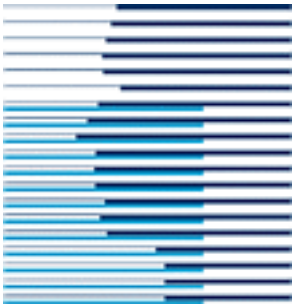
Rapid Retraining on Speech Data with LSTM Recurrent Networks

Alex Graves, Nicole Beringer, Juergen Schmidhuber



Technical Report No. IDSIA-09-05
May 2, 2005

IDSIA / USI-SUPSI
Instituto Dalle Molle di studi sull' intelligenza artificiale
Galleria 2
CH-6900 Manno, Switzerland

Rapid Retraining on Speech Data with LSTM Recurrent Networks

Alex Graves, Nicole Beringer, Juergen Schmidhuber

May 2, 2005

Abstract

A system that could be quickly retrained on different corpora would be of great benefit to speech recognition. Recurrent Neural Networks (RNNs) are able to transfer knowledge by simply storing and then retraining their weights. In this report, we partition the TIDIGITS database into utterances spoken by men, women, boys and girls, and successively retrain a Long Short Term Memory (LSTM) RNN on them. We find that the network rapidly adapts to new subsets of the data, and achieves greater accuracy than when trained on them from scratch. This would be useful for applications requiring either cross corpus adaptation or continually expanding datasets.

1 Introduction

In speech recognition, there are several reasons to consider retraining systems incrementally: large new corpora are costly and time-consuming to create, so it would be preferable to make use of those already in existence; corpora tend to be tailored towards particular tasks, so a system capable of transferring knowledge between them would be more flexible than one that isn't; and should greater specialisation be needed (e.g. towards a particular speaker) such a system could provide it quickly and efficiently. Two areas in particular that would benefit from retraining are speaker adaptation (see [8] for a Hidden Markov Model based attempt) and cross language learning (see e.g. [1, 7]).

Our experiments investigate the potential for incremental retraining on speech data with LSTM recurrent networks. We will successively retrain an LSTM net on disjoint subsets of the TIDIGITS speech corpus [6], containing utterances spoken by men, women, boys, and girls. We will show that LSTM is capable of rapidly adapting to new speaker groups, and of preserving knowledge across multiple retrainings.

The structure of the paper is as follows. In Section 2 we provide a brief overview of the LSTM algorithm and list some of the tasks to which it has successfully been applied. In section 3 we provide experimental results and analysis while conclusions and future work are presented in Section 4.

2 The LSTM Architecture

LSTM is an RNN - first presented in [5] and later extended in [4] - that contains self-connected internal *memory cells* protected by nonlinear multiplicative gates. Unlike other RNN training algorithms, error is back-propagated through the network in such a way that exponential decay is avoided. The unbounded (i.e. unsquashed) cells are used by the network to store information

over long time durations. The gates are used to control the flow of information through the internal states. Each *memory block* has an *input gate* that allows it to selectively ignore incoming activations, an *output gate* that allows a it to selectively take itself offline, shielding it from error, and a *forget gate* that allows the cells to selectively empty their memory contents. One memory block can contain several memory cells. The gates have their own activations in the range $[0, 1]$. See Figure 1.

LSTM's learning algorithm is local in space and time with computational complexity per timestep and weight of $O(1)$ for standard topologies. This locality, in contrast with training algorithms such as Real Time Recurrent Learning [11] and Back Propagation Through Time [12], makes LSTM more biologically plausible than most RNN architectures. Indeed, a recent report by O'Reilly [9] describes a closely related model of working memory in the basal ganglia and prefrontal cortex.

Below are some of the difficult time series problems to which LSTM has successfully been applied. These tasks demonstrated LSTM's facility with timewarped data and long time lags, a key motivation for its application to speech recognition.

1. Recognition of temporally extended patterns in noisy input sequences
2. Recognition of regular, context free and context sensitive languages
3. Recognition of the temporal order of widely separated events in noisy input streams
4. Extraction of information conveyed by the temporal distance between events; stable generation of precisely timed rhythms
5. Robust storage of high-precision real numbers across extended time intervals
6. Reinforcement learning in partially observable environments
7. Music improvisation and music composition

See [3] for the papers containing the above experiments. Further details on the extended LSTM architecture (including full pseudocode) can be found in [4].

3 Experiments

The data used in this paper were taken from the TIDIGITS speech corpus, collected by Texas Instruments from over 300 adults and children. Each utterance was a single spoken digit, from "zero" (or "oh") to "ten" and the task of the network was to correctly identify that digit. For our experiments, the utterances were partitioned into those spoken by men, women, boys and girls. All audio files were preprocessed into mel-frequency cepstrum (MFCC) coefficients, using the HTK toolkit [13] with the following parameters: 12 cepstral coefficients, 1 energy coefficient, and 13 first derivatives, giving 26 coefficients in total. The frame size was 15 ms and the input window was 25 ms. The TIDIGITS corpus comes pre-divided into training and test sets, and all results quoted below were recorded on the test set.

The aim of our investigations was not to demonstrate that LSTM can correctly identify spoken digits (although this was an essential precondition). The aim was show that LSTM can rapidly adapt to significantly different data - in this case the different speech characteristics of males and females, adults and children. The two criteria we were particularly interested in were the speed and accuracy of adaptation, and the retention of previous learning across multiple retrainings. To this end we successively trained an identical network on numerous different sequences of the four data subsets, including some with repetitions (e.g. men then girls then women then men). The training was incremental in the sense that we did not reset the weights between datasets.

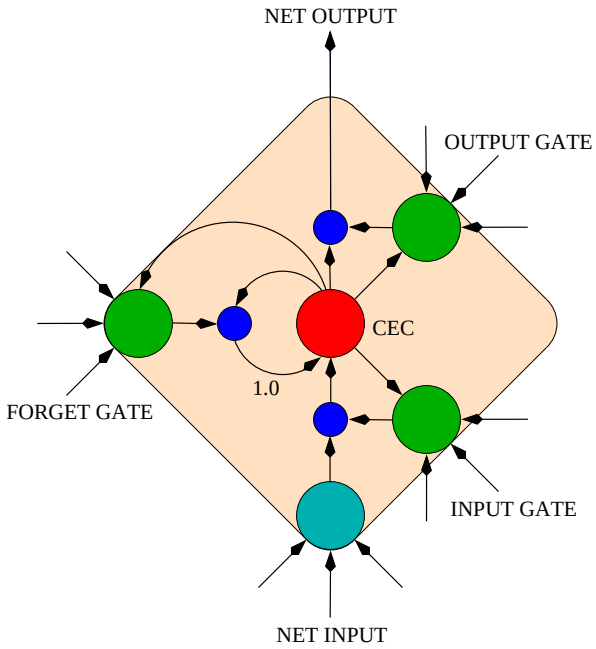


Figure 1: LSTM memory block with one cell. The internal state of the cell is maintained with a recurrent connection of weight 1.0. The three gates collect activations from both inside and outside the block, and control the cell via multiplicative units (small circles). The input and output gates effectively scale the input and output of the cell while the forget gate scales the internal state—for example by resetting it to 0 (making it forget).

3.1 Experimental Setup

We used a neural net with a mix of LSTM and sigmoidal units (with a range of $[0, 1]$). The net had 26 inputs (one for each MFCC coefficient) and 11 sigmoidal output units - one for each possible digit. The classification was based on the most active output layer at the end of an input sequence (i.e. on the last timestep of each spoken digit). A cross-entropy objective function [2] was used at the output layer. The network also had two hidden layers. The first of these was an extended LSTM layer with forget gates and peephole connections (see section 2 for details). The layer contained 20 memory blocks, each with two cells, and therefore contained 100 nodes in total (including gates). The squashing function was logistic with range $[-2, 2]$, and the activation functions of the gates were logistic in range $[0, 1]$. The bias weights to the LSTM forget (input and output) gates were initialised blockwise with positive (negative) values of $+0.5$ (-0.5) for the first block, $+1.5$ (-1.5) for the second block and so on. The second hidden layer consisted of 11 sigmoidal units.

The connection scheme was as follows: all units were biased, except for the input units. The input layer was fully connected to the LSTM layer. The LSTM layer was fully connected to itself, the hidden sigmoidal layer, and the output layer (note that the LSTM layer had only outputs from its cells, and not from its gates). The second hidden layer was fully connected to the output layer. In total there were 121 units (excluding inputs) and 7791 weights.

The learning rate was 10^{-6} and online learning was used, with weight updates at every timestep. The momentum algorithm from [10] was used with a value of 0.9, and the network activations were reset to zero after each pattern presentation. These parameters were experimentally determined, although we have not deviated from them significantly in any of our LSTM speech experiments. Errors were fed back on every timestep, encouraging the net to make correct decisions as early as possible (a useful property for real time applications). Gaussian noise (with mean zero and deviation 1.5) was injected into the training data to prevent overfitting.

3.2 Results

For the following tables, we first trained the network for 1000 epochs on each of the TIDIGITS subsets: men, women, boys and girls (see Section 3). The initial training yielded the following correctness scores (correctness defined as the fraction of correctly identified digits in the whole test set): men 0.997, women 0.991, boys 0.973, girls 0.980. The correctness attained on all four subsets together was 0.993.

In tables 1 and 2, "Previous Subsets" are the subsets on which we have previously trained the net, in the order we presented them. For example "girls, women, men" means the net was trained from scratch on the girls subset, then retrained on women, then retrained on men. "Current Subset" is the subset the net is currently being retrained on. "Initial Score" was the correctness score of the net on the new data set *before* any retraining took place. "Final Score" was the correctness after exactly 200 epochs of retraining. And "Retraining Epochs" was the number of epochs it took the net to equal or better the final score.

3.3 Analysis

3.3.1 Simple Retraining

The results in Table 1 demonstrate that LSTM is capable of rapidly and accurately retraining on speech data with widely varying vocal characteristics. In most cases it actually achieved a higher correctness on the retrained data than it had when trained on that data from scratch; moreover it was usually able to do so in less than 50 epochs. A look at the correctness curves in figure 3

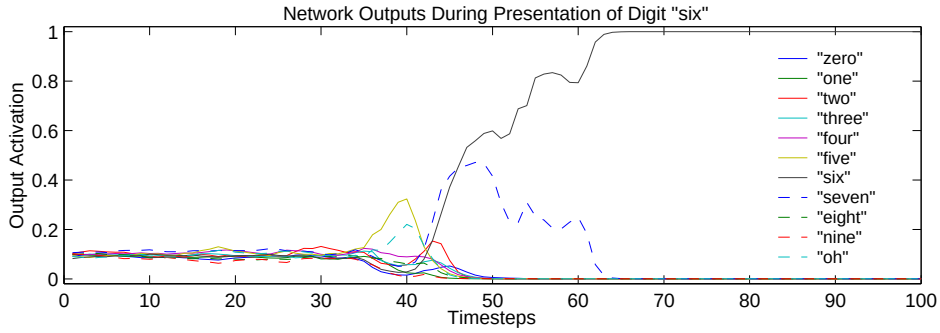


Figure 2: The network correctly identifying the spoken digit six. Early on the output for seven is also active, reflecting the fact that both words begin with a sibilant.

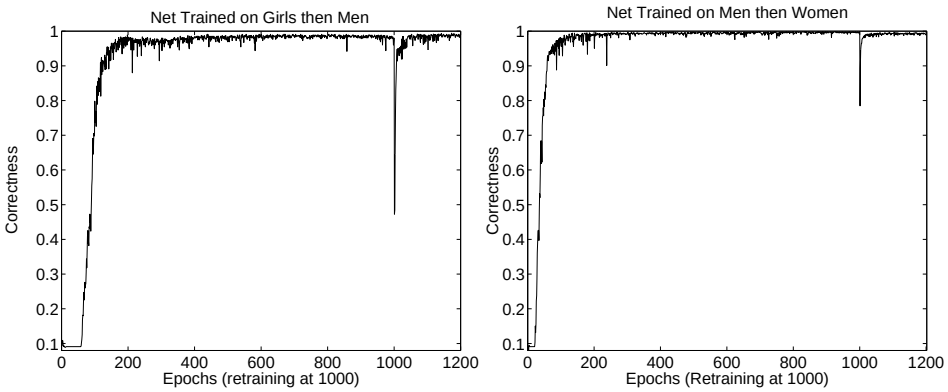


Figure 3: Correctness curves of the network during two retraining sequences. Note the relatively high performance before retraining begins (demonstrating direct transfer of knowledge between the datasets) and the steeper learning curve during retraining than training). Switching from girls to men (left graph) is more difficult than from men to women (right graph) due to the greater change in vocal characteristics.

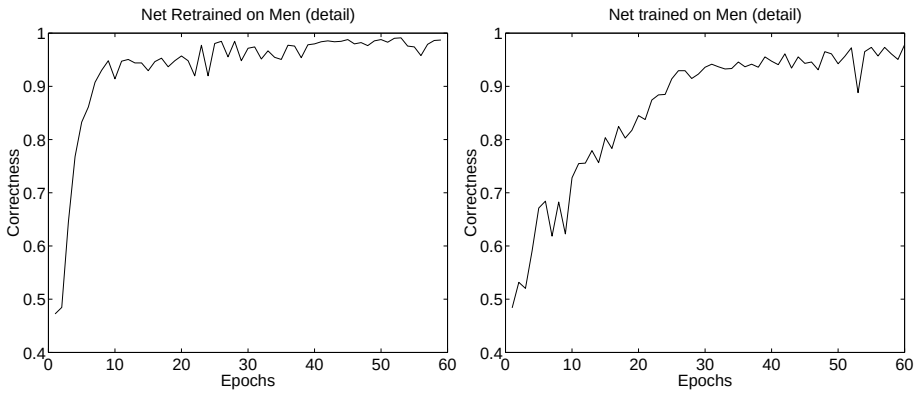


Figure 4: Details from above figure showing the greater speed and accuracy of the network when retraining on men (left) than when training from scratch (right).

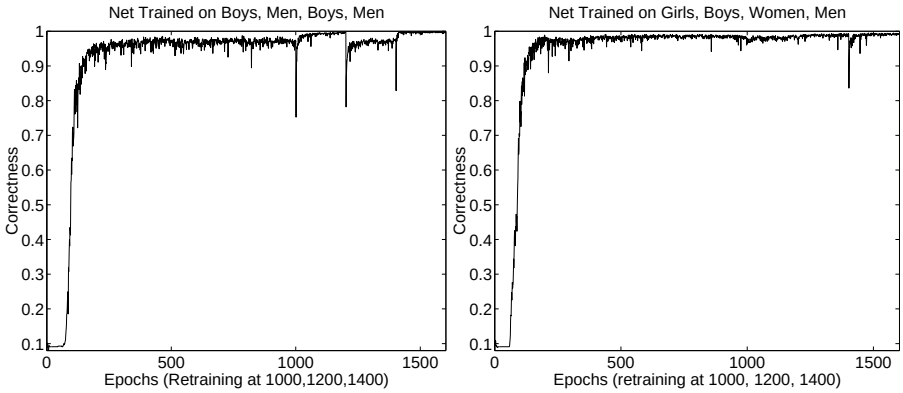


Figure 5: Examples of multiple retraining. In the left graph the second transition from boys to men (after 1400 epochs) is substantially faster than the first (1000 epochs), indicating that the net has retained some knowledge of the men subset. In the right graph (where the whole corpus is traversed in an order chosen to minimize the differences between adjacent subsets) all the retraining times are small.

Table 1: Retraining on TIDIGITS subsets

Previous Subsets	Current Subset	Initial Score	Retraining Epochs	Final Score
men	women	0.785	16	0.990
men	boys	0.713	200	0.991
men	girls	0.496	70	0.982
women	men	0.826	17	0.996
women	boys	0.962	2	0.975
women	girls	0.929	17	0.985
boys	men	0.752	196	0.998
boys	women	0.970	24	0.994
boys	girls	0.985	2	0.991
girls	men	0.472	51	0.989
girls	women	0.920	32	0.990
girls	boys	0.964	16	0.980

reveals not only that the net begins to improve immediately after the new data is presented, but that it does so more rapidly than when training from scratch (see figure 4).

Clearly some speaker groups are easier to move between than others. For example, switching from women to boys, or from boys to girls, gave very little initial error and required virtually no retraining. The transitions between men and children on the other hand, were more time consuming (although still much faster than training from scratch).

3.3.2 Multiple Retraining

See Table 2. The point of the multiple retrainings was to show that the retraining time and difficulty diminished with repetition, and that the net was able to transfer knowledge across several datasets. In addition, we were interested to see how the order in which the datasets were presented affected its retraining ability.

Comparing the net trained on men, then girls, then men, to that trained only on girls then men, we can see that the initial correctness, retraining time, and final correctness were all better for the former. With some exceptions, this was typical of the sequences that contained repetitions of a given dataset: retraining was easier on data that the net has seen before (cf. figure 5, left graph). In other words knowledge of the data was transferred across the intervening retraining. As can be seen by comparing figure 5, right graph, with figure 3, left graph, the order in which the data was presented was also significant: moving from girls to men via boys and women was much smoother than doing so directly.

4 Conclusions and Future Work

We sequentially retrained an LSTM network on disjoint subsets of the TIDIGITS speech corpus, containing utterances from men, women, boys and girls. We found that the net was capable of rapidly and accurately adapting to previously unseen data with very different speech characteristics. We also found that its final performance was generally raised by having been previously trained on different datasets, and that this improvement persisted over multiple retrainings.

In the future we would like to concentrate on more challenging tasks in multi-corpus learning. Two areas in particular we would like to explore are cross language speech recognition (exploiting, for example, phonetic or syllabic similarities between languages), and speaker adaptation - which

Table 2: Multiple Retraining on TIDIGITS subsets

Previous Subsets	Current Subset	Initial Score	Retraining Epochs	Final Score
men, girls	men	0.544	30	0.998
boys, men	boys	0.782	99	0.978
men, women	men	0.887	33	0.999
girls, men	boys	0.773	61	0.989
men, women	boys	0.976	4	0.989
boys, men	girls	0.642	105	0.991
girls, boys	women	0.963	63	0.992
women, girls	men	0.561	36	0.994
girls, men, boys	women	0.982	7	0.994
boys, men, girls	men	0.652	33	0.997
men, boys, girls	men	0.615	43	0.998
girls, women, men	girls	0.593	38	0.985
women, girls, men	girls	0.653	45	0.984
men, girls, men	girls	0.576	58	0.982
boys, men, boys	men	0.829	11	0.995
girls, boys, women	men	0.836	77	0.995

plays a vital part in off-the-shelf dictation software. These will form part of our wider effort to build effective speech recognition systems based on LSTM networks.

References

[1] Beringer, N. and Schiel, F. (2000). The quality of multilingual automatic segmentation using german maus. In *Proc. of the International Conference on Spoken Language Processing, Beijing, China*.

[2] Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Inc.

[3] et al, J. S. http://www.idsia.ch/ juergen/rnn.html.

[4] Gers, F. (2001). *Long Short-Term Memory in Recurrent Neural Networks*. PhD thesis.

[5] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

[6] Leonard, R. G. (1984). A database for speaker-independent digit recognition. In *Proc. of the Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3.

[7] Metze, F., Kemp, T., Schaaf, T., Schultz, T., and Soltau, H. (2000). Confidence measure based language identification. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. The Institute of Electrical and Electronics Engineers (IEEE), Signal Processing Society*, pages 1827–1830.

[8] Neto, J. P., Martins, C. A., and Almeida, L. B. (1996). An incremental speaker-adaptation technique for hybrid -MLP recognizer. In *Proc. ICSLP '96*, volume 3, pages 1293–1296, Philadelphia, PA.

[9] O'Reilly, R. (2003). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. Technical Report ICS-03-03, ICS.

[10] Plaut, D. C., Nowlan, S. J., and Hinton, G. E. (1986). Experiments on learning back propagation. Technical Report CMU–CS–86–126, Carnegie–Mellon University, Pittsburgh, PA.

[11] Robinson, A. J. and Fallside, F. (1987). The utility driven dynamic error propagation network. Technical Report CUED/F-INFENG/TR.1, Cambridge University Engineering Department.

[12] Williams, R. J. and Zipser, D. (1990). Gradient-based learning algorithms for recurrent connectionist networks. In Chauvin, Y. and Rumelhart, D. E., editors, *Backpropagation: Theory, Architectures, and Applications*. Erlbaum, Hillsdale, NJ.

[13] Young, S. (1995/1996). *The HTK Book*. Cambridge University.