Replicator Neural Networks for Outlier Modeling in Segmental Speech Recognition

László Tóth and Gábor Gosztolya

Research Group on Artificial Intelligence
H-6720 Szeged, Aradi vértanúk tere 1., Hungary
{tothl, ggabor}@inf.u-szeged.hu

Abstract. This paper deals with outlier modeling within a very special framework: a segment-based speech recognizer. The recognizer is built on a neural net that, besides classifying speech segments, has to identify outliers as well. One possibility is to artificially generate outlier samples, but this is tedious, error-prone and significantly increases the training time. This study examines the alternative of applying a replicator neural net for this task, originally proposed for outlier modeling in data mining. Our findings show that with a replicator net the recognizer is capable of a very similar performance, but this time without the need for a large amount of outlier data.

1 Introduction – Neural Nets in Speech Recognition

Speech recognition does not naturally fit into the usual pattern classification scheme where the items to be classified are represented by a fixed number of features. Rather, speech is a continuous stream of information where the possible utterances vary in length and their number is practically unlimited. A possible solution is to trace the problem back to the recognition of some properly chosen (fixed-size) building blocks. During recognition these building blocks have to be found, identified, and the information they provide needs to be combined. The most successful solution, Hidden Markov Modeling (HMM) [6], does exactly this. The traditional HMM methodology seeks to model the distribution of the building units by means of Gaussian mixtures. Applying neural nets for this task instead became very popular in the mid-nineties after it became widely known that Artificial Neural Nets (ANN) approximate the class posteriors. It was claimed that ANNs are more flexible, attain a better performance due to their discriminative nature, and require an order of magnitude fewer parameters. Since then ANNs have become a widely accepted alternative to Gaussian-based modeling in the speech community under the name "HMM/ANN hybrids" [1].

In addition to this, many authors have criticized two other important aspects of hidden Markov modeling, namely the choice of units and the way their probability scores are aggregated. Traditionally, the building units are small uniform ($\sim$30ms) signal chunks, and the probabilities assigned to them are combined by multiplication. Both of these simplifications are unrealistic from a perceptual point of view. Several alternatives have been proposed, and one of them is the so-called segmental modeling approach where the building units are longer, variable-length signal intervals [8]. This technology offers

F. Yin, J. Wang, and C. Guo (Eds.): ISNN 2004, LNCS 3173, pp. 996–1001, 2004.

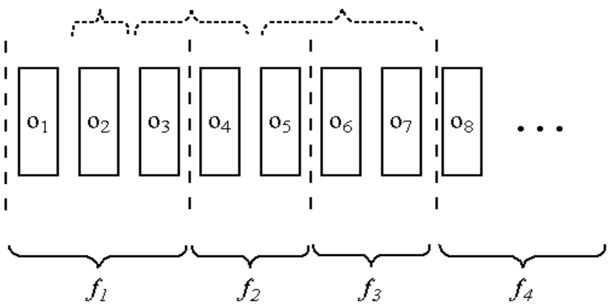

Fig. 1. An illustration of the connection between observation vectors and phonetic segments in speech recognition. The dotted brackets on top denote examples of anti-phone segments

better recognition results, but also introduces special problems. One of these, the problem of outlier segments, is in the focus of this paper. Section 2 really describes the problem in a nutshell, and Section 3 presents one possible solution, replicator neural networks (RNN). In Section 4 we discuss the experiments we preformed to justify the use of RNNs, then we round off with some remarks in Section 5.

2 Segmental Modeling and the Problem of Outliers

The preprocessed speech signal arrives at the input of a speech recognizer in the form of a vector series $o_1, ..., o_T$ (see Fig. 1.). The goal of recognition is both to find the correct phonetic segmentation of the signal (denoted by dashed lines in Fig. 1) and to correctly identify the segments in the form of phonetic labels $f_1, ..., f_N$. Let us assume for the moment that the positions of the segment boundaries have been determined, so the task is only to identify the segments. In the standard frame-based approach the ANN is responsible for supplying label-probabilities to each vector (data frame). The probability associated with the whole segment is then obtained by multiplying the frame-based values. In segmental models the features vectors of a segment are first transformed into a (fixed-dimensional) segmental feature set, and this forms the input data for the ANN. Here, the neural net is responsible for identifying the *whole segment in one*, not merely its individual feature vectors. It was reported by several authors that in segment classification this approach beats the standard frame-based technique even with a surprisingly simple segmental feature set [2][3][7][9].

Thus far, however, we assumed the segment boundaries were given. In real life this this is not so, of course. The solution is to evaluate every reasonable set of segment boundaries, thus incorporating the task of segmental classification in a search process. During this search the recognizer will be confronted with vector subseries that do not correspond to real phonetic segments, that is, they overlap real phonetic boundaries. The frame-based system will automatically handle these "anti-phones": if the frame-based probability of a class label f is high on some part of the segment, but is low on another one, then after multiplication the whole segment will get a small probability of being the phone f. The ANN of the segmental system, however, cannot automatically cope with segments like these. This is because it is usually trained on a manually segmented database, which naturally contains only examples of real phonetic segments. So the anti-phone segments encountered during recognition are not seen during training, and behave as outliers from a classification point of view. Moreover, the ANN has no way of reporting these outliers as it has only outputs for the different class labels. One possible solution might be to insert an additional component into the model that assesses the correctness of the given segmentation [10]. Another option is to introduce an additional class into the segmental classifier that will correspond to the outlier segments [3][9]. In the latter case we are confronted by the problem of how to train the outlier class. In our earlier work we generated outlier examples by taking quasi-random segments from the training corpus so that they overlap real segments or are incorporated within them (e.g. the intervals bounded by the dotted brackets in Fig. 1) [9]. On average, six such anti-phone samples were created per phone, thus seriously increasing the amount of training data needed. In practice, however, we found that these examples are not representative enough in the sense that the recognizer still behaves unexpectedly in many cases (i.e. it accepts obvious outliers as phones). It would have been possible to generate even more outlier examples, but we preferred to avoid this option for several reasons. These are the following:

- Apart from obvious cases (e.g. segments that heavily overlap a real boundary) it is not a trivial matter to see how the anti-phone segments should be generated. It might be, for example, that the segments generated according to Fig. 1 could still sound like one phone so, perceptually, they are not really anti-phones. In addition, the manual segmentation of the training corpus may also contain mistakenly positioned boundaries.
- Generating even more anti-phones per segment would cause the training data to be overwhelmed by one class which, as we observed, has a detrimental effect on the learning process.
- One characteristic of speech recognition is that the training databases are enormous. Even the training corpora that are considered small contain hundreds of thousands of phone instances. Creating dozens of outlier examples for each of these really did not sound appealing, especially regarding the training time.

This is why we were looking for a method that allows 1-class learning, that is learning from positive examples (in our case phonetic segments) only. Unfortunately, standard perceptron-based neural nets are not suitable for this task mainly because their responses are not localized. A network with radial basis functions (RBFN) would have been a possible choice, but we did not want to give up our well-tried multilayer perceptron network. Instead, we were looking for some simple extension to our current system. This is where replicator neural networks come in the picture.

3 Replicator Neural Networks

The basic idea behind a Replicator Neural Net (RNN) [4][5] is simple enough: the input data is also used as the desired output data. Consequently, by minimizing the mean square error during training we force the net to reconstruct its training patterns with the smallest

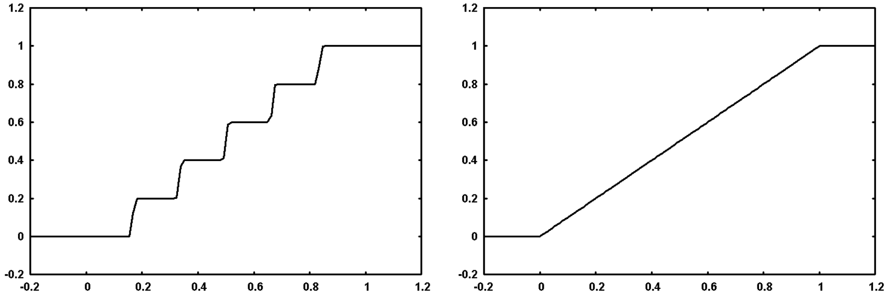

Fig. 2. The staircase-like activation function and the ramp-like one obtained when increasing the number of steps to infinity

error possible. During testing we hope that outlier patterns (patterns not in the training set) will be less well reproduced by the trained RNN and have a higher reconstruction error. Thus the reconstruction error can be used as a measure of 'outlyingness' of a test pattern.

RNNs were originally introduced in the field of data compression [5]. Hawkins et al. proposed it for outlier modeling [4]. In both papers a 5-layer structure is recommended, with a linear output layer and a special staircase-like activation function in the middle layer (see Fig. 2). The role of this activation function is to quantize the vector of middle hidden layer outputs into grid points and so arrange the data points into a number of clusters. Although this component plays a theoretically important role in the performance on the RNN, it makes learning by back-propagation practically impossible because its derivative is close to zero almost everywhere. Fortunately, Hecht-Nielsen argues that, by increasing the number of quantization levels to infinity, we arrive at a ramp-like activation function (see Fig. 2) by which "real-world problems might be solved". In the experiments we tried both the staircase, the ramp-like and the traditional sigmoid activation functions in the middle layer. All the other neurons were tested with both sigmoid and $tanh$ activations. Afterwards, we experimented with varying the number of layers and hidden neurons as well.

4 Experiments and Results

The speech database used in the experiments consisted of isolated words pronounced by children from the lower classes of elementary schools, originally recorded for the purpose of a teaching reading software package. The train/test sets consisted of 4000/920 utterances, both from a 2000-word dictionary. This recognition task proved quite difficult owing to the high variability in the children's voices and recording conditions, and because there were many similar-sounding words in the dictionary.

For the signal representation we tried two different segmental feature sets, one that had proved quite effective in our previous experiments [9] and one suggested by the literature [2]. The former consisted of 77 features per segment, while the latter contained 61 features.

Table 1. Word recognition accuracies on the two feature sets, depending on the anti-phone model used

	Feature Set	
Anti-Phone Model	OASIS	SUMMIT
No anti-phone model	67.17%	68.58%
Anti-phone class /w examples	72.28%	77.28%
RNN	72.39%	75.21%

Our speech recognizer was configured to three different settings. In one case no anti-phone model was used at all – that is, the ANN was trained only on correct phonetic segments. In the second setup ANN was extended with an outlier class and its training examples were generated as described in Section 2. Finally, in the third setup there was again no outlier class in the ANN, but an additional RNN was used to model the anti-phone segments (its reconstruction error being converted into the (0,1) interval by a sigmoid).

With the RNN, we first experimented with the special staircase-like activation function of the middle layer. As expected, we could not get back-propagation to converge when using the staircase-like activation function. However, it converged nicely both with the ramp-like and sigmoid activation and these produced very similar results. In all the other layers both sigmoid and $tanh$ activations were employed, and the sigmoid was found to converge somewhat faster. Based on these findings, we applied sigmoid activations in all layers (apart from the linear output layer) in all the subsequent experiments.

When varying the structure of the net, we found no advantage of using five layers. We obtained similar results with four or three layers only, and with a faster training time, so we settled on using a 3-layer model. When varying the number of hidden neurons in its hidden layer, the optimal performance was found to be about 25 hidden units. It was optimal in the sense that adding more units did not bring any further significant improvement.

The speech recognition results are listed in Table 1, both for our earlier feature set (OASIS) and the one from the literature (SUMMIT). The first thing to notice is that the feature set we developed previously performed worse here than the one suggested by the literature, no matter which anti-phone model was applied. This is probably because our representation was fine-tuned to another, definitely easier and simpler recognition task.

As regards the need for an anti-phone component, the significant improvement they bring over the "no anti-phone model" case clearly justifies their importance. It is hard to see, however, why they were less helpful on one feature set than on the other. This issue needs further examination.

Lastly, let us examine how the RNN performs as an anti-phone model compared to our earlier methodology. In one case it led to exactly the same performance, while in the other it yielded only slightly worse results. This shows that RNN is a viable alternative to our previous method that required the generation of a huge amount of outlier samples, and, consequently, a prolonged training time.

5 Conclusions

This paper investigated the feasibility of using a replicator neural network to assess the outlyingness of hypothesized segments in a segmental speech recognizer. This was motivated by the hope that, by doing this, a relatively simple and efficient model could replace the tedious process of generating and training outlier samples for a traditional MLP. The experiments justified our belief that RNNs indeed have the potential for this task as they yielded a performance similar to our usual methodology. Now, further studies are required to understand under what conditions they may behave worse (as in the second feature set) than our standard system, and whether they can be made to outperform it. Hence we plan to conduct more experiments in the future to precisely identify what these factors are.

References

1. Bourlard, H. A., Morgan, N.: Connectionist Speech Recognition – A Hybrid Approach. Kluwer Academic (1994)
2. Clarkson, P., Moreno, P. J.: On the Use of Support Vector Machines for Phonetic Classification. Proceedings of ICASSP'99 (1999) 585-588
3. Glass, J. R.: A Probabilistic Framework for Feature-Based Speech Recognition. Proceedings of ICSLP'96 (1996) 2277-2280
4. Hawkins, S., He, H. X., Williams, G. J., Baxter,R. A.: Outlier Detection Using Replicator Neural Networks. Proc. DaWak'02 (2002)
5. Hecht-Nielsen, R.: Replicator Neural Networks for Universal Optimal Source Coding. Science, Vol. 269. (1995) 1860-1863
6. Huang, X. D., Acero, A., Hon, H-W.: Spoken Language Processing. Prentice Hall (2001)
7. Kocsor, A., Tóth, L., Kuba Jr., A., Kovács, K., Jelasity, M., Gyimóthy, T., Csirik, J.: A Comparative Study of Several Feature Space Transformation and Learning Methods for Phoneme Classification. International Journal of Speech Technology, Vol. 3, Number 3/4 (2000) 263-276
8. Ostendorf, M., Digalakis, V., Kimball, O. A.: From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition. IEEE Trans. ASSP, Vol. 4. (1996) 360-378
9. Tóth, L., Kocsor, A., Kovács, K.: A Discriminative Segmental Speech Model and its Application to Hungarian Number Recognition. Proc. TSD'2000 (2000) 307-313
10. Verhasselt, J., Illina, I., Martens, J. P., Gong, Y., Haton, J. P.: Assessing the Importance of the Segmentation Probability in Segment-Based Speech Recognition. Speech Communication, Vol. 24, No. 1 (1998) 51-72