AN ADVANCED SYSTEM TO GENERATE PRONUNCIATIONS OF PROPER NOUNS

Neeraj Deshmukh, Julie Ngan, Jonathan Hamaker, Joseph Picone

Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University, Mississippi State, Mississippi 39762
{deshmukh, ngan, hamaker, picone}@isip.msstate.edu

ABSTRACT

Accurate recognition of proper nouns is a critical component of automatic speech recognition (ASR). Since there are no obvious letter-to-sound conversion rules that govern the pronunciation of any large set of proper nouns, this is an open-ended problem that evolves constantly under various sociolinguistic influences. A Boltzmann machine neural network is well-suited for the task of generating the most likely pronunciations of a proper noun. This pronunciation output can be used to build better acoustic models for the noun that result in improved recognition performance. We present here an advanced version of this N-best pronunciations system; and a multiple pronunciations dictionary of 18000 surnames and 25000 pronunciations used as a training database. The database and software are available in the public domain.

1. INTRODUCTION

The quality and effectiveness of a voice interface is often determined by its ability to accurately recognize proper nouns. For instance, in many applications related to medicine [1], the ability to recognize a physician's or patient's name is crucial. Also, it is well-known that a majority of word errors in ASR systems are due to misrecognized proper nouns. Lack of good pronunciation models is the prime factor that affects recognition of these words that often have multiple unfamiliar pronunciations.

We have demonstrated in [2] that a Boltzmann machine network can be successfully used to automatically derive multiple pronunciation models from the text-only spellings of a proper noun. The system architecture allows efficient searches for hypotheses (encoded as combinations of the binary states of various network units or neurons) that maximally satisfy the constraints resulting from the input data and the weighted interaction between individual units [3, 4] by capturing the input statistics seen during training.

We have extended the basic network in [2] by adding more flexibility in architecture and algorithmic features to create a more powerful system capable of efficiently generating an ordered list of the N most likely pronunciations of the input proper noun.

2. SYSTEM ADVANCES

The N-best proper nouns pronunciation system using the Boltzmann machine network architecture requires a text-based spelling of the proper noun as input and generates a network of phonemes that constitute the N most likely pronunciations of peoples' surnames. This architecture can now be automatically extended to other lexical domains by simply feeding the network the new symbol sets.

2.1. System Parameters

While the basic system was constrained to a fixed-context single-hidden-layer architecture, the advanced version is capable of dynamically supporting multiple number of hidden layers and user-defined context lengths. The system allows both single-context and two-context modes (short and long). The number of internal neuron layers and the number of neurons in each of them can also be specified by the user. The long and short contexts, if used in conjunction, may have a different number of hidden layers. This provides the network with a tremendous ability to model the letter-to-sound mappings in an optimal fashion.

2.2. Training

The connection weight values associated with the network are derived using a simulated annealing backpropagation technique [5] and a training database that contains the spelling and all expected pronunciations of each name. The training algorithm is summarized in Figure 1. The user controls the number of training iterations, the stopping criteria and the training (simulated annealing) schedule. The advanced system allows training an already trained network with incremental data. Also, the system learning rate can be made adaptive to the magnitude of the network error to expedite convergence. The network is also capable of simulating a multilayered perceptron (MLP) [6].

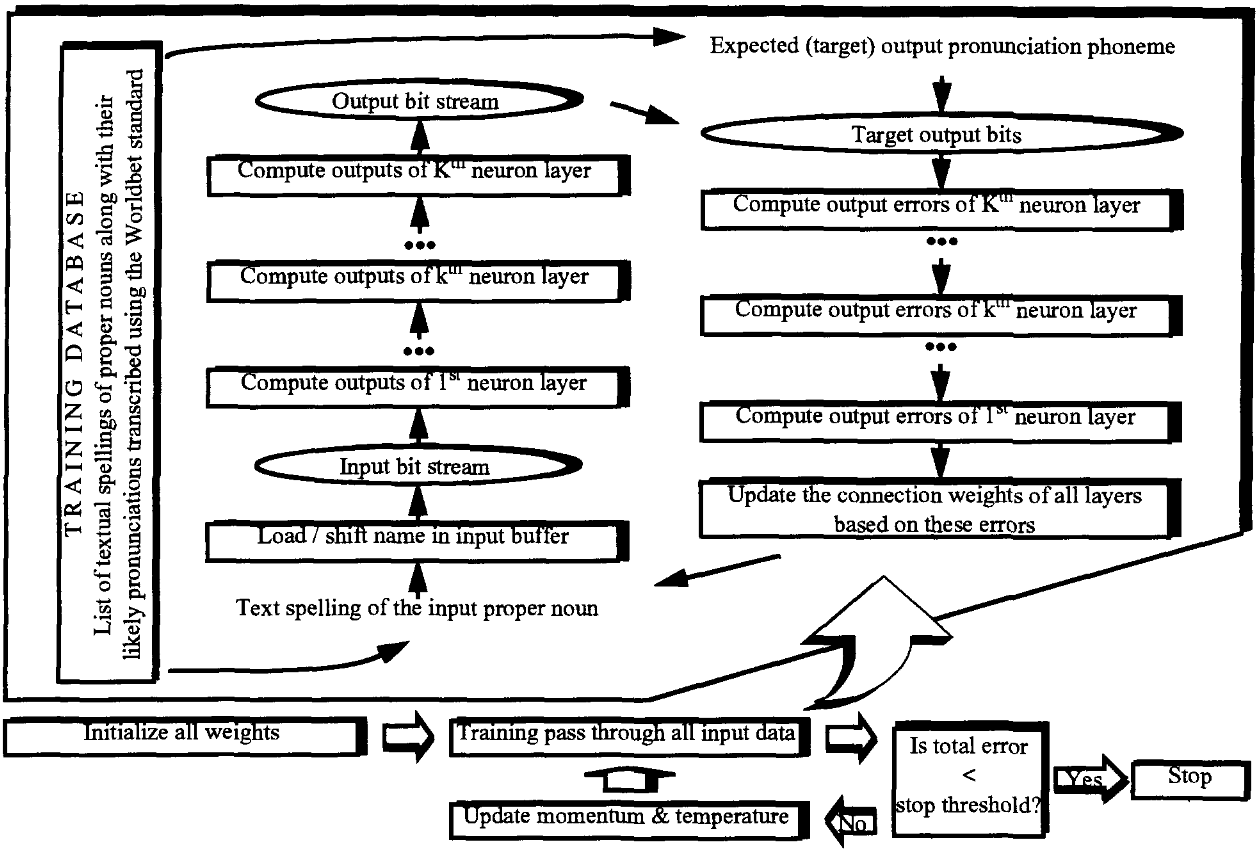
TRAINING DATABASE
List of textual spellings of proper nouns along with their likely pronunciations transcribed using the Worldbet standard
Expected (target) output pronunciation phoneme
Output bit stream
Target output bits
Compute outputs of Kth neuron layer
Compute output errors of Kth neuron layer
•••
Compute outputs of kth neuron layer
Compute output errors of kth neuron layer
•••
Compute outputs of 1st neuron layer
Compute output errors of 1st neuron layer
Input bit stream
Update the connection weights of all layers based on these errors
Load / shift name in input buffer
Text spelling of the input proper noun
Initialize all weights
Training pass through all input data
Is total error < stop threshold?
Yes
Stop
No
Update momentum & temperature

Figure 1. A schematic overview of the simulated annealing used to train the Boltzmann machine neural network. The backpropagation of error during the training pass is displayed in detail in the inset.

2.3. Generation

Once trained, the system can probabilistically perform pattern completion tasks. For a successfully trained network, the set of outputs produced for a given input represents the probability distribution of the pronunciations for the given name in the domain represented by the network. The advanced system generates a specified number of the most likely pronunciations for the input proper noun and outputs a likelihood score for each of them. It is also equipped with the ability to create a phoneme network graph that represents the generation of the different pronunciations.

3. PRONUNCIATION DICTIONARY

A comprehensive development database is an essential part of technology development in speech research. In order to reasonably model the statistics of letter-to-sound conversion it is crucial to train the system on a large dataset representative of the problem. No current databases are amenable to generate multiple proper noun pronunciations. Even though there exists some anecdotal data on this problem, it is not available in any public domain database.

For this task we have developed a comprehensive public domain pronunciation dictionary of people's last names (surnames). It presently consists of 18494 surnames from a diversity of ethnic origins and 25648 corresponding pronunciations. Benchmarking the Boltzmann machine with this database has provided us with significant insights into the problems associated with this task.

3.1. Collection And Transcription

Construction of a representative database of surnames presents some peculiar problems. The distribution of surnames is quite skewed — the 2000 most common surnames account for about 15% of the American population, while the remaining 85% population covers almost 1.5 million surnames. The data, collected from a variety of sources, represents a reasonable mix of commonly found surnames, surnames with infrequent occurrence, and surnames that are known to present problems for letter-to-sound conversion due to complex morphology or difficult stress assignments [7].

The phonetic transcription was performed by hand using the Worldbet standards. Each surname was transcribed to obtain all the correct pronunciations possible. Transcription of name pronunciations was a difficult task as the surnames derive from dozens of source languages having different stress patterns. A number of foreign names have both ethnic as well as anglicized pronunciations and individual pronunciations are often peculiar in defying any kind of typical text-to-speech rules.

3.2. Phone Alignment With Spelling

The Boltzmann machine network is designed to look at each letter of the input name spelling one by one in context of its nearest neighbors, and output a phoneme symbol in accordance. Thus for training purposes it is necessary that there be a phoneme corresponding to every letter in the input spelling. Since in many cases a single phoneme encompasses a small group of letters instead of a single letter, a straightforward alignment is not possible.

To produce a one-to-one alignment of the spellings with the corresponding phonetic expression we have introduced the concept of a *blank phoneme* denoted by the symbol "_". Alignment is performed automatically by a dynamic programming algorithm that introduces such blank phonemes at appropriate places. A phoneme corresponding to a group of letters is aligned with one of the letters according to a strategy that maximizes the total letter-phone alignment score for the entire word. The remaining letters are aligned with blank phonemes. (e.g. The surname 'Wright' is transcribed and aligned as '_ 9r aI _ _ t').

4. EXPERIMENTS AND RESULTS

The performance of the Boltzmann machine system was evaluated on the basis of the number of accurate multiple pronunciations generated. We conducted some pilot experiments to gauge the effect of various parameters on the system performance and used the inferences obtained here in evaluations with extended data sets.

4.1. Character-String Classification

The basic functionality of the system was explored using a simple classification problem involving arbitrary character strings constituting linearly separable classes. The system was tested on 300 held-out strings. The classification performance is displayed in Table 1. It became evident that the context length plays a crucial role in accurate classification and is somewhat related to the length of the input string. The number of hidden layers required can be offset to some extent by choosing a large number of neurons in a smaller number of hidden layers.

Context length	# hidden neurons	# training strings	%error
3	300	1000	11.72
4	300	1000	8.47
4	300	2000	7.47
5	300	1000	9.14
5	500	1000	7.76
5	1000	1000	3.77

Table 1: Character string classification performance as a function of system parameters

Context length	# hidden neurons	% correct pronunciations		
		all	some	none
3	125	5.29	50.93	45.77
4	125	5.29	50.93	43.78
7	300	12.25	54.23	33.51
3 & 7	125 each	17.18	59.71	23.11

Table 2: Performance with 4-letter names

4.2. Fixed-length Proper Nouns

We devised a pilot experiment that consisted of proper noun data with text spellings of a fixed length 4. The training set comprised of 1331 surnames and 1622 pronunciations, while the test set had 334 surnames. Table 2 summarizes some important results for this evaluation.

No significant gain in performance was observed for a system using multiple hidden layers compared to that using only a single internal layer. The number of hidden layer neurons has an optimal value at which the performance is maximized, and the error rate increases if the number deviates from this ideal value. This number varies with the type of input data and is therefore difficult to predetermine analytically. Shorter context lengths were found to be more susceptible to changes in the number of hidden neurons.

The most optimal annealing schedule was found to be an initial temperature of 100 and a temperature decay rate of 0.1 per training iteration. Typically about 60 iterations through the entire training set were found to be necessary and/or sufficient for reasonable performance. The rate of generation of pronunciations was found to be loosely bound to the number of neurons and the context length.

4.3. Full Data Set

Evaluation of the Boltzmann machine on the full data set

Context length	# hidden neurons	% correct pronunciations		
		all	some	none
3	100	22.78	6.78	70.44
3	200	29.33	6.72	63.95
3	500	6.07	1.37	92.56
7	100	23.46	1.37	75.17
7	300	15.65	3.20	81.14
3 & 7	1000 each	3.63	2.98	93.39

Table 3: Performance with full data

(15000 names for training, 3494 for test) produced discouraging results illustrating the problems in scaling up the network to handle more complicated tasks such as variable-length inputs. A performance summary is listed in Table 3. Increasing the network size (number of hidden layer neurons) did not alleviate the performance.

4.4. Multilayered Perceptron Network

To investigate the possible causes for the poor performance we implemented an MLP to output only the single most likely pronunciation. Results obtained for the MLP are displayed in Table 4. It was discovered that at times the network found it difficult to duplicate certain letter-to-phone maps that were observed during training for similar contextual patterns in the test data.

We infer that irrespective of the network architecture, in the current training strategy the network finds it difficult to adapt to certain conflicting constraints represented in the training data. This indicates the need for alternative training strategies that will allow incorporation of contrasting features that correspond to similar input patterns and more representative training data.

5. CONCLUSIONS

# hidden neurons	% correct pronunciations	
	closed loop	open loop
300	80.48	23.28
500	77.76	33.09

Table 4: Performance with 4-letter names on MLP

The Boltzmann machine has the potential to generate an N-best list of possible pronunciations by analyzing the spelling of the proper noun. However, our experiments with various architectures and different lexical domains indicate that the basic neural network itself fails to effectively learn the letter-to-phone distribution in the training data. Thus, contrary to expectations, learning techniques like backpropagation that are prevalently assumed to work for such systems are found to be incapable of modeling the dynamics of this problem. The inherently nonlinear and often conflicting nature of the relevant features highlights the necessity of more powerful training paradigms. Our future research will be focused towards developing more effective training algorithms to address these problems.

We expect the pronunciation dictionary to be a useful resource to the speech research community in fuelling further research in proper noun recognition. The complete database, as well as the system software are available in the public domain at *http://www.isip.msstate.edu/software/.*

6. ACKNOWLEDGMENTS

We are grateful to the Systems and Information Sciences Laboratory of Texas Instruments, especially Dr. Jack Godfrey for his support, insights into the problem and guidance in the development of the training database.

REFERENCES

1. J. Picone, B.J. Wheatley and J. McDaniel, “On the Intelligibility of Text-To-Speech Synthesis of Surnames,” Texas Instruments Technical Report No. CSC-TR-91-002, pp. 1-34, Texas Instruments Inc., Dallas, TX, March 13, 1991.

2 N. Deshmukh, M. Weber and J. Picone, “Automated Generation of N-best Pronunciations of Proper Nouns”, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal processing,* pp. I283-I286, Atlanta GA, May 1996.

3. G. Hinton and T. Sejnowski, “Optimal Perceptual Inference”, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 448-453, Washington D.C., June 1986.

4. T.J. Sejnowski and C.R. Rosenberg, ““NETtalk: A Parallel Network That Learns To Read Aloud,” Tech. Rep. JHU/EECS-86/01, John Hopkins University, Baltimore, MD, 1986.

5. S. Kirkpatrick, C.D. Gellatt and M.P. Vecchi, “Optimization by Simulated Annealing”, in *Science*, Vol. 220, pp. 671-680, 1983.

6. M.L. Minsky and S. Papert, *Perceptrons*, MIT Press, Cambridge, MA, 1969.

7. E.C. Smith, *American Surnames*, Genealogical Publishing Co. Inc., Baltimore, MD, 1986.