Continuous Speech Recognition Using Segmental Neural Nets

S. Austin†, J. Makhoul†, R. Schwartz†, and G. Zavaliagkos‡

†BBN Systems and Technologies, Cambridge, MA 02138
‡Northeastern University, Boston, MA 02115

ABSTRACT

We present the concept of a "Segmental Neural Net" (SNN) for phonetic modeling in continuous speech recognition. The SNN takes as input all the frames of a phonetic segment and gives as output an estimate of the probability of each of the phonemes, given the input segment. By taking into account all the frames of a phonetic segment simultaneously, the SNN overcomes the well-known conditional-independence limitation of hidden Markov models (HMM). However, the problem of automatic segmentation with neural nets is a formidable computing task compared to HMMs. Therefore, to take advantage of the training and decoding speed of HMMs, we have developed a novel hybrid SNN/HMM system that combines the advantages of both types of approaches. In this hybrid system, use is made of the N-best paradigm to generate likely phonetic segmentations, which are then scored by the SNN. The HMM and SNN scores are then combined to optimize performance. In this manner, the recognition accuracy is guaranteed to be no worse than the HMM system alone.

1 Introduction

The current state of the art in continuous speech recognition (CSR) is based on the use of HMMs to model phonemes in context. Two main reasons for the popularity of HMMs is their high performance, in terms of recognition accuracy, and their computational efficiency (after initial signal processing, real-time recognition is possible on a Sun 4 [1]). However, the limitations of HMMs in modeling the speech signal have been known for some time. Two such limitations are (a) the conditional-independence assumption, which prevents a HMM from taking full advantage of the correlation that exists among the frames of a phonetic segment, and (b) the awkwardness with which segmental features (such as duration) can be incorporated into HMM systems. We have developed the concept of Segmental Neural Nets (SNN) to overcome the two HMM limitations just mentioned for phonetic modeling in speech. However, neural nets are known to require a large amount of computation, especially for training. Also, there is no known efficient search technique for finding the best scoring segmentation with neural nets in continuous speech. Therefore, we have developed a hybrid SNN/HMM system that is designed to take full advantage of the good properties of both methods: the phonetic modeling properties of SNNs and the good computational properties of HMMs. The two methods are integrated through a novel use of the N-best paradigm developed in conjunction with the BYBLOS system at BBN.

2 Segmental Neural Net Structure

There have been several recent approaches to the use of neural nets in CSR. The SNN differs from these approaches in that it attempts to recognize each phoneme by using all the frames in a phonetic segment simultaneously to perform the recognition. In fact, we define a SNN as a neural network that takes the frames of a phonetic segment as input and produces as output an estimate of the probability of a phoneme given the input segment. But the SNN requires the availability of some form of phonetic segmentation of the speech. To consider all possible segmentations of the input speech would be computationally prohibitive. We describe in the next section how we use the HMM to obtain likely candidate segmentations. Here, we shall assume that a phonetic segmentation has been made available.

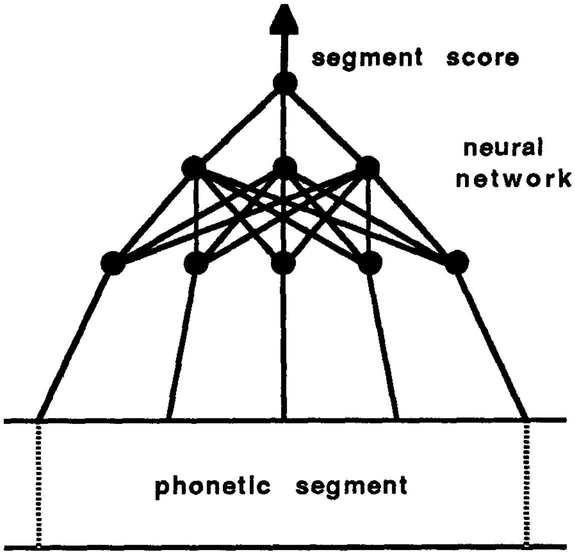

Figure 1: The Segmental Neural Network model samples the frames in a segment and produces a single segment score.

The structure of a typical SNN is shown in Figure 1. The input to the net is a fixed number of frames of speech features (5 frames in our system). The features in each 10-ms frame currently include 14 mel-warped cepstral coefficients, cepstral differences in time, power, and power difference. But the actual number of such frames in a phonetic segment is variable. Therefore, we convert the variable number of frames in each segment to a fixed number of frames (in this case, five frames). In this way, the SNN is able to deal effectively with variable-length segments in continuous speech. The requisite *time warping* is performed by a quasi-linear sampling of the feature vectors comprising the segment. For example, in a 17-frame phonetic segment, we would use frames 1, 5, 9, 13, and 17, as input to the SNN. In a 3-frame segment, the five frames used are 1, 1, 2, 3, 3, with a repetition of the first and third frames. In this sampling, we are using a result from stochastic segment models (SSM) in which it was found that sampling of naturally-occurring frames gives better results than strict linear interpolation [5].

Far from discarding duration information, which is implied in the warping to fixed length, the duration of the original segments can be handed to the neural net as just another feature that can be weighted according to its significance for recognition.

Therefore, by looking at a whole phonetic segment at once, we are able to take advantage of the correlation that exists among frames of a phonetic segment, and by making explicit use of duration as another feature, we are able to fully utilize duration information, thus ameliorating both limitations of HMMs. These properties of the SNN are also shared by the SSM, which was originally developed at BBN [5]. The main difference between the two is in how the probability of a segment is computed. In the SSM, an explicit multi-dimensional probability model has to be used (usually Gaussian) with many simplifying assumptions, so as to reduce the large amount of computation for training and recognition that would be needed in a model that has a complete covariance matrix. In contrast, the SNN has been shown to be capable of implicitly generating an estimate of the posterior probability without the need for an explicit model[2, 3]. In this way, we believe that the neural net will use as much correlation among frames as is needed to enhance performance.

In our initial experiments, we are using a single SNN with 53 outputs, each representing one of the phonemes in our system. The SNN outputs are trained with a 1 for the correct phoneme and a 0 for all the others.

3 Integration of Algorithms Using the N-Best Paradigm

In continuous speech recognition, many systems produce as output a single transcription that best matches the input speech, given some grammar. Because of imperfections in the recognition, the output may not be the correct sentence that was uttered and anything using this output (such as a natural language part of a speech understanding system) will be in error. One way to avoid this is to use a search that produces not only the single best-matching sentence but also the N-best matching sentences [6], where N is taken to be large enough to include the correct sentence most of the time (N is usually anywhere between 20 and 100 in our system, depending on the perplexity of the task; a higher N is needed for higher perplexity). The list of N sentences is ordered by overall score in matching the input utterance. For integration with natural language, we send the list of N sentences to the natural language component, which processes the sentences in the order given and chooses the first sentence that can be understood by the system.

In the hybrid SNN/HMM system, we use this N-best paradigm differently. A spoken utterance is processed by the HMM recognizer to produce a list of the N best-scoring sentence hypotheses. The length of this list is chosen to be long enough to include the correct answer almost always. Thereafter the recognition task is reduced to selecting the best hypothesis from the N-best list. As mentioned above, this list is usually between 20 and 100, which means that the search space of possible word theories is reduced from a huge number (for a 1000 word vocabulary, even a two word utterance has a million possible word hypotheses) to a relatively very small number. This means that each of the N hypotheses can be examined and scored using algorithms which would have been computationally impossible with a combinatorially large set of hypotheses. In addition, it is possible to generate several types of scoring for each hypothesis. This not only provides a very effective way of comparing the effectiveness of different speech models (e.g., SNN versus HMM), but it also provides an easy way to combine several radically different models.

The most obvious way in which the SNN could use the N-best list would be to derive a SNN score for each hypothesis in the N-best list and then reorder this list on the basis of these scores. The proposed answer would be the hypothesis with the best SNN score. However, it is possible to generate several scores for each hypothesis, such as SNN score, HMM score, grammar score, and the hypothesized number of words. We can then generate a composite score by, for example, taking a linear combination of the individual scores. It is also possible to choose the weights for this linear combination by automatically searching for the combination which minimizes a measure of the rank of the correct hypotheses over a training corpus [4].

4 Hybrid SNN/HMM System Using N-Best

As mentioned above, recognition in the hybrid SNN/HMM system is performed by using the SNN scores together with HMM and other scores to reorder the N-best list of likely hypotheses for the utterance. The process is shown schematically in Figure 2. A constrained HMM recognition is performed for each of the N-best hypotheses in turn. This provides both the HMM version of the acoustic score and the segmentation of the utterance for each of the N hypotheses. Of course, only one of these hypotheses can be correct, but this is not a problem since a bad segmentation for the incorrect hypothesis will lead to a correspondingly poor SNN score. This means that the incorrect hypothesis will not only be penalized because of a bad acoustic match, but it will also be penalized because of a malformed segmentation.

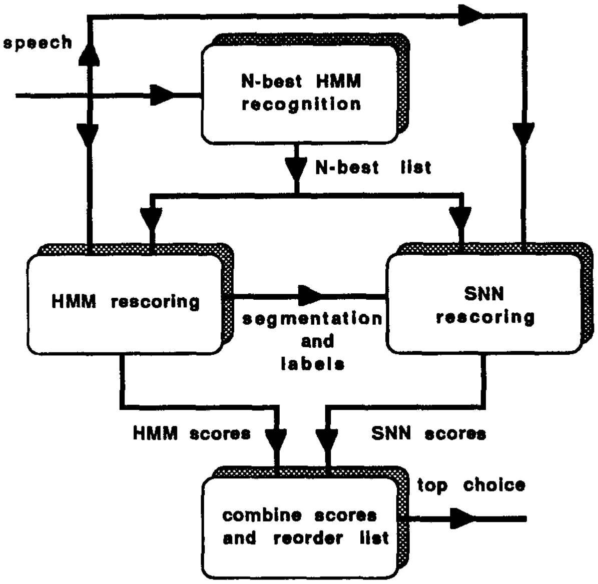

Figure 2: Schematic diagram of the N-best rescoring system using the SNN score.

The SNN uses the segmentation and phonetic sequence produced by the HMM under each hypothesis to construct feature vectors from each segment in the same way as in the training procedure. The neural net produces a score between 0 and 1 for each segment, which gives an estimate of the probability that the segment actually corresponds to the hypothesized phoneme. The logarithm of all these segment scores are computed and added together to produce a SNN score for the particular hypothesis.

For each hypothesis, a total score is then computed by taking a linear combination of the SNN score, HMM score, and other scores computed solely from the text of the hypothesis (e.g., grammar score, number of words). The weights for the linear combination are found by training on a development corpus that is different from the training corpus used to train both the HMM and SNN. A different corpus is used since the acoustic scores generated from training data will be unrealistically optimistic.

It is important to note that, because of the use of weighting to optimize peformance in this hybrid system, overall recognition accuracy can never be worse than with the HMM system alone. How much better the hybrid system will be depends on how well the SNN performs and how different are the errors made by the HMM and SNN systems alone.

5 Results

In our initial experiments, we used a version of the BYBLOS HMM system with non-crossword, context-dependent triphones, to compute the N-best sentence hypotheses. N was set to 20 in our experiments. We used a single context-independent SNN with 53 outputs. The neural net had only a single layer. The training and test data were obtained from the DARPA Resource Management speaker-dependent corpus, which consisted of data from 12 male and female speakers. In order to provide a realistic framework for the recognition, a statistical class grammar with perplexity 100 was used.

Under these conditions, the HMM system alone gave a word error rate of 9.1%, the SNN system alone gave a word error rate of 20.3%, and the hybrid SNN/HMM system gave a word error rate of 8.5%. The small reduction in error rate in the hybrid system over the HMM system is quite reasonable, considering the relatively large error rate of the SNN system alone. The poor performance of the SNN system was expected because the SNN was really primitive, both in terms of structure and the fact that it was context-independent. We expect that, as we enhance the structure of the SNN and make it context dependent, the performance of the SNN will improve and so will that of the hybrid system.

6 Conclusions and Further Work

The ultimate purpose of investigating new speech recognition algorithms is to improve on the performance of existing algorithms. Our hybrid SNN/HMM system has the advantage that its performance cannot be inferior to that of the corresponding HMM system alone. The neural network in this initial version of the SNN is a very simple model. It uses a one-layer neural net modelling context-independent phonemes. Even so, it produces a slight increase in accuracy over the context-dependent HMMs. Future developments of the SNN system will include the modelling of context-dependent phoneme segments, will use more sophicsticated neural networks, and will add additional features in order to model phoneme segments more closely.

Acknowledgments

The authors would like to thank Amro El-Jaroudi of the University of Pittsburgh for his help in running some of the experiments. This work was sponsored by DARPA.

References

[1] Austin, S., Peterson, P., Placeway, P., Schwartz, R., Vandegrift, J., "Towards a Real-Time Spoken Language System Using Commercial Hardware," *Proceedings of the DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, June 1990.

[2] El-Jaroudi, A. and Makhoul, J., "A New Error Criterion for Posterior Probability Estimation with Neural Nets," *International Joint Conference on Neural Networks*, San Diego, CA, June 1990, Vol III, pp. 185-192.

[3] Gish, H., "A Probabilistic Approach to the Understanding and Training of Neural Network Classifiers," ICASSP-90, April 1990, Albuquerque, NM, pp. 1361-1368.

[4] Ostendorf, M., Kannan, A., Austin, S., Kimball, O., Schwartz, R., Rohlicek, J.R., "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses," *Proceedings of the DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, February 1991.

[5] Ostendorf, M. and Roukos S., "A Stochastic Segment Model for Phoneme-based Continuous Speech Recognition," *IEEE Trans. Acoustic Speech and Signal Processing*, Vol. ASSP-37(12), pp. 1857–1869, December 1989.

[6] Schwartz, R. and Y.L. Chow (1990) "The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses," ICASSP-90, April 1990, Albuquerque, NM, pp. 81-84. Also in *Proceedings of the DARPA Speech and Natural Language Workshop*, Cape Cod, MA, Oct. 1989.