Understanding Speech Recognition Using Correlation-Generated Neural Network Targets

Yonghong Yan

***Abstract*— Training neural networks (NN's) with variable targets for speech recognition systems has been shown to be effective in improving word accuracy. In this correspondence, a new and simple method for estimating variable targets for a given training pattern is presented. It uses estimated correlations between different output nodes of an NN to create a set of variable targets for each training pattern. Experimental results show that the word error is reduced by more than 20% when these new correlation-based targets are compared to more conventional zero/one targets with a squared-error cost function. Performance with these new targets approaches that of high-performance hidden Markov model (HMM) recognizers but requires far fewer parameters.**

***Index Terms*—Neural networks, speech recognition systems.**

I. Introduction

It has been argued that if a neural network (NN) is trained to minimize a least square error or a cross entropy distortion measure, the output of the NN will approximate the underlying posterior probability [2], [7], [13]. This is usually how NN's are used in speech recognition systems. To train an NN, a target must be given for each training pattern (the speech vector). In general (as in our baseline approach), for each feature vector, only a single class is assigned a variable target. This binary-target approach can be a problem: Since phonetic models (output nodes) in a NN system are trained simultaneously so that examples of each category are also counter examples of other categories, it is desirable that acoustically similar pairs (e.g., the end of /iy/ and the end of /ei/) should somehow have training targets reflecting this similarity. The target-assignment issue has been raised by other researchers. Hampshire and Waibel [8] proposed the classification figure of merit (CFM), which attempts to minimize the number of classification errors in a way similar to the perceptron convergence procedure. Caruana *et al.* [4] reported significant improvement on medical risk evaluation by using variable targets, and Senior *et al.* [14] had a similar conclusion for a handwriting recognition system. Our previous work [17] showed the effectiveness in improving system performance using NN's trained with variable targets generated by the forward–backward algorithm, and the conclusion was in line with previously reported similar approaches (e.g., [9] and [10]). In this work, a novel and simple technique for computing multiple variable training targets for NN-based phoneme probability estimators is presented. It uses correlations between output values of an initial NN to create a set of variable training targets for each output category. In this approach, acoustical similarities among different nodes (which represent different subphone units) are measured by their correlations, which can be interpreted as a kind of soft data sharing [6], since multiple output nodes are active for each training pattern. Unlike the previously reported approaches, our approach uses the statistics (correlations) obtained from the whole training set instead of each individual utterance (as used in [10] and [14]) to generate the target for the given training utterance. The proposed approach was evaluated on a telephone digit-string recognition task and a speaker- and vocabulary-independent isolated-word recognition task. The results showed that a greater than 20% word-error reduction was achieved. This is comparable to the improvement achieved using other variable target approaches (e.g., [10] and [17]) in terms of relative error reductions on different tasks.

Manuscript received July 16, 1997; revised August 7, 1998. This work was supported by the Office of Naval Research, the National Science Foundation, DARPA, and the member companies of CSLU. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard C. Rose.

The author is with the Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, Portland, OR 97291-1000 USA (e-mail: yan@cse.ogi.edu).

Publisher Item Identifier S 1063-6676(99)02731-5.

II. Target Generation Using Correlations

In theory, an NN learns the average effect of the training patterns. The normalized output values (the activations) of a trained NN given a test pattern reflect the relative likelihood of the test pattern belonging to different classification categories. Since the models (output nodes) in a NN system are trained simultaneously, examples of every category are also counter examples of other categories. With multiple output nodes active for each training pattern, acoustic similarity is explicitly represented. Also, given the fact that an NN learns the mapping function from the input patterns to the corresponding targets, a set of variable targets for each training pattern can also be viewed as soft data sharing.

In statistics, the correlation coefficient measures the degree of association between two random variables. In our study, the probabilistic targets are generated based on the correlations of output nodes. Since a trained NN is a mapping function between training patterns and targets (supposedly, the probabilities), the correlations between NN output activations can be viewed as a measurement of their acoustic similarity.

In our implementation, the standard correlation formulation is used. The correlation coefficient between two activation variables X and Y is calculated as

$$\rho_{(XY)} = \frac{\mathrm{Cov}(X, Y)}{D(X)D(Y)} \tag{1}$$

where $\mathrm{Cov}(X, Y)$ is the covariance, and $D(X)$ and $D(Y)$ are the standard deviations. For each training pattern, the correct output class is assigned a target value of 1.0 and other classes are assigned a target value equal to their correlation coefficient with the target class ($\rho_{(XY)}$). For practical reasons (storage space and numerical precision), only a small portion (about 1.5%) of the nodes are assigned variable targets. These are the nodes with the top N highest correlation coefficients (with cut-off threshold 0.2), given the actual linguistic unit (the only active node in the zero/one target NN). Further, in order to avoid loss of discriminative power, the activation for the correct target class is multiplied by a scaling factor α (which is set in the range of 1.2–1.5). Our experiments indicate that α is not sensitive in terms of significantly changing the system performance. This scaling factor is determined empirically. The sum of targets for each training pattern is then normalized to one.

During implementation, a binary NN (for each training pattern, only one output node has a target value set to 1.0, the rest are set to zero) is trained first as an initialization NN. The correlation coefficients are calculated based on the output activations of this NN over the training data. A new NN with variable targets is trained once the target set based on correlations has been generated.

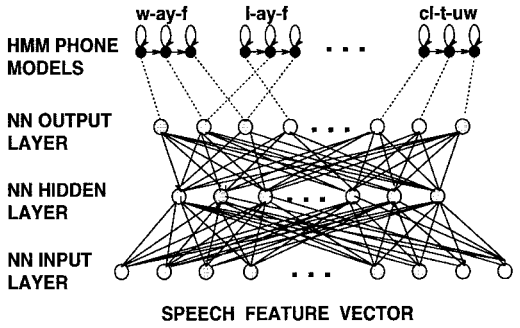

Fig. 1. Relation between NN output nodes and the phone models.

III. The NN-Based Speech Recognition System

The experiment platform for this study was implemented using the CSLU speech toolkit, which is an NN-based approach (multilayer perceptron) resulting from many people's work in the past seven years.[1] The tasks and the NN architectures used in this study are standard for the CSLU toolkit, which can be downloaded from the CSLU website. The NN systems reported in this work all run in real time.

Like most of the existing approaches, our acoustic modeling unit is the phone. Each phone is represented as an n state HMM, where n can be one, two, or three based on their spectral structure. This is a derived from [1]. Each state in a phone model corresponds to a NN output node.

The relation between context-dependent phone models and the output nodes is illustrated in Fig. 1. In the figure, L-PH-R (such as w-ay-f) denotes phone PH in the context of phone L (left) and R (right). As shown, the context-dependent phones from the same monophone share the middle state. The left (or right) state for each model only depends on the left (or right) context. Thus in Fig. 1, both the middle states and the end states of the w-ay-f and l-ay-f models use the same NN output.

In the experiments reported in this paper (unless specified), the speech waveform was parameterized in 16 ms windows with a 6 ms overlap between contiguous frames. For each frame, a 26-dimensional feature vector was calculated: 12 mel scale cepstral coefficients (MFCC's) and normalized energy plus their deltas. Cepstral mean subtraction was employed. A training pattern was a large (130-dimensional) vector of juxtaposed frame features for five contiguous analysis frames [1].

IV. Evaluation

In order to evaluate the new approaches, several experiments were conducted. This section presents the tasks and the results.

A. Databases

Three databases used in this study are described as follows.

1) *PhoneBook Database:* This is a phonetically rich isolated-word telephone-speech database collected by NYNEX Science and Technology, Inc., [12]. It has nearly 8000 distinct words, selected for complete coverage of phoneme contexts. There are more than 92 000 utterances from over 1300 native speakers in the database. Only word transcriptions are available in this database.
2) *OGI Multilanguage Telephone Speech Database (OGI_TS):* This is a telephone speech database with unconstrained vocabulary [11]. This database was originally designed for automatic language identification research. We used the "story-bt" part from (American) English, in which each utterance contains 45 s of unconstrained continuous speech. These utterances are phonetically transcribed.

   This database was used as a supplement to the PhoneBook database. The phonetic transcriptions were used to train the original NN and some of the data were used to provide more balanced contextual coverage.
3) *OGI_30K Numbers Database:* This is a telephone speech database collected by CSLU [5]. It consists of over 30 000 numbers utterances. Any phrase that can be considered a number is placed in the corpus: cardinal numbers, ordinal numbers, and digit strings.

Since callers were recruited through public advertisement, and were instructed to call the data collection phone number at any time or place, the database is close to a real-world application environment. False starts, pauses, repetition, and background noise are common in the database.

[1] For more detail, see http://www.cse.ogi.edu/CSLU.

TABLE I
Word Error Rates for the Isolated Work Recognition Task

Data Set	Baseline	Correlation	HMM
Test Set	16.0%	12.7%	12.0%

B. Vocabulary-Independent Task

This is an isolated-word recognition task intended for speaker- and vocabulary-independent applications. The experimental system was trained mainly using the PhoneBook database (enhanced by the OGI_TS database). The training data contain 3990 distinct words. The active vocabulary used during recognition was the entire set of 7979 words in phonebook.

A baseline system with binary targets was trained first and used as the initialization NN for the correlation-based NN training. The two systems had the same NN architecture with 130 input nodes, 200 hidden nodes, and 534 output nodes.

For the correlation-based NN, the correlation matrix was calculated based on the output activations of the initial NN when used on the training patterns. The system was retrained using the newly generated targets, with nine variable output nodes for each training pattern. For comparison purposes, a continuous HMM recognizer was also implemented. The HMM system was built using a commercially available software package (HTK V2.0 [16]). Each phone was represented as a three-state left-to-right model using diagonal covariances. The speech signal was parameterized every 12.8 ms with a 25.6 ms Hamming window. Twelve mel scale cepstral coefficients plus normalized log-energy, together with their delta and acceleration coefficients, were calculated to form a 39-dimensional frame feature. Several HMM systems were trained using a decision tree [15] for state clustering, with the total number of states ranging from 300 to 3000. The best HMM system had 1583 states with eight Gaussians per state. This HMM system had about 1 M parameters, while the resulting NN system had 133 k parameters. Results are summarized in Table I.

C. Continuous Digit Recognition

The vocabulary for this task is: *zero*, *oh*, and *one* through *nine*. The digit strings contain one to ten continuously pronounced digits; the grammar treats all strings of any length as equally likely (no constraint). The digits part of the OGI_30K numbers database was randomly divided into three sets, with 2090 utterances in the training set, 500 utterances in the development set, and 1600 utterances in the final evaluation set.

TABLE II
WORD AND STRING (SENTENCE) ERROR RATES FOR THE DIGIT STRING TASK

UNIT	BASELINE	CORRELATION
Word	5.7%	3.8%
String	16.0%	13.6%

Two NN systems were built for evaluating the proposed target generating methods. The baseline system used a binary-target NN. This NN was used as the initialization NN.[2]

The second system was trained using correlation-based targets, where three nodes were allowed to be variables for each training pattern. These two systems both have 130 input nodes, 200 hidden nodes, and 209 output nodes. Results are summarized in Table II.

V. CONCLUDING REMARKS

Generating multiple variable output targets based on correlations between output units is a simple yet effective way to improve the recognition accuracy for NN-based speech system. It also provides a "soft" data sharing method. Experiments show this approach, without adding parameters into the final system, achieved more than a 20% word error reduction for the two benchmark tasks. In the vocabulary independent experiment, given the fact that the parameter size of the NN based system is significantly smaller than that of the HMM system, the proposed method effectively closed the performance gap without adding any extra parameter into the system. The experiment confirms the belief that a more reasonable target set simplifies learning (since there are fewer inconsistencies in the training patterns) and increases the generalization ability of the trained NN for a given amount of data.

ACKNOWLEDGMENT

The author wishes to thank Dr. R. A. Cole, Dr. M. Fanty, X. Wu, and J. Schalkwyk for help in setting up the experiments and for many useful discussions, and J. P. Hosom for proofreading this manuscript.

REFERENCES

[1] E. Barnard, R. A. Cole, M. Fanty, and P. Vermeulen, "Real-world speech recognition with neural networks," in *Proc. Int. Symp. Aerospace/Defense Sensing and Control and Dual-Use Photonics, International Society for Optical Engineering, Technical Conf. 2492*, 1995.

[2] A. Barron, "Statistical properties of artificial neural networks," in *Proc. IEEE Conf. Decision and Control*, 1989, pp. 280–285.

[3] D. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Proc. Int. Conf. Spoken Language Processing*, 1996, pp. 1145–1148.

[4] R. Caruana, S. Baluja, and T. Mitchell, "Using the future to sort out the present: Rankprop and multitask learning for medical risk evaluation," in *Advances in Neural Information Processing Systems 8*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996, pp. 959–965.

[5] R. A. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at CSLU," in *Proc. 4th Europ. Conf. Speech Communication and Technology*, Madrid, Spain, Sept. 1995, pp. 821–824.

[6] A. M. Derouault, "Context-dependent phonetic Markov models for large vocabulary speech recognition," in *Proc. ICASSP*, 1987, pp. 360–363.

[7] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1990, pp. 1361–1364.

[2] To shed light on the relative difficulty of the OGI_30K numbers database, CSLU implemented two baseline systems using the same technique for the TIDIGITS and the OGI_30K numbers. The word error rates for the TIDIGITS and OGI_30K numbers were 0.7% [3] and 6.0%, respectively.

[8] J. B. Hampshire and A. H. Waibel, "A novel objective function for improved phoneme recognition using time-delay neural networks," *IEEE Trans. Neural Networks*, vol. 1, pp. 216–228, 1990.

[9] Y. Komori, "A neural fuzzy training approach for continuous speech recognition improvement," in *Proc. 1992 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Mar. 1992, pp. 405–408.

[10] Y. Konig, H. Bourlard, and N. Morgan, "Remap-experiments with speech recognition," in *Proc. 1996 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Atlanta, GA, May 1996, pp. 3350–3353.

[11] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proc. Int. Conf. Spoken Language Processing*, Oct. 1992, pp. 895–897.

[12] J. Pitrelli *et al.*, "PhoneBook: A phonetically-rich isolated-word telephone-speech database," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1995, pp. 101–104.

[13] M. D. Richard and R. P. Lippmann, "Neural network classifiers estimate Bayesian *a posteriori* probabilities," in *Proc. Conf. Neural Computation 3*, 1991, pp. 461–483.

[14] A. Senior and T. Robinson, "Forward-backward retraining of recurrent neural networks," in *Advances Neural Information Processing Systems 8*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996, pp. 743–749.

[15] S. J. Young and P. C. Woodland, "Tree-based state-tying for high accuracy acoustic modeling," in *Proc. Human Language Technology Workshop*, Mar. 1994, pp. 307–312.

[16] S. Young *et al.*, *The HTK Book* Cambridge Res. Lab., Cambridge Univ., U.K.

[17] Y. Yan, M. Fanty, and R. Cole, "Speech recognition using neural networks with forward–backward probability generated targets," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1997, pp. 3241–3244.