SPEECH WORD RECOGNITION WITH BACKPROPAGATION AND FUZZY-ARTMAP NEURAL NETWORKS

Lipo Wang
School of Computing and Mathematics
Deakin University
662 Blackburn Road
Clayton, Victoria 3168 Australia
Email: lwang@deakin.edu.au

ABSTRACT

We present the initial results on speaker-independent speech word recognition using a modified version of the QuickProp network, a back-error-propagation variant, and the Fuzzy-ARTMAP network, an ART variant. The TIMIT speech database is used for training and testing. While both networks achieve results better than existing results on speech phonetic recognition, the Fuzzy-ARTMAP network runs much faster, but yields generally inferior results, in comparison to the modified QuickProp network.

INTRODUCTION

Neural networks have shown some promise in speech processing lately [e.g., BDFK92, DG91, MZ91, WC91, DRO91, BF90] and have yielded comparable or better results in comparison with the hidden Markov model (HMM) approaches [e.g., LH89, PC91]. These work are done using phonemes in the TIMIT speech database [ZSG90]. Speaker-independent phonetic recognition accuracy ranging from 40%-80% has been reported.

In this brief paper, we present our initial results on speaker-independent word recognition, also using the TIMIT speech database. We shall use two types of neural networks, i.e., a modified version of the QuickProp network [Fah88], a backpropagation [RM86] variant, and the Fuzzy-ARTMAP network [CG92], an ART [CG86] variant. We deal with speech words, instead of phonemes. Specifically, we use speech data for the 11 words in the following sentence: "She had your dark suit in greasy wash water all year."

SPECIFICATIONS AND RESULTS

The training sets consist of speech data recorded from a total of 400 speakers from 8 main dialect regions in the United States [ZSG90]. The testing sets consist of speech data from 230 speakers not included in the training set.

Since the number of data in a digitised speech recording is usually very large, e.g., it includes about 60,000 data points per sentence, preprocessing of the raw data is necessary before they are suitable for neural networks. We did the following preprocessing with all training and testing data:

(1) The speech signal is high-frequency pre-emphasized with transfer function (1 - 0.95/z);
(2) The speech signal is then windowed using a 32 msec Hamming window, with a 10 ms frame spacing;
(3) The magnitude spectrum is computed using a 512 point Fast Fourier Transform for each frame;
(4) The spectrum is then log amplitude scaled, frequency warped with a bilinear transform with a coefficient of 0.6;
(5) Fifteen sepstral coefficients are then computed over a frequency range of 150 Hz to 6000 Hz.

We have modified the quick-prop network [Fah88] to enable it to output "unclassified" or "don't know" answers. Specifically, we have added the following rules during testing after training. When a testing pattern is presented to the network after training, we calculate a sum of squared error for each output node. If the sum is less than a preset error threshold, the test pattern is considered to be "classified", and it can be either "correctly classified" or "incorrectly classified". Otherwise the test pattern is

"unclassified". We note that this classification error threshold is not necessarily equal to the error threshold for training to stop.

After training the two networks with the preprocessing training data, we test the networks with the preprocessed testing data. Both the recognition accuracy and the computer time are recorded. Our work is carried out in a Convex C240 Super Computer. The results are presented in Table 1.

Note that the "time" used above refers to the "user time", rather than the "real time" or the "system time".

The Fuzzy-ARTMAP results in the table are obtained using a 2-voter scheme [CG92]; however, we note that similar results are obtained with a 1-voter scheme.

In the present work, we use a three-layer feed-forward backpropagation network. We chose the number of output nodes to be the same as the number of classes, i.e., 11. We chose the number of input nodes to be the same as the largest number of input data point for a speech word after preprocessing, which depends on dialect region and the sex of the speakers. We use 400 hidden units for all cases, which yields good results compared with other choices for the number of hidden units. We chose the classification threshold ERROR=0.2 and the error threshold for training to stop TOTALERRORT=10.0. We let the maximum number of epox be 400.

CONCLUSIONS

We present the initial results on speaker-independent speech word recognition using a modified version of the QuickProp network, a backpropagation variant, and the Fuzzy-ARTMAP network, an ART variant. The TIMIT speech database is used for training and testing. While both networks achieve results better than existing results on speech phonetic recognition, the Fuzzy-ARTMAP network runs much faster, but yields generally inferior results, in comparison to the modified QuickProp network.

REFERENCES

[BDFK92] Y. Bengio, R. De Mori, G. Flammia, and R. Kompe. Global optimiztion of a neural network-hidden Markov model hybrid. IEEE Trans. on Neural Networks, 3: 252-259, 1992.

[BF90] D. Bijl and F. Fallside. A speech application of a probabilistically conditioned neural network. in C.P. Tsang (ed.) Proc. 4th Australian Joint Conf. Artificial Intelligence: 265-273, 1990.

[CG86] G. A. Carpenter and S. Grossberg. Neural dynamics of category learning and recognition: Attention, memory consolidation, and amnesia. in Brain Structure, Learning, and Memory, J. Davis, R. Newburgh, and E. Wegman (eds.), AAAS Symposium Series, 1986.

[CG92] G.A. Carpenter, S. Grossberg, N. Markuzon, J.H.

Table 1 Performance Comparison on Speech Recognition

Dialect Region & Sex of Speaker	Fuzzy-ARTMAP Classification Accuracy Overall Correct	Fuzzy-ARTMAP Computer Time in Seconds	Modified QuickProp Classification Accuracy Correct in Classified (Overall, Unclassified)	Modified QuickProp Computer Time in Seconds
1 f	81.8%	56.0	90.9% (68.2%, 25.0%)	16343.6
1 m	83.1%	104.3	96.9% (81.8%, 15.6%)	45386.5
2 f	94.3%	77.9	98.6% (80.7%, 18.2%)	30077.9
2 m	87.9%	201.2	96.6% (86.9%, 10.1%)	142878.5
3 f	100.0%	56.5	100.0% (78.8%, 21.2%)	19190
3 m	96.8%	205.9	98.3% (90.9%, 7.5%)	128075.5
4 f	92.0%	73.5	97.1% (75.6%, 22.2%)	12322.9
4 m	94.3%	239.5	98.8% (89.8%, 9.1%)	171333.0
5 f	90.9%	106.1	100.0% (82.6%,17.4%)	38334.3
5 m	90.9%	221.7	98.8%(86.1%, 12.8%)	118969.3
6 f	97.0%	43.1	92.3% (72.7%, 21.2%)	11654.1
6 m	85.2%	93.3	94.7% (81.8%, 13.6%)	33369.6
7 f	94.3%	59.5	98.6% (83.0%, 15.9%)	17742.8
7 m	86.1%	177.3	95.9% (85.5%, 10.9%)	134079.0
8 f	84.8%	26.1	88.9% (72.7%, 18.2%)	4687.4
8 m	84.1%	56.5	97.1% (76.1%, 21.6%)	17037.4

Reynolds, and D.B. Rosen. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Trans. Neural Networks, 3: 698-713, 1992.

[DG91] M.R. Davenport and H. Garudadri, A neural net acoustic phonetic feature extractor based on wavelets. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (May 9-10, 1991), 449-452 (1991)

[DRO91] V. Digalakis, J.r. Rohlicek, and M. Ostendorf. A dynamical system approach to continuous speech recognition, IEEE Conf. Proc. CH2977-7/91/0000-0289: 289-291, 1991.

[Fah88] S. Fahlman. Faster-Learning Variations on Back-Propagation: An Empirical Study. in Proceedings of 1988 Connectionist Models Summer School, Morgan Kaufmann, 1988.

[LH89] K.-F. Lee and H.-W. Hon. Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoustic, Speech, and signal Processing, 37: 1641-1648, 1989.

[MZ91] H.M. Meng and V.W. Zue. Signal representation comparison for phonetic classification, IEEE Conf. Proc. CH2977-7/91/0000-0285: 285-288, 1991.

[PC91] D.J. Pepper and M.A. Clements. On the phonetic structure of a large hidden Markov model. IEEE Conf. Proc. CH2977-7/91/0000-0465: 465-468, 1991.

[RM86] D.E. Rumelhart and J.L. McClelland. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, I and II, MIT Press, Cambridge MA, 1986

[WC91] J.-X. Wu and C. Chan. Recognition of phonetic labels of the TIMIT speech corpus by means of an atificail neurla netowrk. Pattern Recognition, 24: 1085-1091, 1991.

[ZSG90] V. Zue, S. Seneff, and J. Glass. Speech database development at MIT: TIMIT and beyond. Speech Comm., 9: 351-356, 1990.