GDNN: A Gender-Dependent Neural Network for Continuous Speech Recognition

Yochai Konig [*] [†]
Nelson Morgan [*] [†]
Claudia Chandra [†]

TR-91-071

December 1991

Abstract

Conventional speaker-independent speech recognition systems do not consider speaker-dependent parameters in the probability estimation of phonemes. These recognition systems are instead tuned to the ensemble statistics over many speakers. Most parametric representations of speech, however, are highly speaker dependent, and probability distributions suitable for a certain speaker may not perform as well for other speakers. It would be desirable to incorporate constraints on analysis that rely on the same speaker producing all the frames in an utterance. Our experiments take a first step towards this speaker consistency modeling by using a classification network to help generate gender-dependent phonetic probabilities for a statistical recognition system. Our results show a good classification rate for the gender classification net. Simple use of such a model to augment an existing larger network that estimates phonetic probabilities does not help speech recognition performance. However, when the new net is properly integrated in an HMM recognizer, it provides significant improvement in word accuracy.

[*] International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704

[†] EECS Department, University of California at Berkeley, Berkeley, CA 94720

1 Introduction

Reports over the last few years have described improvements in continuous speech recognition using a Multilayer Perceptron (MLP) to estimate output probabilities for Hidden Markov Models (HMMs). Bourlard and Wellekens [2] showed the relationship between minimizing a squared-error cost function and estimating Bayesian probabilities for the multiclass case (see Richard and Lippmann [6] for a good overview of theory and results for that problem). In particular, experiments by Bourlard and Morgan [1] have shown that the MLP outputs (when divided by the prior classification probabilities) can be used as estimates of the emission probability of HMM's. Particularly when this approach is used to incorporate contextual information at the input of the MLP, significant performance improvement over non-discriminant HMM training procedures is observed. However, in common approaches to speaker-independent recognition (both pure HMM and hybrid MLP-HMM), speaker-dependent parameters are not considered during training. The probability estimators are instead trained using the entire ensemble of data from many speakers. Most parametric representations of speech, however, are highly speaker dependent, and probability distributions suitable for a certain speaker may not perform as well for other speakers. Examples of speaker-dependent parameters are differences in regional accents and the length of male and female vocal tracts.

A more general view of this limitation is the lack of consistency constraints. Each time that one estimates HMM output probabilities, the analysis ignores the fact that the speaker is consistent for the whole utterance. That is, each frame-wise analysis should be constrained to models that assume the same vocal tract and basic glottal mechanism. This suggests that speaker-dependent feature estimation should be done over a large window of the input waveform, since these features are not local properties of a small input window.

As has been observed for some mainstream HMM systems [5], given enough training data, separate phonetic models for male and female speakers can be used to improve performance. Our first attack on consistency, then, is to train an MLP to estimate the probability of gender. This probability is then integrated into an existing MLP-HMM hybrid recognizer.

2 The Classification Net

The first task is to build a net that predicts whether the speaker of a speech input is male or female. The output of the classification net is taken as the probability $P(Gender|Data)$, and will be used in some manner to be described later to improve overall recognition.

In our experiments we tried to answer the following questions:

- What is the best set of features, that we want to extract from speech?
- What is the best configuration for the classification network?

Experiments were done using the 109-speaker DARPA Resource Management corpus. As we have done in our group for previous experiments with this corpus, 3511 sentences were used for the training set and 479 in the cross validation set.

In our first experiment we used 12 mel-cepstral features and energy along with their first derivatives. The features were calculated from 20ms of speech. The networks that we use only have one hidden layer. The number of input features are 26, and there are two output units. One for each gender. The results are summarized in table 2.

From these first set of results, we tried to improve the recognition rate by using linear prediction coefficients instead of mel-cepstrum since they capture the speaker-dependent features better. We used 12 coefficients and energy along with their first derivatives. We started with an analysis window of 20ms.

At this point we realized that we had to use a bigger analysis window to properly model consistency and to augment the feature set with features that separate males and females. After some

Number of Hidden Units	*Temporal Window Length*	*Cross Validation Set Error*
32	5	28.7%
32	9	28.0%
16	5	25.1%
16	9	25.1%

Table 1: Results for Mel-Cepstrum Features

Number of Hidden Units	*Temporal Window Length*	*Cross Validation Set Error*
32	5	28.4%
32	9	27.9%
16	5	29.5%

Table 2: Results for the LPC12 Features

experimentation, the LPC coefficients and energy were calculated over a 500 msec window, and were augmented by the fundamental frequency (plus derivatives of all these features), for a total of 28 inputs. The MLP had one hidden layer. Experiments showed little difference in performance over a range of hidden layer sizes, and 16 hidden units were used for the rest of the experiments. There were two output units, one for each gender. This final system incorrectly predicts the gender for cross-validation set frames less than 16% of the time.

On a sentence level, it is possible to identify the talker's gender with much lower error. Table 3 shows the classification rate over the cross-validation set using either the mean or median of MLP output probabilities over the sentence. We also tested the sentence mean on the test set of the *DARPA* resource management data base, comprising of 300 sentences from male and female speakers. The performance remained high with a 6.3% sentence error. In the course of dynamic programming, an estimate of a geometric mean is essentially used, which also should give a similar strong result by the end of the sentence. While this result could perhaps be improved, we were satisfied that at least we had a fairly reliable gender indicator.

3 Using Gender-Dependent Features in Speaker-Independent Recognizers

3.1 Introduction

Traditional HMM's have tried to deal with speaker-dependencies through speaker adaptation or speaker clustering. Speaker clustering assumes that speakers can be divided into clusters with similar speech parameters. HMM's are then trained for each cluster. The recognition process is divided

Combination Method	*Sentence Error*
Average	5.8%
Median	7.1%

Table 3: Results for Male-Female Classification with Sentence Smoothing

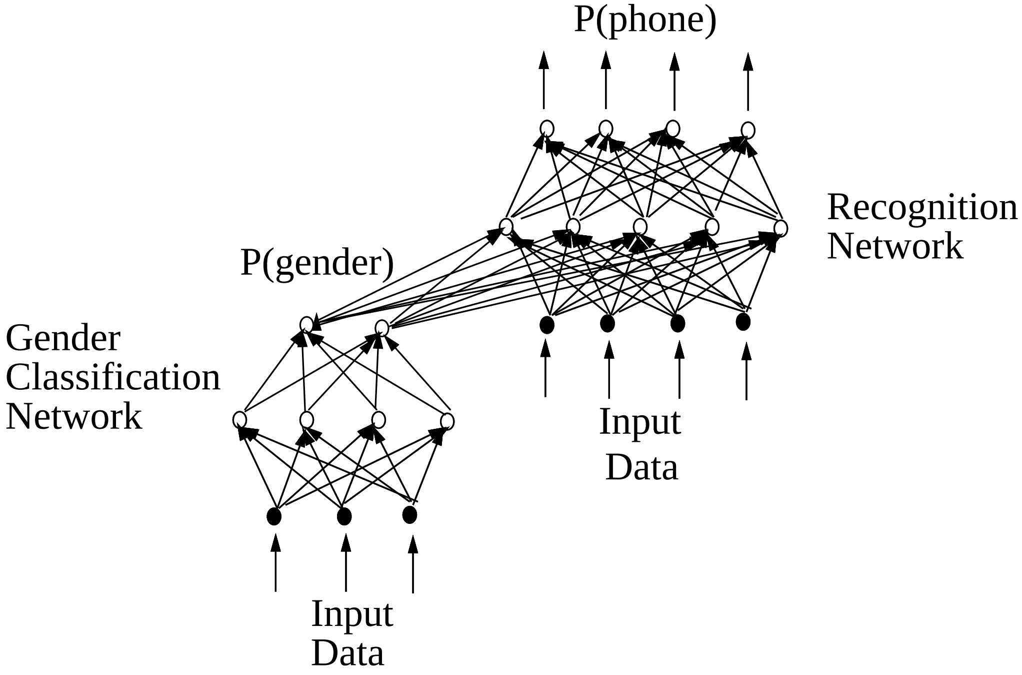

Figure 1: **The outputs of the gender net become additional features to the hidden layer of the general recognition net**

into two stages. The first stage is the *cluster identification* stage, where the speaker is classified into the cluster that most resembles his speech characteristics. In the second stage, the HMM corresponding to the cluster is used to identify the speaker's sentence. For the *cluster identification* stage, the speaker has to provide a training sentence with its correct identity, in order to find the correct cluster. K.F. Lee [4] experimented with two algorithms for speaker adaptation. However, his results are not better than speaker-independent recognition. In a more recent work Huang et al [3] presented a codeword-dependent network(CDNN) for speaker normalization. The network is used as a nonlinear mapping function to transform speech data between two speakers. The goal is to minimize speaker variations with a limited amount of training sentences. Their results using this method are comparable to that of the best speaker-independent performance on the same test set.

In the hybrid MLP-HMM approach, speaker-dependencies can be dealt with by taking into account speaker-dependent speech parameters in the training of the MLP network.

3.2 The Integration of the Classification in the Speech Recognizer

We then attempted to incorporate the gender classification net into the phonetic probability estimation process. In the first approach, the gender subnet is used to generate an extra pair of inputs to the main phonetic probability network. The latter network (which uses 9 frames of 26 mel cepstral and delta mel cepstral coefficients) is then trained and used for HMM likelihood estimation, that is to generate (after division by priors) output probabilities for hypothesized HMMs in the recognition process (see [1] for a more detailed description of this method) . This straightforward approach would generate the same probabilities as in the gender-independent system, that is $P(phone|data)$. However, the hope would be that the separate training of a small application specific net would allow the overall net to learn a better minimum. The architecture of this approach is shown in Figure 1.

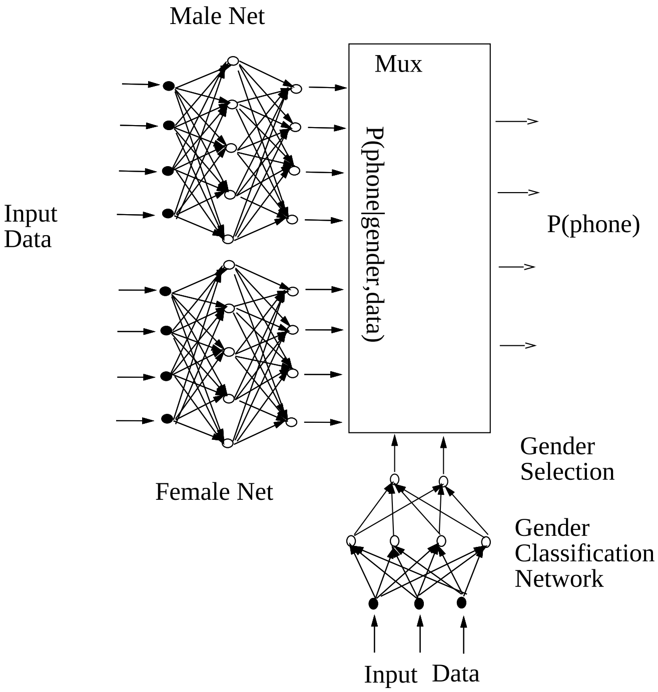

Figure 2: **The output of the classification net selects which net to use, or else provides a weighting for each net's outputs**

In a second approach, the main network is replaced by two networks, one to be trained only with sentences from males, and one only with females. To keep the number of parameters similar, the input context for each is reduced to 5 frames. As shown in Figure 2, the gender network makes a hard decision and only the "correct" net is used. For simplicity's sake, we ignore real-time considerations here and use the sentence-average estimate of male/female probability to determine the setting of the "switch". This approach is similar to K.F. Lee's approach as mentioned above.

In a subtler form of this approach the output of the gender net is used as a probability to weight the phonetic probabilities out of the two other networks, so as to estimate the total phonetic probabilities given the male/female decomposition. The equation for determining the probability of each phone is then:

$$P(phone|data) = P(phone|male, data) * P(male|data) + P(phone|female, data) * P(female|data) \tag{1}$$

All of these approaches in one way or another simply decompose the desired phonetic probability into gender-dependent pieces. Alternatively, one could estimate a probability that was explicitly conditioned on gender, as if the phonetic models were simply doubled to permit male and female forms of each phoneme. This would lead to yet another formulation in which the outputs of the gender-dependent and gender-independent networks were multiplied, but not added together - rather the male and female cases would each have an optimal path computed by the Viterbi process, and the overall probabilities could be compared at the end of the utterance.

We can express $P(male, phone|data)$ (which is then divided by priors to get the corresponding data likelihood) by expansion to $P(phone|male, data) \times P(male|data)$. These factors are just the outputs of the male-trained phonetic net and the male output of the gender net. The final likelihood

Method	*Word error*
Baseline	17.2%
Gender-net as input	17.2%
Switched between gender-dependent nets	18.2%
Weighted Gender-dependent nets	17.0%
Gender consistency over utterance	15.1%

Table 4: Results of Speech Recognition Task

for the male case can be expressed as:

$$P(data|phone, male) = \frac{P(phone|male, data) \times P(male|data) \times P(data)^{1}}{P(phone|male) \times P(male)} \tag{2}$$

Similarly, a female-assumed probability can be computed for each hypothesized phone. These male and female-assumed probabilities can then be used in separate Viterbi calculations (since we do not permit any hypothesis to switch gender in the midst of an utterance). In other words, dynamic programming is used with the framewise network outputs to evaluate the best hypothesized utterance assuming male gender, and then the same is done for the female case. The case with the lowest cost (highest probability) is then chosen.

4 Experimental Results

Each of the approaches described above was tested using a simple HMM recognizer that used a single density per phone and a single pronunciation per word (our best results have used a more complex SRI system that uses multiple pronunciations). The baseline speaker-independent net, as seen in the first row of Table 4, gave a word error rate (using a perplexity 60 wordpair grammar, and including insertions, deletions, and substitutions) of 17.2% on a 300 sentence development set. The table shows that all of the male/female experiments except the last had negative results; simply incorporating male/female information to better estimate the same probabilities gave no significant improvement. In the case of the hard decision switch, there was actually a significant degradation.

On the other hand, the probabilistic approach which used the MLPs to estimate the likelihood of gender-specific phonetic models appears to have been quite successful(significant at the 0.05 level), having eliminated about 1/8 of the errors.

5 Conclusions and Future Work

In order to get better results in the speaker independent task, the speech recognition system has to adapt its speaker-dependent parameters according to the speaker. Furthermore, the knowledge about the speaker has to be acquired incrementally, and integrated in the recognition process.

Our work takes a first step towards integrating speaker-dependent features in a speaker-independent speech recognition system. From our experiments, we draw the following conclusions:

- The features that identify the speaker can be different from the features that are used to recognize the spoken utterance. In particular, the fundamental frequency, which is typically of little value in English phonetic discrimination, is an important feature for gender classification. Similarly, a larger analysis window of 0.5 sec. was useful in modeling consistency.

[1] During recognition, P(data) can be ignored

- Simply adding modular nets to estimate the same ultimate probability did not improve performance. Reformulating the problem to estimate probabilities that explicitly include gender led to significant improvements.

The final successful approach, however, required a separate network for each consistency category. Therefore it does not seem scalable to a large number of speaker characteristics. However, for major and clearly distinguishable classes with a significant amount of training data, it appears to be a viable technique. Other approaches will need to be investigated to incorporate subtle shadings such as regional accent and pace.

6 Acknowledgements

Thanks to Jerry Feldman and Steve Renals for their comments along the way. Chuck Wooters provided assistance in much of the data handling. Computations were done on the RAP machine, with support from software guru Phil Kohn. Michael Cohen of SRI gave useful feedback during the formulation of the successful male/female approach. This work was partially funded by DARPA contract MDA904-90-C-5253.

References

[1] Bourlard, H., Morgan, .N., " Merging Multilayer Perceptrons and Hidden Markov Models: Some Experiments in Continuous Speech Recognition", *Neural Networks: Advances and Applications*, Elsevier Science Publishers B.V., North Holland, 1991.

[2] Bourlard, H., Wellekens, C. J., "Links Between Markov Models and Multilayer Perceptrons" , *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol 12, No. 12, December 1990.

[3] Huang, X. D., Lee, K. F., and Waibel, A., "Connectionist Speaker Normalization and Its Applications To Speech Recognition ", *Neural Networks for Signal Processing*, Proc. of the 1991 IEEE Workshop, Princeton, New Jersey, October 1991.

[4] Lee, K. F., *Automatic Speech Recognition: The Development of SPHINX System*, Kluwer Academic Publishers, 1989.

[5] Murveit, H., Weintraub, M., and Cohen, M., *Training Set Issues in SRI's DECIPHER Speech Recognition System*, Proc. Speech and Natural Language Workshop, June 1990, pp.337-340

[6] Richard, M. D., Lippmann R. P., " Neural Network Classifiers Estimate Bayesian a posteriori Probalilities" , *Neural Computation*, 3, 461-483, 1991.