IMPROVED NEURAL NETWORK TRAINING OF INTER-WORD CONTEXT UNITS FOR CONNECTED DIGIT RECOGNITION

Wei Wei and Sarel van Vuuren

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
PO Box 91000, Portland, OR 97291-1000, USA

ABSTRACT

For connected digit recognition the relative frequency of occurrence for context-dependent phonetic units at inter-word boundaries depends on the ordering of the spoken digits and may or may not include silence or pause. If these units represent classes in a model this means that the distribution of samples between classes (the class prior) may be extremely nonuniform and that the distribution over many utterances in a training set may be very different from the rather flat distribution over any single test utterance. Using a neural network to model context-dependent phonetic units we show how to compensate for this problem. We do this by roughly flattening the class prior for infrequently occurring context units by a suitable weighting of the neural network cost function. This is based entirely on training set statistics. We show that this leads to improved classification of infrequent classes and translates into improved overall recognition performance. We give results for telephone speech on the OGI Numbers Corpus. Flattening the prior for infrequently occurring context units resulted in a 12.37% reduction of the sentence-level error rate (from 16.17% to 14.76%) and a 9.93% reduction of the word-level error rate (from 4.23% to 3.81%) compared to not doing any compensation.

1. INTRODUCTION

Under certain conditions a neural network trained on acoustic samples (speech feature vectors) of phonetic units can be used as a class probability estimator [1, 2]. During recognition the estimates provided by network outputs can be used by a Viterbi search [3] to find the most probable path through the input speech feature vectors, as shown in Figure 1. For the path that represents the actual spoken utterance to be probable it is important that the neural network estimate good posterior class probabilities. The posterior class probabilities can be seen to depend on the prior class probabilities using Bayes' rule: for an input vector X and classes C_i, $P(C_i|X) = p(X|C_i)P(C_i)/p(X)$. For optimum performance it is important that the prior class probabilities match the actual frequency of occurrences expected during recognition.

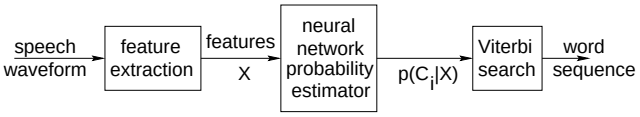

Figure 1: A neural network based speech recognition system.

In our experiments, classes are *context-dependent phonetic* units. These units are based on different acoustic realizations of a phone. We model the phonetic units which take account of the left or right contexts but not both. This is motivated in that inclusion of context-dependent phonetic models can improve discrimination (compared with using only context-independent phonetic models such as monophones). This follows because the portions of the human vocal tract that produce sound are constantly in motion and causes the phonetic units to vary depending on context. Figure 2 demonstrates context-dependent phonetic modeling for the digit /eight/.

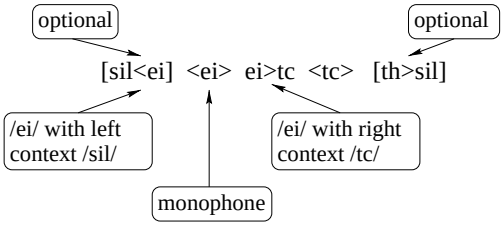

Figure 2: Context dependent phonetic modeling. For example, one model of digit /eight/, where /sil/ represents silence or pause and [ ] means optional.

For instance the phone /ei/ would be split into three

parts, modeling it in the context of a sound to the left (silence), a stationary part and in the context of a sound to the right (t closure)[1].

The use of context-dependent phonetic modeling results in disproportionate sample distributions. For example in a training set that is composed of many utterances of connected digits there are comparatively many more instances of a particular context occurring inside words (inner-word) than a particular context occurring between words (inter-word). This is of course obvious, since if there are 11 words in the vocabulary, the probability of seeing an inner-word context unit is about 1/11 whereas the probability of seeing an inter-word context unit is about $1/11^2$. Therefore in the training set inter-word context units may occur about an order of magnitude less frequently than inner-word context units. In the following we demonstrate this argument using statistics from a training set obtained from the OGI Numbers Corpus [4] which we use for the experiments in this paper. We describe the corpus in more detail in the results section.

For the digit /eight/ the number of instances for inner-word context units in the training set are:

1878	4096	3683	10616	1801
sil<ei	<ei>	ei>sil	<tc>	[th>sil]

On the other hand suppose the digit /eight/ is followed by the digit /zero/. This would introduce one or more inter-word context units. Here /eight/ and /zero/ may optionally be separated by silence or pause. For example inter-word context units include:

			813	
...	<tc>	[th>sil]	[sil<z]	...
			124	
...	ei>sil	<tc>	[tc<z]	...

Now suppose that /eight zero/ is the actual transcription of an utterance to be recognized. Then for *this utterance* it is easy to see that *all* the implied context units have roughly the same frequency of occurrence – they occur only about once irrespective of whether they are inner-word or inter-word context units. One can argue that all of these context units are roughly equally important for successful decoding of the utterance. (Post-analysis of the test set shows that the combination /eight zero/ actually occurred 30 times.)

During training of our neural network model we balance the frequency of occurrence of context units in the training set to reflect this flatter prior present in the test utterances.

[1]In our model a t closure occurring in right context gets absorbed into silence occurring in right context.

We will call the training samples of a class its in-class samples and the in-class samples of other classes its out-of-class samples. One way to balance the frequencies of classes during network training, is to blindly throw out training samples of the frequent classes. However, this is a counter-productive way because it may destroy the natural prior class probabilities that may be represented by the training samples (at least for frequent classes) and will not fully use the available training data. Other methods described in previous work includes duplicating the sample number of infrequent classes, weighing the in-class sample training of infrequent classes and compensating for different prior probabilities by the multiplication of class priors from the test set and the division of class priors from the training set [2]. Frequency balancing was also used in an image recognition task [5].

For connected digit recognition we propose in this paper roughly flattening the prior for infrequent classes. In Section 2 we propose separating classes into infrequent classes and frequent classes based on information obtained from training data, with each class being assigned a coefficient that is informative of its prior. In Section 3 we show how the neural network may be trained with a modified prior and how to modify the neural network weight-update functions based on these coefficients. In Section 4 we present experimental results on a connected digit recognition task.

2. SEPARATING CLASSES INTO INFREQUENT AND FREQUENT CLASSES

For a multi-layer perceptron (MLP) neural network, let X represent an input vector with elements $\{x_i : i = 1, \cdots, D\}$, $\{C_i : i = 1, \cdots, M\}$ M classes, $\{y_i(X) : i = 1, \cdots, M\}$ the network outputs, $\{d_i : i = 1, \cdots, M\}$ the desired outputs for all output nodes, and $\{h_j : j = 1, \cdots, K\}$ the outputs for all hidden nodes. Also, for a 1 of M classification problem, let $d_i = 1$ if X belongs to C_i and 0 otherwise.

Classes can be separated into two categories: a set of frequent classes (denoted as A_1) and a set of infrequent classes (denoted as A_2), based on the estimates of class priors. We then assign coefficients b_i that are informative of class priors to classes. If $C_i \in A_1$, let $b_i = 1$; if $C_i \in A_2$, let $0 < b_i < 1$. The information of class priors can be obtained for example by relative frequencies of occurrence of classes' training samples.

Let N represent the total number of training samples and $n_i (i = 1, \cdots, M)$ the number of in-class training samples of class C_i. A method for using relative frequencies of classes' training samples is:

Use a class' relative ratio of the number of out-of-

class training samples and in-class training samples, $(N-n_i)/n_i$, as a reference to separate infrequent classes and frequent classes. Ideally, if classes have equal amounts of training samples, the ratio of a class' out-of-class sample number and in-class sample number is $(M-1)/1$. If $(N-n_i)/n_i \leq (M-1)/1$, $C_i \in A_1$; otherwise, $C_i \in A_2$. Therefore, coefficients b_i can be defined as:

$$b_i = \begin{cases} 1 & \text{if } C_i \in A_1 \\ \frac{(M-1)/1}{(N-n_i)/n_i} & \text{if } C_i \in A_2 \end{cases}$$

For infrequent classes ($C_i \in A_2$) and $0 < b_i < 1$.

3. NEURAL NETWORK TRAINING WITH A MODIFIED PRIOR

In the Introduction we showed that infrequent classes may be under represented in the training set and that they have a disproportionate number of out-of-class samples compared to in-class samples. Another problem may occur when infrequent classes are short of training samples, in which case their in-class sample learning will be poor. We cannot do much about the latter problem since it is difficult to get enough training samples for infrequent classes because of their naturally rare frequencies. However de-weighing the contribution of out-of-class samples during training of infrequent classes may compensate for under-represented classes.

We propose a modified cross-entropy cost function to balance infrequent classes' relatively larger amounts of out-of-class sample learning:

$$\varepsilon = -E\{\sum_{i=1}^{M}[d_i log y_i(X) + b_i(1-d_i)log(1-y_i(X))]\} \quad (1)$$

Let w_{ij} represent the weight between output neuron i and hidden neuron j, h_j the output of hidden neuron j, and $\Delta w_{ij}^{(n)}$ the weight update of w_{ij} when the nth sample is trained using stochastic training.

$$\Delta w_{ij}^{(n)} = -\eta \xi_i h_j = \begin{cases} -\eta(y_i - d_i)h_j, & \text{if } d_i = 1 \\ -\eta b_i(y_i - d_i)h_j, & \text{if } d_i = 0 \end{cases}$$

If $\forall i, b_i = 1$ the modified cross-entropy cost function becomes the standard cross-entropy cost function [6]. Note that $0 < b_i \leq 1$ so that the learning rate of class i's output neuron is no greater than the global learning rate η. This guarantees that the usual constraints on the learning rate [7] is satisfied.

A similar procedure holds for batch training. In fact, for batch training summing over the training samples gives

$$\Delta w_{ij} = \sum_{n_i}(-(y_i - d_i)h_j) + b_i \sum_{N-n_i}(-(y_i - d_i)h_j) \quad (2)$$

which shows how b_i adjusts the prior.

If the network parameters are chosen to minimize the modified cross-entropy cost function (Equation 1), it can be shown that the outputs estimate the conditional expectations of the desired outputs with an adjusted value for infrequent classes:

$$y_i(X) = E\{d_i|X\}/(E\{d_i|X\} + b_i(1 - E\{d_i|X\})).$$

For a 1 of M problem, d_i equals one if the input X belongs to class C_i and zero otherwise. Therefore class C_i's conditional expectation of the desired output is

$$E\{d_i|X\} = \sum_{j=1}^{M} d_j p(C_j|X) = p(C_i|X)$$

. Therefore in the expectation

$$\begin{aligned} y_i &= p(C_i|X) \text{ if } b_i = 1, \text{ i.e., } C_i \in A_1 \\ y_i &> p(C_i|X) \text{ if } 0 < b_i < 1, \text{ i.e., } C_i \in A_2 \end{aligned} \quad (3)$$

which balances the ratio of out-of-class samples to in-class samples for infrequent classes.

4. EXPERIMENTAL RESULTS ON A CONNECTED DIGIT RECOGNITION TASK

Using the speech recognition system shown in Figure 1, we constructed a three layer MLP neural network model as the probability estimator. The neural network has 56 input nodes, 200 hidden nodes and 209 output nodes that correspond to 209 context-dependent phonetic classes. It was trained using stochastic backpropagation and a cross-entropy cost function. We separated the data set into three parts by the ratios 3:1:1 for training, cross-validation and test, respectively. The task is connected digit recognition using the OGI Numbers Corpus – a telephone speech corpus that contains "fluent" numbers. Callers were asked to leave their phone number, birth date, or zip code. Examples of digit utterances are /oh one oh nine three oh/, /two eight zero oh seven/ and /two one zero four two five three seven three five three zero/. Utterances contain 1 to 12 digits.

We trained three neural networks: (1) using throw-away policy to balance training samples and standard cross-entropy cost function, called *net*1 ; (2) using all the training data and standard cross-entropy cost function, called *net*2; (3) using all the training data with the modified cross-entropy function shown as Equation 1, called *net*3. The classes were separated into 146 infrequent classes and 63 frequent classes.

We did experiments to compare *net3* with *net2* in terms of the classification rate at class level based on the cross-validation set. Among the 209 classes, the

ESTIMATOR	*net1*	*net2*	*net3*
error(sentence)	18.45%	16.17%	14.76%
error(word)	4.94%	4.23%	3.81%

Table 1: Digit Recognition results on OGI Numbers Corpus (doing evaluation on 1626-sentence test set).

classification rates of 171 classes (including 130 infrequent classes) were increased, 28 classes (including 6 infrequent classes) decreased, and 10 classes (including 10 infrequent classes) did not have available cross-validation data. These results show that the modified cross-entropy cost function, which changes the prior for infrequent classes, also improves their classification accuracy. Figure 3 shows these results in terms of relative differences of classification accuracy.

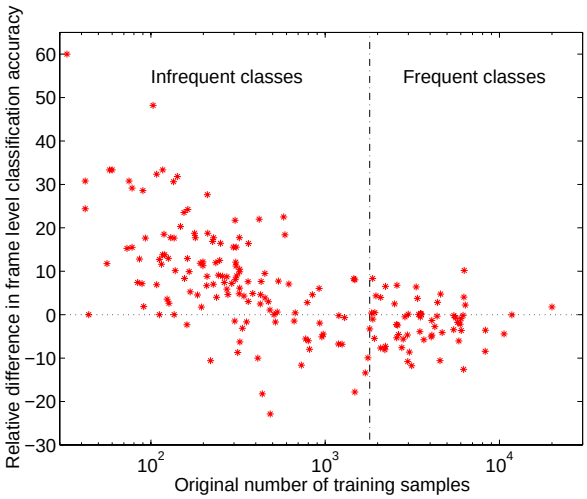

Figure 3: Frame level classification accuracy between *net*3 and *net*2 [absolute difference in percent accuracy] as a function of number of training samples. The prior for infrequent classes were changed during training.

Experimental results on the test set, Table 1, show that *net*3 resulted in 20% reduction in recognition errors at the sentence level and 22.87% at the word level as compared to *net*1, and 12.37% and 9.93% at sentence level and word level respectively, as compared to *net*2. The number of word insertion and deletion decreased from 248 (*net*1) and 182 (*net*2) to 152 (*net*3). The results of McNemar's significance test [8] show that the observed differences would arise by chance on occasion about 0.1% from *net1* to *net3* and 1.4% from *net2* to *net3*. These results indicate that the improvement of recognition performance is statistically significant.

5. CONCLUSION

We observed that the prior over a training set for infrequently occurring context units may not be matched to any one particular test utterance. We proposed to balance the prior for such infrequently occurring context units by de-weighing the contribution of out-of-class samples in the cross-entropy cost function. A frame level analysis showed that a model trained using this modified prior models the inter-word context units better. We showed that this translates into statistically significant reduction in word-level and more importantly sentence-level error rate.

Acknowledgements

We would like to acknowledge Mark Fanty for providing a platform for some of the neural-network based context dependent phonetic modeling. This research was supported in part by grant IRI-9314959 from the National Science Foundation NSF.

6. REFERENCES

[1] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. Wiley-Interscience, 1973.

[2] M. D. Richard and R. P. Lippmann, "Neural network classifiers estimate Bayesian a posteriori probabilities," *Neural Computation*, no. 3, pp. 461–483, 1991.

[3] N. Morgan and H. Bourlard, "Continuous speech recognition," *IEEE Signal Processing Magazine*, pp. 25–42, May 1995.

[4] R. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at CSLU," *Proceedings of the Fourth European Conference on Speech Communication and Technology*, vol. 1, pp. 821–824, 1995.

[5] R. Lyon and L. Yaeger, "On-line hand-printing recognition with neural networks," *Fifth International Conference on Microelectronics for Neural Networks and Fuzzy Systems*, Feb. 1996. Lausanne, Switzerland.

[6] J. B. I. Hampshire and B. A. Perlmutter, "Equivalence proofs for multilayer perceptron classifiers and the Bayesian discriminant function," *Proceedings of the 1990 Connectionist Models Summer School*, 1990. D. Touretzky, J. Elman, T. Sejnowski, and G. Hinton, eds. Morgan Kaufmann, San Mateo, CA.

[7] Y. Le Cun, P. Simard, and B. Pearlmutter, "Automatic learning rate maximization by on-line estimation of the Hessian's eigenvectors," pp. 156–163, 1993.

[8] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 532–535, 1989.