaker gender. Training gender dependent input to hidden layer weights allows the MLP to better extract these features from the input speech, taking advantage of any differences between the purely male or female cepstral data. In the future, we will try to confirm this hypothesis by computing correlations between hidden layer activations and speaker gender.

Finally, our gender adaptation scheme worked better when not using any grammar than when we used the DARPA word-pair grammar. The word-pair grammar strongly constrains word order within a sentence, probably enough to mask small improvements in MLP performance.

7. CONCLUSIONS

We have shown a new way to train GD MLPs for speech recognition with minimum cost beyond the effort of training an initial GI network. This approach allows the smoothing of GI and GD parameters. A range of MLP architectures, sharing different subsets of weights between genders, was tested. In all cases, we saw a small but consistent improvement in word recognition accuracy.

In the future, we plan to finish investigating training that makes model parameters a function of both acoustics and speaker gender. We will also combine our GD MLPs with GD HMM models and investigate the robustness of the final hybrid to new speech, comparing it to GD HMM models obtained with deleted interpolation. Finally, we plan to explore the use of smoothed priors, and combine our results with the context-dependent modeling scheme described by Franco [4].

ACKNOWLEDGMENTS

Support for this research project was provided by DARPA contract MDA 904-90-C-5253. The first author was also partially supported by a National Science Foundation Graduate Fellowship. Any opinions, findings, conclusions or recommendations are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Figure 1: Block Diagram of a Gender-Dependent Hybrid HMM/MLP Speech Recognizer.

Figure 2: MLP Architecture: The GI MLP has 234 inputs, 512 hidden units, and 69 outputs. Each output unit represents one context-independent phone class. The GD networks G1—G5 are initialized with a set of fully trained GI weights; during GD training, only those weights covered by the grey bars labeled with the appropriate network names *(G1—G5)* are adjusted, all others remain fixed at their GI values.

REFERENCES

1. H. Bourlard and N. Morgan, ''Merging Multilayer Perceptrons and Hidden Markov Models: Some Experiments in Continuous Speech Recognition,'' in *Neural Networks: Advances and Applications*, ed. E. Gelenbe, North Holland Press, Amsterdam (1990).
2. M. Cohen, H. Franco, N. Morgan, D. Rumelhart, and V. Abrash, ''Multiple-State Context-Dependent Phonetic Modeling with MLPs,'' *Speech Research Symposium XII*, (June 1992).
3. M. Cohen, H. Franco, N. Morgan, D. Rumelhart, and V. Abrash, ''Hybrid Neural Network/Hidden Markov Model Continuous Speech Recognition,'' *ICSLP*, (1992).
4. H. Franco, M. Cohen, N. Morgan, D. Rumelhart, and V. Abrash, ''Context-Dependent Connectionist Probability Estimation in a Hybrid HMM-Neural Net Speech Recognition System,'' Submitted to *IJCNN*, (November 1992).
5. X. Huang, Y. Ariki, and M. Jack, '',,'' pp. 212 in *Hidden Markov Models for Speech Recognition*, Edinburgh Information Technology Series, Edinburgh University Press, Edinburgh (1990).
6. Y. Konig, N. Morgan, C. Wooters, and V. Abrash, ''Modeling Consistency in a Speaker Independent Continuous Speech Recognition System,'' Submitted to *NIPS*, (November 1992).
7. Y. Konig and N. Morgan, ''GDNN: A Gender-Dependent Neural Network for Continuous Speech Recognition,'' *IJCNN*, (June, 1992).

8. N. Morgan and H. Bourlard, ‘‘Continuous Speech Recognition Using Multilayer Perceptrons with Hidden Markov Models,’’ *ICASSP*, pp. 413-416 (1990).

9. H. Murveit, M. Cohen, P. Price, G. Baldwin, M. Weintraub, and J. Bernstein, ‘‘SRIs DECIPHER System,’’ *DARPA Speech and Natural Language Workshop*, (February 1989).

10. H. Murveit, M. Weintraub, and M. Cohen, ‘‘Training Set Issues in SRI’s DECIPHER Speech Recognition System,’’ *DARPA Speech and Natural Language Workshop*, (June 1990).

11. D. Paul, ‘‘The Lincoln Continuous Speech Recognition System: Recent Developments and Results,’’ *DARPA Speech and Natural Language Workshop*, (February 1989).

12. S. Renals, N. Morgan, M. Cohen, and H. Franco, ‘‘Connectionist Probability Estimation in the DECIPHER Speech Recognition System,’’ *ICASSP* **1** pp. 601-604 (1992).

13. M. Richard and R. Lippmann, ‘‘Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities,’’ *Neural Computation* **3**(4) pp. 461-483 (Winter 1991).