Phoneme recognition in TIMIT with BLSTM-CTC

Santiago Fernández, Alex Graves, Jürgen Schmidhuber

Technical Report No. IDSIA-04-08
April 15, 2008

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

IDSIA is a joint institute of both University of Lugano (USI) and University of Applied Sciences of Southern Switzerland (SUPSI), and was founded in 1988 by the Dalle Molle Foundation which promoted quality of life.

Phoneme recognition in TIMIT with BLSTM-CTC

Santiago Fernández, Alex Graves, Jürgen Schmidhuber [*]

April 15, 2008

Abstract

We compare the performance of a recurrent neural network with the best results published so far on phoneme recognition in the TIMIT database. These published results have been obtained with a combination of classifiers. However, in this paper we apply a single recurrent neural network to the same task. Our recurrent neural network attains an error rate of 24.6%. This result is not significantly different from that obtained by the other best methods, but they rely on a combination of classifiers for achieving comparable performance.

1 Introduction

Spontaneous speech production is a continuous and dynamic process. This continuity is reflected in the acoustics of speech sounds and, in particular, in the transitions from one speech sound to another. As a consequence, the boundaries between speech sounds are not clearly defined. This fact significantly contributes to making segmentation and labelling of speech data interrelated tasks. Because of this interrelation, automatic speech recognition is best performed with methods such as hidden Markov models (HMM) that do not require segmented data for development. On the contrary, developing neural networks has traditionally relied on segmented data. The objective functions require a network output target value at every or specific time-steps in the data sequence. Connectionist temporal classification (CTC) overcomes this limitation. CTC allows developing neural network classifiers using a sequence of labels as the desired output target [5]. Labels correspond to events occurring in the input data sequence, such as phones in a speech data stream. The number of labels in a target labelling is, therefore, typically much shorter than the number of time-steps in the input data sequence. Also, there is not timing information in a target labelling, except for labels being in the same order in which events occur in the input data sequence.

Recurrent neural networks are an interesting alternative to HMMs for speech recognition. Their continuous internal state is naturally well suited for modelling speech dynamics. Moreover, their capability to model data dependencies has potential for modelling coarticulatory

[*] Santiago Fernández and Jürgen Schmidhuber are with IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland (email: [santiago, juergen]@idsia.ch). Alex Graves and Jürgen Schmidhuber are with TU Munich, Boltzmannstr. 3, D-85748 Garching, Münich, Germany (email: [graves,juergen.schmidhuber]@in.tum.de).

effects in speech. In contrast, HMMs are built on a number of independence assumptions about the data.

We showed in [5] that CTC-based recurrent neural networks outperform state-of-the-art algorithms on phoneme recognition in the TIMIT database. In contrast with the algorithms compared in [5], which rely on a single type of classifier to perform the task, Glass' uses a committee-based classifier [4], whereas Deng *et al.*'s combines the scores from two related algorithms [1]. These two systems achieved the best phoneme recognition rates published so far for TIMIT. In this paper, we compare the performance of a single CTC-based recurrent neural network with that of Glass' and Deng *et al.*'s systems. The main differences with respect to the experimental setup used in [5] are: first, the data are divided into training, validation and test sets as described in [8], and second, a standard set of 39 phonetic categories, instead of 61, is used [10]. This new experimental setup allows a direct comparison of the three systems.

2 Materials

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) contains recordings of prompted English speech accompanied by manually segmented phonetic transcripts [2]. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States.

For the experiments, the SA sentences were discarded and the remaining data were split into a training set, a validation set and a test set according to [8]. The training set contains 3696 sentences (462 speakers), the validation set contains 400 sentences (50 speakers) and the test set contains 192 sentences (24 speakers).

TIMIT transcriptions are based on 61 phones. Typically, 48 phones are selected for modelling. Confusions among a number of these 48 phones are not counted as errors. Therefore, results are presented for 39 phonetic categories. We decided to train the network on transcriptions based on this lexicon of 39 phones. The 61 categories were folded onto 39 categories as described by Lee and Hon [10]. This is shown in table 1.

Speech data was transformed into Mel frequency cepstral coefficients (MFCC) with the HTK software package [11]. Spectral analysis was carried out with a 40 channel Mel filter bank from 64 Hz to 8 kHz. A pre-emphasis coefficient of 0.97 was used to correct spectral tilt. Twelve MFCC plus the 0th order coefficient were computed on Hamming windows 25 ms long, every 10 ms. Delta and Acceleration coefficients were added giving a vector of 39 coefficients in total. For the network, the coefficients were normalised to have mean zero and standard deviation one over the training set.

The division of TIMIT into the three aforementioned data sets and the presentation of results for 39 phones, was also adopted in [4, 1]. As acoustic features, Deng *et al.* used frequency-warped LPC cepstra [1], instead of MFCC. For his part, Glass tried a number of variations and combinations of MFCC, perceptual linear prediction (PLP) cepstral coefficients, energy and duration [8, 4]. Glass' system built 61 models, one for each of the 61 phones in TIMIT, and results were tabulated using the standard set of 39 phones.

aa	aa, ao
ah	ah, ax, ax-h
er	er, axr
hh	hh, hv
ih	ih, ix
l	l, el
m	m, em
n	n, en, nx
ng	ng, eng
sh	sh, zh
sil	pcl, tcl, kcl, bcl, dcl, gcl, h#, pau, epi
uw	uw, ux
—	q

Table 1: Folding the 61 categories in TIMIT onto 39 categories (from [10]). The phones in the right column are folded onto their corresponding category in the left column (the phone 'q' is discarded). All other TIMIT phones are left intact.

3 Method

The method employed is the same described in [5]. Briefly, phoneme recognition is performed with a recurrent neural network. The long short-term memory recurrent neural network (LSTM) was used because of its ability to bridge long time delays [9, 3]. The hidden units in an LSTM network are called memory blocks. Each memory block has one or more memory cells controlled by an input, an output and a forget gate. When the input gate is open incoming data is stored in the memory cell, and when the output gate is open data stored in the memory cell is sent to the output layer. The forget gate resets the memory cell. Gates can optionally have access to the data stored in the memory cell (*peephole* connections). Gates and the memory block input are typically connected to the same units in the network. These connections are trainable, thus the behaviour of the gates is not pre-determined, but rather learned during training.

For phoneme recognition, where both anticipatory and carry-over coarticulatory effects are important, a bi-directional neural network is suitable. The bi-directional LSTM (BLSTM) [7, 6] has two separate recurrent hidden layers, both of them connected to the same input and output layers. The *forward* recurrent network is presented with sequential data forward in time, from the beginning of the data sequence to time-step t. The *backward* recurrent network is presented with sequential data backwards in time, from the end of the data sequence to time-step t. At any time-step t, the network has access to all information in the data sequence.

The BLSTM recurrent neural network was trained with the CTC algorithm using the list of phones in the speech utterances as target labellings [5]. Once the network has been trained, the predicted labelling for a new speech utterance can be directly read from its outputs. This method (best path decoding) is, however, not guaranteed to find the most probable labelling. A second method (prefix search decoding) consists in calculating the probabilities of successive extensions of labelling prefixes, which can then be used to find the most probable labelling. However, because this procedure is computationally intensive, it was separately calculated for sections of the output sequence. As a consequence, prefix search decoding is not guaranteed to find the most probable labelling but, in practice, it generally outperforms best path decoding [5].

In the experiments reported in this paper, the BLSTM-CTC network had an input layer of size 39, the forward and backward hidden layers had 128 blocks each, and the output layer was size 40 (39 phones plus blank). The gates used a logistic sigmoid function in the range $[0, 1]$. The input layer was fully connected to the hidden layer and the hidden layer was fully connected to itself and the output layer. The total number of weights was 183,080.

Training of the BLSTM-CTC network was done by gradient descent with weight updates after every training example. In all cases, the learning rate was 10^{-4}, momentum was 0.9, weights were initialized randomly in the range $[-0.1, 0.1]$ and, during training, Gaussian noise with a standard deviation of 0.6 was added to the inputs to improve generalisation. For prefix search decoding, an activation threshold of 0.9999 was used (see [5] for a description of this parameter).

Performance was measured as the normalised edit distance (label error rate; LER) between the target label sequence and the output label sequence given by the system.

Deng *et al.*'s hidden trajectory models (HTM) are a type of probabilistic generative model aimed at modelling speech dynamics and adding long-contextual-span capabilities that are missing in hidden Markov models (HMM) [1]. A thorough description of this system is available in [12]. HTM uses a bi-directional filter to estimate probabilistic speech data trajectories given a hypothesized phone sequence. This estimate is then used to compute the model likelihood score for the observed speech data. The search for the phone sequence with the highest likelihood is performed with an A* based lattice search and rescoring algorithm specifically developed for HTM.

Glass's system is a segment-based speech recogniser (as opposed to frame-based recognisers) based on the detection of *landmarks* in the speech signal [4]. Acoustic features are computed over hypothesized segments and at their boundaries. The standard decoding framework is modified and extended to deal with this paradigm shift.

4 Results

Results are shown in table 2. Error rates include errors due to substitutions, insertions and deletions with respect to the reference transcription. Deng *et al.*'s best result was achieved with a lattice-constrained A* search with weighted HTM, HMM, and language model scores [1]. Glass's best results were achieved with many heterogeneous information sources and classifier combinations [4]. A single BLSTM-CTC recurrent neural network attains an error rate of 24.6%, which is not significantly different from Deng *et al.*'s or Glass's best results. It is likely that BLSTM-CTC can achieve improved performance when more sources of information are added and when they are combined with other classifiers. The results shown in table 2 are the best results reported in the literature on phoneme recognition in TIMIT.

28.57%	Deng *et al.*'s baseline HMM [1]
25.17%, s.e. 0.20%	BLSTM-CTC (best path decoding)
24.93%	Deng *et al.*'s HTM-HMM [1]
24.93%	Deng *et al.*'s HTM-HMM [1]
24.58%, s.e. 0.20%	BLSTM-CTC (prefix search decoding)
24.4%	Glass's committee-based classifier [4]

Table 2: Error rates on TIMIT. Results for BLSTM-CTC are the average and standard error (s.e.) over 10 runs. On average, the networks were trained for 112.5 epochs (s.e. = 6.4). The horizontal lines divide the list of systems into groups performing significantly different than the networks. BLSTM-CTC with best path decoding is significantly different from Deng *et al.*'s baseline HMM (two-sided t-test, $p < 3 \cdot 10^{-8}$), from BLSTM-CTC with prefix search decoding ($p < 0.05$) and from Glass's classifier ($p < 0.004$). BLSTM-CTC with prefix search decoding is not significantly different from either Deng *et al.*'s HTM-HMM or Glass's classifier.

5 Conclusions

We have provided results for phoneme recognition with BLSTM-CTC using the TIMIT database. The experiments use the same standard data sets and phonetic inventory employed by the systems reportedly having the best performance to date. Finally, we have compared BLSTM-CTC's performance to that achieved by these systems [4, 1]. BLSTM-CTC achieves comparable performance without relying on a combination of multiple classifiers. Also, BLSTM-CTC makes fewer assumptions about the task domain.

Acknowledgments

The authors would like to thank James R. Glass for providing the list of speakers for each data set. This research was funded by SNF grant 200021-111968/1.

References

[1] Li Deng, Dong Yu, and Alex Acero. A generative modeling framework for structured hidden speech dynamics. In *Proceedings of NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*, Whistler, BC, Canada, December 2005.

[2] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgrena, and Victor Zue. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, 1993.

[3] Felix A. Gers, Nicol Schraudolph, and Jürgen Schmidhuber. Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*, 3:115–143, 2002.

[4] James R. Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 17:137–152, 2003.

[5] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, USA, June 2006.

[6] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Bidirectional LSTM networks for improved phoneme classification and recognition. In *Proceedings of the 2005 International Conference on Artificial Neural Networks*, Warsaw, Poland, 2005.

[7] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6):602–610, June/July 2005.

[8] Andrew K. Halberstadt. *Heterogeneous acoustic measurements and multiple classifiers for speech recognition.* PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1998.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[10] Kai-Fu Lee and Hsiao-Wuen Hon. Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648, November 1989.

[11] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK Book.* Cambridge University Engineering Department, HTK version 3.4 edition, December 2006.

[12] Dong Yu, Li Deng, and Alex Acero. A lattice search technique for a long-contextual-span hidden trajectory model of speech. *Speech Communication*, 48:1214–1226, 2006.