COMBINING ANNS TO IMPROVE PHONE RECOGNITION

Brian Mak

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
20000 N.W. Walker Road, Portland, OR 97006, USA.
mak@cse.ogi.edu

ABSTRACT

In applying neural networks to speech recognition, one often finds that slightly different training configurations lead to significantly different networks. Thus different training sessions using different setups will likely end up in "mixed" network configurations representing different solutions in different regions of the data space. This sensitivity to the initial weights assigned, the training parameters and the training data can be used to enhance performance, using a committee of neural networks. In this paper, we study various ways to combine context-dependent (CD) and context-independent (CI) neural network phone estimators to improve phone recognition. As a result, we obtain 6.3% and 2.2% increase in accuracy in phone recognition using monophones and biphones respectively.

1. INTRODUCTION

In the past decade, a number of connectionist approaches have enabled a new computing paradigm for speech recognition with some success [1, 6, 12, 15]. In these ANN-based speech recognizers, the neural network is usually employed as an estimator for the posterior probabilities of phones or other subword units. Powerful as neural networks are, training neural networks that generalize well with unseen data remains an ongoing research topic. Neural network training is sensitive to the initial assignment of weights , the training parameters and most importantly the training data. As pointed out by Hansen and Salamon [7], a significant problem lies in the many local minima of the objective function used in tuning the weights; thus different training sessions using different setups will likely end up in "mixed" configurations representing different solutions in different regions of the data space.

One way to improve neural-network estimation is to use an ensemble of neural networks [2, 7, 8], hereafter called a neural network committee. In theory, the committee generalization error is guaranteed to be less than the (weighted) average of member-network errors; the smaller the correlation between the member networks, the smaller the committee error. In this paper, we study various ways to improve speech recognition by forming ensembles of neural network phone estimators in our hybrid speech recognizer. More specifically, context-dependent (CD) and context-independent (CI) neural network phone estimators are combined by interpolation and/or by collapsing CD network outputs into CI network outputs.

In the next section, we will briefly present the neural network committee theory and outline several methods in forming committees from various neural network phone estimators. Then in Section 3, we will describe our baseline systems and their performance on phone recognition. Phone recognition results using various committees are summarized in Section 4. Finally we discuss the significance of our findings in Section 5.

2. PHONE NETWORK COMMITTEE

2.1. Relationship Between Committee Error and Member Error

The training of our phone networks can be considered as learning the functional mapping between the input speech features and the posterior probabilities of the phones underlying the speech features. Let us consider a committee of K such networks using the following notations:

$\mathbf{x}$	:	speech vector of dimension N
$f_i(\mathbf{x})$	:	mapping function learned by the i-th member network
ω_i	:	weighting of the i-th member network in forming the committee
$\overline{f(\mathbf{x})}$	:	mapping function of the committee
$t(\mathbf{x})$	:	the true mapping function
$E[\cdot]$	:	expectation

Then it can be proved (see e.g. [8] for a detailed proof) that

$$\mathcal{E}\left[(t(\mathbf{x})-\overline{f(\mathbf{x})})^2\right] = \sum_{i=1}^{K}\omega_i\mathcal{E}\left[(t(\mathbf{x})-f_i(\mathbf{x}))^2\right] - \sum_{i=1}^{K}\omega_i\mathcal{E}\left[(f_i(\mathbf{x})-\overline{f(\mathbf{x})})^2\right]$$

where, $\overline{f(\mathbf{x})} = \sum_{i=1}^{K}\omega_i f_i(\mathbf{x})$ *and* $\sum_{i=1}^{K}\omega_i = 1$. That is,

$$\textit{Committee Error} = \textit{Weighted Member Error} - \textit{Weighted Member Variance}\ . \quad (1)$$

Since all the errors and variances are positive, the committee error must be smaller than or equal to the weighted member error. Hence in theory a committee, on average, always gives a better estimate than a member network.

2.2. Three Types of Network Committee

Due to memory and computation limitations in a practical speech application, only committees consisting of two members are studied. Let **X** be any CI monophone and **Y** be any CD phone of **X**; we will write **Y** $\in$ **X** to denote such a relationship. Also let X_{ci} be the output score of phone **X** of a CI network and Y_{cd} the output score of class **Y** of a CD network. Their committee outputs will be denoted as X_{com} and Y_{com} respectively and we will add superscripts 1, 2 to denote different networks of the same kind.

Type I: Homogeneous Committee (Homo)

A homogeneous committee is formed from two CI phone networks or two CD phone networks by linearly combining their corresponding outputs. i.e.

$$X_{com} = \alpha X_{ci}^{(1)} + (1-\alpha) X_{ci}^{(2)}, \quad 0 \leq \alpha \leq 1$$

Similarly,

$$Y_{com} = \alpha Y_{cd}^{(1)} + (1-\alpha) Y_{cd}^{(2)}, \quad 0 \leq \alpha \leq 1 .$$

Type II: Heterogeneous Committee (Hetero)

A heterogeneous committee is formed by combining a CI and a CD phone network. Since the CI network is usually more robust (due to more training data for each output), while the CD network is more precise (because of the detailed modeling of each context of the base phones), a combination may be both robust and precise.

(A) If the resulting committee emulates a CI network, outputs from all CD classes **Y** of base phone **X** of the CD member network are summed before interpolating with **X**'s score of the CI member network.

$$X_{com} = \alpha X_{ci} + (1-\alpha) \sum_{\mathbf{Y} \in \mathbf{X}} Y_{cd}$$

The special case when $\alpha = 0$ is of particular interest and will be called a "Collapsed Committee" (Collapse). Here the "committee members" are in fact different outputs from the same network which correspond to the same base phone.

(B) When the resulting committee is to emulate a CD network, each output Y_{cd} (whose base phone is **X**) of the CD member network is interpolated with **X**'s score in the CI network.

$$Y_{com} = \alpha Y_{cd} + (1-\alpha) X_{ci} , \quad where \ \mathbf{Y} \in \mathbf{X}$$

Type III: Background Committee (Backgr)

Two CD networks are combined by first collapsing the second member CD network to a CI network and then forming a heterogeneous committee with the first CD network. The resulting committee emulates a CD network.

$$Y_{com} = \alpha Y_{cd}^{\ (1)} + (1-\alpha) \sum_{\mathbf{Y}' \in \mathbf{X}} Y_{cd}'^{\ (2)}, \quad where \ \mathbf{Y} \in \mathbf{X}$$

Although the CD network may not be robust with respect to the CD phones, by summing up its outputs corresponding to each base phone, the resulting CI phone scores can be more robust and accurate. By interpolating the collapsed CI phone posterior probabilities with itself or another CD network, the collapsed CI probabilities provide robust "background" values which enhance the discriminability among CD phones belonging to different base phones. Furthermore, when the same CD network is used to form the background committee, the committee will be called a "Self-Background Committee" (S-Backgr); and, when a different network is used, it will be called a "Cross-Background Committee" (C-Backgr).

3. BASELINE SYSTEM

We have been working with both CI and CD phone networks for some time. The CI network has 40 OGIBET monophone outputs, while the CD network has 429 generalized biphone outputs; both are trained on the OGI_TS [10] (telephone speech) Corpus. The generalized biphones are derived from the same database with a data-driven approach using the Bhattacharyya distance as described in [9].

3.1. Training of the Phone Networks

Both types of phone networks have 56 input nodes and 50 hidden nodes. The 56 inputs represent 7-th order PLP coefficients and normalized energies from seven successive frames centered around the frame under investigation. Out of the 208 speech files in the OGI_TS Corpus, 148 speech files are used to derive training vectors, 30 for cross validation and another 30 for testing. Five thousand frames of training vectors are randomly selected for each CI class while 500 such frames are selected for each CD class [1]. They are trained for a maximum of 30 epochs using standard Backpropagation with the cross-entropy cost function. The best net is chosen with the cross validation data. Table 1 summarizes the architecture and some training parameters for the two types of phone networks.

3.2. Phone Recognition

Phone recognition is performed on the OGI_TS test set using Viterbi search. CI decoding uses both a bigram language model and gamma duration models as described in [4]. CD decoding uses only a bigram language model plus the generalized biphones constraint. The various thresholds and weightings of the language model and duration models are determined empirically using the validation data. The %insertions are kept to ~10% by adjusting the transition penalty.

[1] In the CD case, each of the 3 phones, "sil", "ucl" and "vcl" has more (5500) training frames to account for their greater variability.

Table 1. Architecture of the CI and CD phone networks. I=#inputs, H=#hidden units, C=#output classes, N=#training frame per class, P=total #parameters, R=#training frames per parameter(see footnote 1)

UNIT	I	H	C	N	P	R
CI	56	50	40	5000	4.8K	41.7
CD	56	50	429	500	24.3K	9.46

Table 2. Phone/frame recognition results of the baseline system

PHONE UNIT	TRAINING SET NO.	FRAME %CORRECT	PHONE %ACCURATE
CI	1	30.1	41.4
CI	2	30.3	39.3
CD	1	33.5	46.0
CD	2	34.4	46.4

Table 2 shows the phone and frame recognition results of various networks on the test dataset where phone accuracy is computed as 100% - (%substitutions + %deletions + %insertions).

4. COMMITTEE PHONE RECOGNITION

If the committee is to work, from Eq.(1), we see that the member errors have to be small and the member networks should exhibit a large variance. Usually this is achieved by using non-overlapping or partially-overlapping training data for the member networks; for example, crossnet [2], bootnet [11] and bagging [3]. In this study, due to the limited amount of training data, partially-overlapping training data are derived for the two member networks as follows: out of the 208 OGI_TS files, 30 are reserved for testing all member networks. Of the remaining 178 files, 30 files are randomly picked to be held out for cross validation and the rest is used for training the first member network. To derive training data for the second member network, another 30 files different from the 30 validation files used in the first member network are randomly picked for cross validation and the rest is used for training. In any case, training vectors are randomly sampled without replacement from the whole training dataset.

Each of the two partially-overlapping sets of training files is used to train one CD and one CI neural net. The three types of neural committee as described in Section 2 are then composed from the four networks so that the two member networks in a committee always come from different training datasets. Table 3 shows the performance of various committees on phone recognition on the test dataset using optimal weightings determined empirically in ascending order of recognition accuracies. The weightings are now the same for all phones. As another check on the committee theory, frame classification is also performed and its results are listed along in Table 3.

From Table 3, a conventional approach to combining two similar CI or CD phone networks into a homogeneous committee helps improve the phone recognition accuracy by 2.7% in monophone decoding and 1.5% in biphone decoding. Simply collapsing a CD network into a CI network increases phone recognition accuracy by 5.2% and the result is close to the phone recognition accuracy using a CD network with biphone decoding. This confirms that the CD phone network is superior due to context modeling. Still better performance of monophone decoding is obtained by forming a heterogeneous committee of a CI network with a collapsed CD network, which improves recognition accuracy by 6.3% and outperforms a CD network using biphone decoding.

A CD heterogeneous committee does not improve as much as a CI heterogeneous committee and is outperformed by a CD homogeneous committee. This shows again that the CI phone network is inferior. The biggest gain 2.2%, however, comes from the cross-background committee in which CD-1 is interpolated with CD-2 after CD-2 is collapsed into a CI network. This is probably due to the fact that the collapsed CI network is more robust, and the resulting committee can take advantage of both the preciseness of CD-1 and robustness of the collapsed CD-2.

There is no direct translation between frame classification results and phone recognition results. The language model and decoding constraints employed in phone recognition greatly affect phone recognition accuracy which may then deviate from the corresponding frame classification result. In our experiments, the frame classification results actually follow the phone recognition results closely, except in the case of the CD Self-Background committee. Most importantly, a consistent picture is seen to emerge – thus improving our confidence in the validity of the results observed.

5. DISCUSSION & FUTURE WORK

In the course of research, several phone neural networks are usually trained. Instead of throwing away all but the best network, we may make good use of the poorer networks by combining them with the best one to form a committee. Due to computation and memory constraints, a committee of two networks is probably the practical choice. In this paper we have shown that both CI and CD phone network committees perform better phone recognition. For a single network, CD phone decoding clearly gives better recognition results. This is, however, achieved at the expense of much more memory usage and computation time. For example, on average, monophone decoding using CI-1 takes only ~10s per test file (which is composed of ~50s of telephone speech) while biphone decoding using CD-1 takes ~80s on a DEC AlphaStation 200 4/166. For the same task, a CI homogeneous committee takes ~20s and a CI heterogeneous committee takes ~30s. In about one third of the computation time, a CI heterogeneous committee per-

Table 3. **Phone/frame recognition results of various phone network committees (For easy comparison, the result of member networks are included here and labelled as "Single" for Committee Type)**

Network Combination	Committee Type	Frame %cor	Phone				
			%acc	%cor	%sub	%del	%ins
CI-2	Single	30.3	39.3	49.0	35.3	15.8	9.66
CI-1	Single	30.1	41.4	51.2	34.7	14.1	9.76
CI-1,CI-2	Homo	32.3	43.0	52.7	33.8	13.5	9.74
CD-2	Collapse	37.9	45.5	55.6	30.1	14.3	10.0
CI-1,CD-2	Hetero(CI)	38.3	46.6	56.6	29.8	13.6	9.98
CD-1	Single	33.5	46.0	56.0	30.7	13.3	9.99
CD-2	Single	34.4	46.4	56.4	29.6	14.0	9.97
CD-2,CI-1	Hetero(CD)	35.4	47.2	56.8	29.4	13.8	9.61
CD-2	S-Backgr	37.1	47.2	56.9	28.9	14.3	9.68
CD-1,CD-2	Homo	35.9	47.7	57.3	29.6	13.1	9.63
CD-1,CD-2	C-Backgr	38.2	48.4	58.1	28.9	13.1	9.75

forms better than a single CD network. The technique is thus useful in time-critical applications.

In this paper the committee weightings are kept the same for all phones to reduce the complexity of the system. Phone-dependent (non-constant) weighting functions should give better performance. To further improve our understanding, we are putting the techniques into practical applications aimed at word recognition.

6. ACKNOWLEDGMENT

This work was supported through a grant in the Young Investigator Program of the Office of Naval Research. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

[1] E. Barnard, R. A. Cole, M Fanty, and P. Vermeulen. "Real-World Speech Recognition With Neural Networks". In R. J. Alspector and T. X. Brown, editors, *Applications of Neural Networks to Telecommunications (IWANNT95)*, volume 2, pages 186–193. Lawrence Erlbaum Assoc., Hillsdale, New Jersey, 1995.

[2] L. Breiman. "Stacked Regressions". Technical report, University of California, Berkeley, 1992.

[3] L. Breiman. "Bagging Predictors". Technical report, University of California, Berkeley, 1994.

[4] D. Burshtein. "Robust Parametric Modeling of Durations in Hidden Markov Models". *Proceedings of IEEE ICASSP*, pages 548–551, May 1995.

[5] R. A. Cole, M. Noel, T. Lander, and T. Durham. "New Telephone Speech Corpora at CSLU". *Proceedings of Eurospeech*, pages 821–824, Sep 1995.

[6] A. Waibel et al. "Phoneme Recognition Using Time-Delay Neural Networks". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(3):328–339, 1989.

[7] L. K. Hansen and P. Salamon. "Neural Network Ensembles". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, October 1990.

[8] A. Krogh and J. Vedelsby. "Neural Network Ensembles, Cross Validation and Active Learning". In D.S. Touretzky G. Tesauro and T. K. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 231–238. The MIT Press, 1995.

[9] B. Mak and E. Barnard. "Phone Clustering Using The Bhattacharyya Distance". *Proceedings of ICSLP*, pages 2005–2008, October 1996.

[10] Y.K. Muthusamy, R.A. Cole, and B. T. Oshika. "The OGI Multi-Language Telephone Speech Corpus". *Proceedings of ICSLP*, II:895–898, Oct 1992.

[11] B. Parmanto, P. Munro, and H. Doyle. "Improving Committee Diagnosis with Resampling Techniques". In M. Mozer D. S. Touretzky and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. The MIT Press, 1996.

[12] T. Robinson and F. Fallside. "A Recurrent Error Propagation Network Speech Recognition System". *Computer Speech and Language*, 5(3):259–274, 1991.

[13] V. Tresp and M. Taniguchi. "Combining Estimators Using Non-Constant Weighting Functions". In D.S. Touretzky G. Tesauro and T. K. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 419–426. The MIT Press, 1995.

[14] D. H. Wolpert. "Stacked Generalization". In *Neural Networks*, volume 5, pages 241–259. 1992.

[15] G. Zavaliagkos, Y. Zhao, R. Schwartz, and J. Makhoul. "Integration of Segmental Neural Nets with Hidden Markov Models". *Proceedings of ARPA/MTO CSR Workshop*, pages 71–76, 1991.