Continuous Speech Recognition on the Resource Management Database Using Connectionist Probability Estimation

N. Morgan[1], C. Wooters[1],
H. Bourlard[1,2], M. Cohen[3]

TR-90-044

September 7, 1990

Abstract

Previous work has shown the ability of Multilayer Perceptrons (MLPs) to estimate emission probabilities for a Hidden Markov Model (HMM). The advantage to this approach is the ability to incorporate multiple sources of evidence (features, temporal context) without restrictive assumptions of distribution or statistical independence.

In our earlier publications on this topic, a hybrid MLP/HMM continuous speech recognition algorithm was tested on the SPICOS German-language data base. In our recent work, we have shifted to the speaker-dependent portion of DARPA's English language Resource Management (RM) data base. Both consist of continuous utterances (sentences) and incorporate a lexicon of roughly 1000 words. Preliminary results appear to support the previously reported utility of MLP probability estimation for continuous speech recognition (at least for the case of this simple form of HMM).

[1]International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704, USA

[2]Philips Research Laboratory Brussels, Av. Van Becelaere 2, Box 8, B-1170 Brussels, Belgium

[3]SRI International, Menlo Park, CA, USA

INTRODUCTION

We have been working on continuous speech recognition using moderately large vocabularies (1000 words)[1][2][3][4]. While some of our research has been in speaker-independent recognition [5], we have primarily used a German speaker-dependent database called SPICOS [6]. Multilayer Perceptrons (MLPs) trained with back-propagation-styled learning schemes have previously been shown to be useful for the recognition of voiced-unvoiced speech segments [7], isolated phonemes [8], [9], [10], or of isolated words [11]. These results indicate that "neural network" approaches can, for some problems, perform pattern classification as well or better than traditional approaches. Our earlier frame classification results [13], which showed static classification performance for 10 millisecond frames of speech, are consistent with these conclusions. However, these results are not particularly mysterious. When traditional statistical assumptions (distribution, independence of multiple features, etc.) are not valid, systems which do not rely on these assumptions can work better (as discussed in [12]). Furthermore, networks provide an easy way to incorporate multiple sources of evidence (multiple features, contextual windows, etc.) without restrictive assumptions.

Recognition of words in continuous speech requires a system which does more than static pattern classification. In previously reported work, we developed a hybrid MLP-HMM algorithm for this problem, in which an MLP is trained to generate the output probabilities of an HMM [3]. Given speaker-dependent training, we have been able to recognize 50-60% of the words in the SPICOS test sentences. While this is not a state-of-the-art level of performance, it was accomplished with single-state phoneme models, no triphone or allophone representations, no function word modeling, etc., and so may be regarded as a "baseline" system. The main point to using such a system is simplicity for comparison of the effectiveness of alternate probability estimation techniques; our system has very few "knobs" to turn. While we are working on extending our technique to a more complex system, the current paper describes the application of the baseline system (with a few changes, such as different VQ features) to the speaker-dependent portion of the English language Resource Management database [19]. This exercise is primarily intended to confirm that the previous result, which showed the utility of MLPs for the estimation of HMM output probabilities, was not restricted to the limited data set of our first experiments, and that it works for English much as it does for German.

METHODS

As shown by both theoretical [2] and experimental [13] results, MLP output values may be considered to be estimates of Maximum A Posteriori (MAP) probabilities for pattern classification. Either these, or some other related quantity (such as the output normalized by the prior probability of the corresponding class) may be used in a Viterbi search to determine the best time-warped succession of states (speech sounds) to explain the observed speech measurements. This hybrid approach (MLP to estimate probabilities, HMM to incorporate them to recognize continuous speech as a succession of words) has the potential of exploiting the interpolating capabilities of MLPs while using a Dynamic Time Warping (DTW) procedure to capture the dynamics of speech. As described in [3], the practical application of the technique for continuous speech recognition requires cross-validation during training to determine the stopping point, division by priors at the output to generate likelihoods, optimized word transition penalties, and training sentence alignment via iterations of the Viterbi algorithm.

For the Resource Management data, initial development was done on a single speaker to confirm that the techniques we developed for the German data base were still applicable. Although we experimented slightly with this data, the system we ended up with was substantially unchanged, with the exception of the program modifications required to use different VQ features (described below). Final reported scores are given for the 11 speakers which were left out in the development. For each speaker, we used 400 sentences for training, 100 for cross-validation, and a final 100 for recognition. A transcription for each sentence was derived from the most likely pronunciations observed in a large speaker-independent database. For each speaker, we initialized a Viterbi algorithm by assuming a segmentation obtained by assigning a length to each phoneme in the phonetic transcription which came from a table of average phoneme length (normalized to the length of the actual sentence). Relative frequency (i.e., counting) was then used to estimate the emission probabilities. The Viterbi was then used iteratively to generate a final labeling. The final labels were used to train an MLP on the 400 sentences for that person. Input features used were based on the front end for SRI's DECIPHER system [14], including vector quantized mel-cepstrum (12 coefficients), vector quantized difference of mel-cepstrum, quantized energy, and quantized difference of energy. Both

vector quantization codebooks contain 256 prototypes. Energy and delta energy are each quantized into 25 levels. A feature vector is calculated for each 10 ms of input speech.

Each feature was represented by a simple binary input layer with only one bit 'on'. Some experiments were run with no context (i.e., only one frame was input to the network for each classification). Other experiments were run with nine frames of input to the network, allowing four frames of contextual information on each side of the frame being classified. As we found in our SPICOS experiments, a hidden layer was not useful for this problem. The size of the output layer was kept fixed at 61 units, corresponding to the 61 phonemes to be recognized. The input field contained 9×562=5058 units, and the total number of possible inputs was equal to 3×10^{68}. There were typically about 130000 training patterns (from the 400 training sentences). Of course, this represented only a very small fraction of the possible inputs, (or even of the inputs which are plausible for real speech), and generalization was thus potentially difficult. Training was done by an error-back propagation algorithm [15][16], using an entropy criterion [17][18]. In each iteration, the complete training set was presented, and the parameters were updated after each training pattern. To avoid overtraining of the MLP, improvement on the cross-validation set was checked after each iteration. If the classification rate on the cross-validation set had not improved more than a small threshold, the adaptation parameter of the gradient procedure was decreased; otherwise it was kept constant. Training ended when improvement on the cross-validation set went below a second threshold. Performance was insensitive to the exact values of these thresholds. After some experiments with our development speaker (dtd05), we settled on an initial adaptation constant of .05 with no momentum term. The threshold for changing the learning constant was initially set at .5% improvement on the cross-validation set, but was then reset to an infinite value (i.e., change the learning constant after every iteration) after the reduction from the first learning constant. This heuristic appeared to cut learning time roughly in half without adversely affecting performance. The learning constant was reduced by a factor of two for each change. The final stopping parameter was set at .5% improvement on the cross-validation set.

The output layer of the MLP was evaluated for each frame, and (after division by the prior probability of each phoneme) was used as the emission probability in a discrete HMM system. In this system, each phoneme was modeled with a single conditional density, repeated $D/2$ times, where D was a prior estimate of the duration of the phoneme. Only selfloops and sequential transitions were permitted. A Viterbi decoding was then used for recognition of the first thirty sentences of the cross-validation set (on which word transition penalties were optimized). The trained system was then tested on the final 100 sentences for each speaker. Note that this same simplified HMM was used for both the Maximum Likelihood (ML) reference system (estimating probabilities directly from relative frequencies) and the MLP system, and that the same input features were used for both.

RECOGNITION RESULTS

Table I shows the frame classification performance for the MLP and for single-frame Maximum a Posteriori (MAP) estimates, where the latter were calculated from counting relative frequencies of the features for each phoneme. For the explicit MAP estimate, it is not possible to use contextual information, because of the high dimensionality of the input space. Further, we must assume the independence of the four feature likelihoods so that we can estimate the joint density by their product. These restrictions explain why the MAP (also called Bayes) classifier, which is inherently optimal for a given pattern classification problem, is shown here as yielding a lower performance than the potentially suboptimal MLP. The difference between the frame classification performance for the MAP classifier and the single-frame MLP's (second column) is statistically significant ($p < .001$ using a normal approximation to a binary distribution for the null hypothesis), both for all individual speakers and for the multispeaker comparison. Thus, for these VQ features, the MLP is somewhat better at estimating the joint density than the simple multiplicative method. Statistics aside, the improvement is slight. However, the large improvement shown for the 9 frames of context (last column) indicates that the MLP has found temporal correlations which generalize to new data.

Table II shows the recognition rate (100% – error rate, where errors include insertions, deletions, and substitutions) for the 100 test sentences. As observed for the frame-level results of Table I, the incorporation of context (last column) was the major effect. However, the improvement using the single-frame window was statistically significant for the multi-speaker comparison ($p<.001$), and was also significant ($p<.01$) for 7 of the individuals. This technique was statistically equivalent to using ML probability estimates (from counting) for another 3 of the speakers, The single-frame MLP gave significantly worse

results for only one of the 11 speakers (das12), and inspection of this case showed that the word-transition optimization had done a poor job, resulting in an excessive number of insertions for this speaker. Finally, the incorporation of 9 frames of context (the last column of Table II) provided significant improvement ($p < .001$) for every individual case, as well as for the pooled data. Thus, the MLP-based methods consistently show measurable improvement over the simpler estimation technique.

DISCUSSION

These results (all obtained with no language model, i.e., with a perplexity of 1000 for a 1000 word vocabulary) show some of the improvement using MLPs for continuous recognition (over simpler probability estimators) which one might expect from the frame level results (Table I). As described above, each table shows results for 11 speakers (plus the multi-speaker mean) and for cases of 1 and 9 frames of input context for the network. MLPs can sometimes make better frame level discriminations than simple statistical classifiers, because they can easily incorporate multiple sources of evidence (multiple frames, multiple features), which is otherwise difficult to do in HMMs without major simplifying assumptions. In general, the relation between the MLP and ML word recognition is more complex, because of interdependence over time of the input features. The incorporation of multiple frames of context for the MLP might seem to be an "unfair" basis of comparison to the ML case; however, no claim is made for the MLP as the unique method of incorporating temporal context in probabilistic estimation for HMM's, only that it appears to be a good method. An alternate approach is to use dynamic features, which in fact were already in use in our example (the SRI features include delta features). Other experiments not reported here [5] have shown that higher order dynamic features (e.g., acceleration) may also be used to good advantage by the MLP. Our current belief is that the further use of such features can improve the performance of the MLP/HMM system for single-frame inputs. Nonetheless, it is clear that the MLP can make good use of temporal context to generate probability estimates.

The features we have been using were chosen for their effectiveness in HMM systems, and different combinations may prove to be better for MLP inputs. In particular, we would expect that feature combinations that have not been vector-quantized should have more useful dependencies (both within-frame and over time) that the MLP may be able to learn and exploit. We are exploring this possibility currently. However, despite the potential performance advantages for continuous features, VQ features are still of interest because of significant advantages for hardware implementation (e.g., no multiplications are required).

Results reported here are somewhat better than those we reported for the SPICOS study. Since we are using essentially the same system, we attribute the difference to the improved features (four features rather than 1, including delta features) and to the fact that the SPICOS test set included many words that never occur in the training set.

SUMMARY

Previous results [3], using a German speaker-dependent database, showed that an MLP could make effective use of temporal context to generate effective estimates of likelihoods for use in an HMM-based continuous speech recognition procedure. A new experiment has been performed that shows that this result holds for an English language data base, the speaker-dependent part of Resource Management [19]. In particular, for each of the 11 speakers that were not used for development, the contextual MLP reduced word recognition error significantly (from 48% to 39% on the average). This was observed using a simplified HMM system with single-state monophone models, no skips or insertions, no language model, and no function word models. Now that we have confirmed the principle, we are beginning to develop a complete system, which will incorporate context-dependent sound units. We are also working on further incorporation of speech knowledge in the design of the MLP.

ACKNOWLEDGMENTS

Thanks to Phil Kohn of ICSI for his many hours of work providing software support on our new multiprocessor machine (the RAP), which crunched many teraflops on this study. Support from the International Computer Science Institute (ICSI), Philips Research, and SRI International for this work is also gratefully acknowledged.

speaker name	MAP	MLP(1)	MLP(9)
jws04	54.6	56.9	69.0
bef03	51.4	52.9	65.0
cmr02	51.4	52.9	67.2
dtb03	51.3	52.4	64.6
das12	53.5	55.9	72.2
ers07	50.9	52.0	64.4
dms04	55.8	57.0	72.1
tab07	54.8	56.3	70.3
hxs06	52.8	54.8	69.9
rkm05	45.0	45.8	56.8
pgh01	46.2	52.1	64.9
mean	51.6	53.5	67.0

Table I — Frame classification in % correct for MAP estimate from counting, MLP with 1 frame of speech input, and MLP with 9 frames of context as input.

speaker name	ML	MLP(1)	MLP(9)
jws04	48.2	53.6	56.0
bef03	39.3	46.6	52.3
cmr02	59.5	64.4	66.6
dtb03	49.8	52.0	58.1
das12	63.8	55.8	71.7
ers07	45.4	49.2	53.8
dms04	58.0	58.8	67.1
tab07	60.8	59.6	66.6
hxs06	60.9	65.5	71.3
rkm05	37.9	46.4	48.7
pgh01	50.4	57.3	60.4
mean	52.2	55.4	61.1

Table II — Word recognition, in % correct for Maximum Likelihood estimate from counting, MLP with 1 frame of speech input, and MLP with 9 frames of context as input.

REFERENCES

[1] H. Bourlard, N. Morgan, and C.J. Wellekens. "Statistical Inference in Multilayer Perceptrons and Hidden Markov Models with Applications in Continuous Speech Recognition", to appear in *Neuro Computing, Algorithms, and Applications*, NATO ASI Series, 1989

[2] H. Bourlard, and C.J. Wellekens. "Speech Pattern Discrimination and Multilayer Perceptrons", in *Computer, Speech, and Language*, vol. 3, pp.1-19, 1989

[3] N. Morgan, and H. Bourlard. "Continuous Speech Recognition Using Multilayer Perceptrons with Hidden Markov Models", *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, pp. 413-416, Albuquerque, New Mexico, 1990.

[4] H. Bourlard, and N. Morgan. "Merging Multilayer Perceptrons & Hidden Markov Models: Some Experiments in Continuous Speech Recognition" in *Artificial Neural Networks: Advances and Applications*, North Holland Press, 1990, E. Gelenbe editor (In Press)

[5] N. Morgan, H. Hermansky, C. Wooters, P. Kohn, and H. Bourlard. "Phonetically-based Speaker Independent Continuous Speech Recognition using PLP Analysis with Multilayer Perceptrons" *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, Submitted for 1991

[6] H. Ney, and A. Noll.. "Phoneme Modeling Using Continuous Mixture Densities", *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, Vol. 1, pp. 437-440, New York, 1988.

[7] A. Gevins, and N. Morgan. "Ignorance-Based Systems", *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, Vol. 3, 39A5.1-39A5.4, San Diego, 1984

[8] R.L. Watrous, & L. Shastri. "Learning phonetic features using connectionist networks: an experiment in speech recognition", *Proceedings of the First Intl. Conference on Neural Networks*, IV-381-388, San Diego, CA, 1987

[9] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. "Phoneme Recognition: Neural Networks vs. Hidden Markov Models", *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, Vol. 1, pp. 107-110, New York, 1988.

[10] S. Makino, T. Kawabata, and K. Kido. "Recognition of consonants based on the Perceptron Model", *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, Vol. 2, pp. 738-741, Boston, Mass., 1983

[11] S.M. Peeling, R.K. Moore. "Experiments in Isolated Digit Recognition Using the Multi-Layer Perceptron", Royal Speech and Radar Establishment, Technical Report 4073, Malvern, Worcester, 1988

[12] L. Niles, H. Silverman, G. Tajchman, and M. Bush. "How Limited Training Data Can Allow a Neural Network Classifier to Outperform an 'Optimal' Statistical Classifier", *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, Vol. 1, pp. 17-20, Glasgow, Scotland, 1989

[13] H. Bourlard, and N. Morgan. "Merging Multilayer Perceptrons and Hidden Markov Models: Some Experiments in Continuous Speech Recognition" International Computer Science Institute TR-89-033, 1989

[14] H. Murveit, and M. Weintraub.. "1000-Word Speaker-Independent Continuous-Speech Recognition Using Hidden Markov Models" *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, Vol. 1, pp. 115-118, New York, 1988.

[15] P.J. Werbos, 1974. "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences", Ph.D. thesis, Dept. of Applied Mathematics, Harvard University, 1974

[16] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. "Learning Internal Representations by Error Propagation," in Parallel Distributed Processing. vol. 1: Foundations, Ed. D.E.Rumelhart and J.L.McClelland, MIT Press, 1986.

[17] S.A. Solla, E. Levin, & M. Fleisher. "Accelerated Learning in Layered Neural Networks", AT&T Bell Labs Manuscript, 1988

[18] G. Hinton . "Connectionist Learning Procedures", Technical Report CMU-CS-87-115, Carnegie Mellon, 1987

[19] P. Price, W. Fisher, J. Bernstein, and D. Pallet. "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition", *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, Vol. 1, pp. 651-654, New York, 1988.