Classifying unprompted speech by retraining LSTM Nets

Nicole Beringer, Alex Graves, Jürgen Schmidhuber
IDSIA
Galleria 2
6928 Manno-Lugano
Switzerland

Florian Schiel
Schiel BAS Services
Schellingstr. 3
80799 Munich
Germany

Technical Report No. IDSIA-07-05
March 2005

IDSIA / USI-SUPSI
Dalle Molle Institute for Artificial Intelligence
Galleria 2, 6928 Manno, Switzerland

Classifying unprompted speech by retraining LSTM Nets

Nicole Beringer, Alex Graves, Jürgen Schmidhuber
IDSIA
Galleria 2
6928 Manno-Lugano
Switzerland

Florian Schiel
Schiel BAS Services
Schellingstr. 3
80799 Munich
Germany

March 2005

Abstract

We apply Long Short-Term Memory (LSTM) recurrent neural networks to a large corpus of unprompted speech- the German part of the VERBMOBIL corpus. Training first on a fraction of the data, then retraining on another fraction, both reduces time costs and significantly improves recognition rates. For comparison we show recognition rates of Hidden Markov Models (HMMs) on the same corpus, and provide a promising extrapolation for HMM-LSTM hybrids.

1 Introduction

It would be desirable to retrain an Automatic Speech Recognition (ASR) system on new data without losing the benefits of previous learning. For example, it may be necessary to adapt quickly to new input, or to use information gained from a previous task, e.g., recognize read speech, in order to solve the next task, e.g., quasi-spontaneous (= unprompted) speech. In task/domain independent recognition [15], systems that are (pre-)trained under certain conditions and/or certain dialogue specifications are required to adapt to utterances recorded under different conditions or with different dialogue specifications. It has also become standard practice to train Hidden Markov Models (HMMs) on multiple corpora, in order to improve their robustness also with respect to new data. However, methods for adapting HMM's are complex, unintuitive, and time-consuming [11]. Most modern systems use a hybrid of HMMs and maximum likelihood linear regression to adapt to new training material.

An artificial neural net (ANN) lends itself to a very simple form of retraining: train on one dataset, then continue training on another. One type of ANNs - Recurrent Neural Nets (RNNs) - is particularly promising for speech processing because it has the potential to learn a dynamic model of speech that incorporates multiple time scales without using time windows or fixed time delays. Unlike traditional RNNs, Long Short-Term Memory nets (LSTM) [10] can also handle long time lag correlations between inputs and errors, also in the context of speech applications

[8]. Recent experiments with plain LSTM on speaker adaptation [7] suggest that retraining is fast and effective on **small** corpora, and that results of previous learning and generalization improve with retaining on randomly chosen subsets of the data. In this paper we apply this approach for the first time to Bidirectional LSTM [9] and a **large** corpus of unprompted speech.

The following section gives an overview of LSTM. Section 3 briefly describes the VERBMOBIL data used for both LSTM and HMM experiments. Section 4 describes the experimental setup. Section 5 analyses the experimental results of baseline and retrained LSTM for framewise phoneme prediction and gives results for the entire phonemes for plain HMMs on the same test set to show the task difficulty (unprompted speech). Section 6 provides an extrapolation of the framebased results for a HMM-LSTM hybrid (under development) based on previous comparisons of framewise and phoneme error rates on various corpora of read speech.

2 LSTM

"Long Short-Term Memory" [10, 6] is a general purpose algorithm for extracting statistical regularities from noisy time series. It learns from scratch, typically with more adjustable parameters (the weights), a larger search space, and less initial bias [5] than HMMs, which incorporate prior linguistic knowledge.

2.1 Bidirectional LSTM (BLSTM)

The output of typical RNNs is based on the complete history of *previous* inputs. However, there are many sequence processing tasks where future inputs are also useful because reverse correlations exist. In speech, for example, the articulatory system is already preparing future utterances as it shapes the current one. A solution is bidirectional training [13, 1, 2]: the input is presented forwards and backwards to two separate recurrent nets, both of which are connected to the same output layer. In this way, errors can be injected as normal and backpropagated through the nets. Current results with BLSTM [9] show that it outperforms normal LSTM, as well as support vector machines and previous bidirectional RNNs on speech recognition tasks.

2.2 Retraining with bidirectional LSTM

An in-depth investigation of retraining with LSTM [7] (i.e. presenting new data to an already trained network) showed that LSTM is capable of fast and effective relearning on speakers with widely varying vocal characteristics. The net was trained and successively retrained on disjoint subsets of the TIDIGITS database. The retraining time and difficulty diminished with repetition, and the net was able to transfer knowledge across several datasets. The final performance of the net was generally raised by having been previously trained on different datasets, and this improvement persisted over multiple retrainings.

3 Corpus description

Our present investigation uses a database of *unprompted* speech- the VERBMOBIL (VM) corpus [17], which is more difficult to recognize than read speech such as the TIMIT corpus. The VM corpus is divided into VM1 (recordings up to 1996) and VM2 (recording after 1996). Both sets differ in recording conditions and tasks. The corpus consists mainly of three language portions: German, American English and Japanese. The German VM portion contains sufficient speech data for training and testing (35136 turns[1]). For this study only the German portion was used. The

[1] One turn in the VM database has about 22.8 words in average.

database-scenario deals with scheduling appointments with a business partner: real-life-situations with currently used speech. The "formal situation" setup ensures that speech contains fewer and weaker regional variants than it would contain if personal affairs were discussed.

The training (TRAIN), development (DEV) and test (TEST) sets currently used in our experiments on the VM corpus were created with the following constraints (see table 1 for exact numbers): each speaker is allowed in only one set (hard constraint), for each speaker there must be at least one complete dialogue (to allow speaker adaptation algorithms to be applied; hard constraint), speakers should be distributed equally across sexes in all sets (soft constraint), recordings should be distributed equally across recording sites in all sets (to cover possible accents preferences in one site; soft constraint),

Table 1: Basic numbers for the subsets in VM1 and VM2

	VM1_DEV	VM1_TEST	VM1_TRAIN	VM2_DEV	VM2_TEST	VM2_TRAIN
WORDS	15084	14615	285280	11905	9855	153438
TURNS	630	631	12600	592	592	11835
LEX	1537	1342	6472	1397	1264	5238
SPEAKER	35	33	629	13	13	119

The HMM system uses the full data for training and testing. The LSTM classification network uses only one fourth of the training set in its baseline training, another fourth of the training set is used for retraining. The full test set is used as described above.

4 Experimental setup

4.1 HMM system for evaluating the task difficulty

The HMM[12] phone recognizer was built up with the Hidden Markov Toolkit [18]. It uses the above defined subsets and a bigram trained solely on the training corpus ($VM1\_TRAIN + VM2\_TRAIN$). It was tested on the $VM1\_DEV + VM2\_DEV$ sets with the corresponding lexicon (total: 5540 lexical entries).

The acoustic models are based on 12 Standard MFCC + Energy + velocity + acceleration (39), Diagonal covariance matrices, 3-5 states per phoneme, 43 phoneme classes (extended German SAMPA) + garbage + voice garbage + silence + laugh + breath (48), Models initialized using the Munich Automatic Segmentation (MAU) tier of the BAS Partiture Format (BPF) from 1/4 of TRAIN, Re-estimation and splitting mixtures after 6 iterations on total TRAIN, testing after every two iterations on DEV, weight of language model fixed to 6.5; beam search width 100.0.

4.2 BLSTM: experimental setup

Preliminary experiments with LSTM standard nets with 25, 50, 100 and 200 blocks (2 cells each) showed that although the duration of the epochs doubled each time, comparable results occurred in far fewer epochs. Nevertheless all experiments converged at around 50% framewise phoneme correctness. When comparing LSTM bidirectional nets to standard nets with comparable weights (50 000) we found that BLSTM needs less epochs to obtain comparable results to standard nets and reaches higher framewise phoneme correctness (58.87%). Both bidirectional and standard nets reach their peak around the 120th epoch.

Based on these findings we used a two-step retraining procedure as follows: LSTM training and retraining sets were each around 1/4 of the whole VM training set. Both training and retraining set are distinct from each other but were randomly chosen from the whole training set. The whole VM test set was used.

Our bidirectional LSTM network contained two hidden LSTM layers (for the forward and reverse nets), each with 200 blocks of 2 cells. It had 26 input nodes and a softmax output layer containing 52 nodes. A cross entropy objective function was used. The input layer was connected to the hidden layers, both of which were connected to themselves and to the output layer. There were 907112 weights in total. Note that unlike HMMs BLSTM has no structural bias and more weights - a disadvantage according to the bias-variance dilemma [5].

5 Experimental results

Our experiments are divided into two main parts: The first shows the recognition results of a plain HMM phone recognizer which was trained both on monophones and triphones (also across words). Part two gives the plain LSTM classification for frame by frame recognition results.

Table 2: For comparison: Phoneme error rate for plain HMMs

System	training set size	phoneme error rate on the test set	epochs
Monophone	full	34.29%	52
Triphone crossword	full	35.49%	37

Monophones contain 512 Gaussian mixtures per state. Triphones have the same number of parameters as the monophone system, 8 mixtures per state and are trained also across word boundaries. HMMs were trained on the full training set (BLSTM just on one fourth, retraining on another fourth). Both systems use the same test set. Table 3 shows the main results of the plain BLSTM net.

Table 3: Recognition results: frame by frame phoneme error rate for plain BLSTM

System	training set size	frame by frame phoneme error rate on the test set	epochs
baseline	1/4 (randomly chosen)	38.40%	50
retraining	1/4(distinct from baseline)	33.36%	67

BLSTM retraining led to a 5% improvement on the full test set. Using 1/4 of the training set at a time greatly reduces total training time.

6 Predicting the phoneme error rate: an extrapolation for a HMM-LSTM hybrid approach

Although we cannot compare the framewise phoneme error of BLSTM directly with the phoneme error of the HMM we expect that a BLSTM-HMM hybrid (under construction) will outperform both plain BLSTM on frame by frame and plain HMMs on the phoneme level, inheriting the best of both worlds, namely reduction of training material and training time (BLSTM), as well as more built-in structural bias (HMMs).This expectation is encouraged by experiments on **read speech** by Chen and Jamieson [3], Shire [14], Waterhouse, Kershaw and Robinson [16], and Elenius and Blomberg [4]. They all achieved better results on the phoneme level using an ANN-HMM hybrid approach, as shown in table 4 for framewise and phoneme error rates for several systems on various corpora. *improvement factor* shows the relative ratio of framewise and phoneme error. *LIN* stands for Linear Input Network, *MLIN* for Mixtures of LINs for adaptation (with *_2* = 2 experts; *_4* = 4 experts). *MLP* stands for Multilayer Perceptron nets.

Table 4: Framewise and phoneme errors on read speech corpora

System	corpus	frame	phoneme (ANN-	improvement
		(plain ANNs)	HMM hybrids)	factor
Backprop [5]	Swedish speakers	30.0%	24.5%	1.22
RNN 0 pass [15]	MUM[2] Task	22.8%	18.1%	1.26
LIN 1 pass [15]	MUM Task	20.1%	16.5%	1.22
LIN 2 pass [15]	MUM Task	19.9%	15.9%	1.25
MLIN_2 1 pass [15]	MUM Task	19.2%	16.5%	1.16
MLIN_2 2 pass [15]	MUM Task	18.9%	16.1%	1.17
MLIN_4 2 pass [15]	MUM Task	18.2%	15.8%	1.15
MLIN_4 3 pass [15]	MUM Task	18.0%	15.7%	1.15
MLP [14]	clean speech[4]	28.97%	7.3%	3.97
MLP [14]	clean sp. no border[4]	29.80%	7.7%	3.87
MLP [14]	factory noise[3]	42.84%	15.5%	2.76
MLP [14]	factory noise no border[4]	42.88%	15.0%	2.86
RNN [3]	TIMIT	26.3%	20.21%	1.30

As can be seen from table 4 the framewise errors are quite high for noisy input sequences (several microphones or enriched with background noise) as opposed to clean speech. The HMM part of the hybrids is able to drastically reduce the error on the phoneme level due to structural bias of the HMM. This means that on unprompted speech with background noise, speaker overlaps and other perturbations we can expect a much lower phoneme error.

With the **worst** improvement factor (1.15) of table 4 we can conservatively predict a phoneme error rate of 29.01% for a retrained BLSTM-HMM hybrid on VERBMOBIL (33.39% for the standard BLSTM respectively). An optimistic calculation with the **best** improvement factor (3.97) for read speech in table 4 would give us 8.4% for the retrained BLSTM-HMM hybrid (9.67% for the baseline respectively). Of course, to figure out the precise improvement we really have to implement a BLSTM-HMM hybrid.

7 Conclusions and outlook

We examined the retraining ability of LSTM recurrent nets in a frame by frame phoneme classification task of unprompted speech. We compared recognition results of a normally trained BLSTM system to those of a retrained one. Retraining both significantly reduced both time costs

[2] ARPA 1995 H3 multiple unknown microphones

[3] NUMBERS95

and training set size and improved recognition results. An extrapolation based on previous work on read speech [16, 3, 14, 4] promises significant additional improvements on the phoneme level through a BLSTM-HMM hybrid, which we are currently implementing.

8 Acknowledgements

This work was supported by the SNF under grant number 200020-100249.

References

[1] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G.L. Pollastri. Exploiting the past and the future in protein secondary structure prediction. *BIOINF: Bioinformatics*, 15, 1999.

[2] J. Chen and N. S. Chaudhari. Capturing long-term dependencies for protein secondary structure prediction. In *Advances in Neural Networks - ISNN*, Lecture Notes in Computer Science. Springer, 2004.

[3] R. Chen and L. Jamieson. Experiments on the impementation of recurrent neural networks for speech phone recognition. *Proc. Thirtieth Annual Asilomar Conference on Signals, Systems and Computers*, pages 779–782, 1996.

[4] K. Elenius and M. Blomberg. Comparing phoneme and feature based speech recognition using artificial neural networks. *Proc. ICSLP*, 1992.

[5] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 1992.

[6] F. A. Gers and J. Schmidhuber. Long Short-Term Memory learns simple context free and context sensitive languages. *Proc. IEEE TNN*, 2001.

[7] A. Graves, N. Beringer, and J. Schmidhuber. Rapid retraining on speech data with lstm recurrent networks. Technical Report IDSIA-05-05, IDSIA, www.idsia.ch/techrep.html, 2005.

[8] A. Graves, D. Eck, N. Beringer, and J. Schmidhuber. Biologically plausible speech recognition with LSTM neural nets. *Proc. Bio-ADIT*, 2004.

[9] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm networks. In *IJCNN, under review*, 2005.

[10] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 1997.

[11] John McDonough and A. Waibel. Performance comparisons of all-pass transform adaption with maximum likelihood linear regression. *Proc. ICSLP*, 2004.

[12] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. 77(2):257–286, 1989.

[13] M. Schuster and K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 1997.

[14] M. Shire. Relating frame accuracy with word error in hybrid ann-hmm asr. *Proc. EUROSPEECH*, 2001.

[15] W. Wahlster. SmartKom: Symmetric multimodality in an adaptive and reusable dialogue shell. *Krahl, R., Günther, D. (eds): Proceedings of the Human Computer Interaction Status Conference*, 2003.

[16] S. Waterhouse, D. Kershaw, and T. Robinson. Smoothed local adaptation of connectionist systems. *Proc. ICSLP*, 1996.

[17] K. Weilhammer, F. Schiel, and U. Reichel. Multi-Tier annotations in the Verbmobil corpus. *Proc. LREC*, 2002.

[18] S. Young. *The HTK Book*. Cambridge University Press, 1995.