Multitask Learning in Connectionist Robust ASR using Recurrent Neural Networks

Shahla Parveen and Phil Green

Department of Computer Science, University of Sheffield, Sheffield S14DP, UK
s.parveen@dcs.shef.ac.uk, p.green@dcs.shef.ac.uk

Abstract

The use of prior knowledge in machine learning techniques has been proved to give better generalisation performance for unseen data. However, this idea has not been investigated so far for robust ASR. Training several related tasks simultaneously is also called multitask learning (MTL): the extra tasks effectively incorporate prior knowledge. In this work we present an application of MTL in robust ASR. We have used an RNN architecture to integrate classification and enhancement of noisy speech in an MTL framework. Enhancement is used as an extra task to get higher recognition performance on unseen data. We report our results on an isolated word recognition task. The reduction in error rate relative to multicondition training with HMMs for subway, babble, car and exhibition noises was 53.37%, 21.99%, 37.01% and 44.13% respectively.

1. Introduction

Automatic Speech Recognition in the presence of additive background noise is a challenging task because of the mismatch between the acoustic models and incoming data caused by the noise [1]. Conventional techniques for improving recognition robustness (reviewed by Furui [2]) seek to eliminate or reduce the mismatch, for instance by enhancement of the noisy speech, by adapting statistical models for speech units to the noise condition or simply by training in different noise conditions. Success with these techniques has been moderate compared to human performance (see for instance the sessions on Noise Robust Recognition in Eurospeech 2001). Psychological studies show that humans employ prior knowledge for efficient learning and can also learn several related tasks simultaneously [3]. The drawback of traditional machine learning techniques including ASR is that they are based on training set learning and lack the characteristics of human learning, resulting in non-optimal generalisation performance. Several benefits can be achieved (e.g. better generalisation and reduced learning time) by incorporating prior knowledge in the learning process. In this work we will exploit prior knowledge in a robust ASR task using multitask learning. Our goal is to achieve high performance on both classification and enhancement tasks.

2. Prior knowledge in learning

It is evident from psychological studies that humans can learn even from a single example. The reason is that human learning is not just a result of exposure to random examples. Rather, humans benefit from built in mechanisms (e.g. pain) and teachers (such as family or the school) [4]. Human beings face each new learning task equipped with knowledge gained from previous learning tasks [5]. The use of prior knowledge in machine learning is an active area of research for several reasons including faster learning, better generalisation error and training with insufficient data. There has been active interest in Artificial Neural Network techniques for using prior knowledge in learning [6], [7], [8], [9]. The prior knowledge in ANNs can be incorporated in a number of ways e.g., by programming net weights [6] by creating virtual examples [10], [8] by using extra output targets or catalyst [11], or by task specific changes in the learning algorithm [12]. Since parallel learning of multiple tasks explored by Caruana [7] is at the forefront of learning methods we have adapted this idea for robust ASR.

2.1. Multitask learning

One of the key aspects of the learning problem faced by humans, which differs from the vast majority of problems studied in the field of neural network learning, is the fact that humans encounter a whole stream of learning problems over their entire lifetime [3]. Humans take advantage of the opportunity of comparing and contrasting similar categories in learning to classify a new example within these categories [5].

Neural networks and other learning algorithms e.g. decision trees have difficulty in learning if given a single, isolated and difficult task. Hinton [13] proposed that if networks are trained to represent underlying regularities of the domain, generalisation of the network will be better. Use of extra targets associated with additional tasks (figure 1), also known as adding *catalyst* output units, is an interesting way to incorporate prior knowledge [11]. This idea is also referred to as MultiTask Learning (MTL) [7, 14] and provides the learning algorithm with a better chance of capturing the domain regularities. The argument behind this is that sharing the information among the tasks learned can help those tasks together more efficiently and easily than in isolation [14].

Figure 1 shows an MTL feedforward multilayer network with a hidden layer and an output associated with each task. These outputs are fully connected to the hidden layer. Input-to-hidden links represent task domain common components whereas hidden-to-output links are associated with task specific components [15]. The hidden layer of this net is shared by all tasks. This sharing of a common hidden layer makes the internal representation developed in the hidden layer available to all tasks. This is the central idea in MTL: to share the

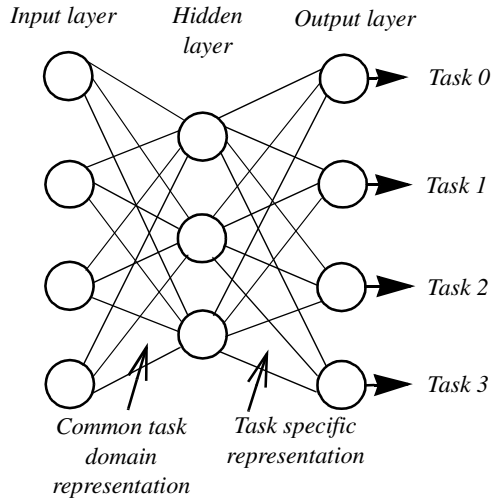

Figure 1: An MTL network with one main task and three extra tasks. There is an output unit for each task being learned in parallel

learned information while the tasks are learned in parallel. Compared to this, when each task is learned in isolation, there is no sharing of information among the tasks.

MTL uses the idea to create extra tasks that get trained on the same net with the main task. It is a form of inductive transfer that improves performance on the main task by using the information contained in the training signals of other related tasks. The main purpose of extra tasks is to help in the learning of the main task. However, when extra tasks are also required to be optimised there are several ways to achieve this goal e.g. by using a separate learning rate for each task, separate output error weighting or having a private hidden layer for the major task [14].

3. Robust ASR and prior knowledge

Human listeners have the capability to recognise speech subject to many types of variability. Apart from recognition based on a small utterance, our brain can do several other tasks simultaneously without great effort:

- we can deal with drastic degradation in speech caused by additive or multiplicative noise,
- at the same time, we know the gender of the speaker,
- we can discriminate between the pitch of different speakers and use this information to identify them,
- we can differentiate between the accent and dialects of the speakers,
- we can analyse emotions and intention,

Since state-of-the-art robust ASR systems are trained on a single task they don't benefit from the domain specific prior knowledge contained in the training data. The performance of robust ASR systems can further be boosted by employing available sources of additional and prior knowledge.

Use of prior knowledge in the area of word segmentation has been reported by [11]. Niyogi [10] discussed the use of prior knowledge in speech recognition in the form of 'virtual examples', which are basically a transformed version of existing training examples and help to increase the size of training data.

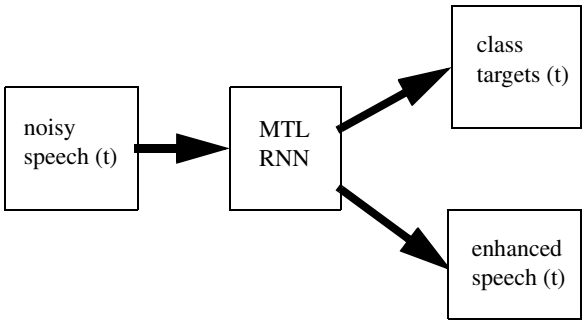

Figure 2: Speech enhancement and classification with a single net

4. Recurrent Neural Nets for robust ASR using MTL

Prevailing robust ASR techniques are based on Continuous Density Hidden Markov Models. Here, we consider a connectionist alternative. One motivation is that CDHMMs are generative models which do not give direct estimates of posterior probabilities of the classes given the acoustics. Neural Networks, unlike HMMs, are discriminative models which do give direct estimates of posterior probabilities and have been used with success in hybrid ANN/HMM speech recognition systems [16].

Recurrent Neural Networks have the potential to capture long-term contextual effects over time, which CDHMM based techniques do not do: the estimated likelihood of the data at a particular time in a CDHMM system depends on the observed acoustics and the state distribution. RNNs also allow a single net to perform both enhancement and classification, with the potential of combining these processes to mutual benefit. Previous work on robust ASR using RNN has been reported by [17].

The use of RNNs in robust ASR to train several related tasks simultaneously has not been addressed so far. In this work we present an application of MTL in robust ASR. We have used RNN architecture for the problem of integrating classification and enhancement of noisy speech in an MTL framework. Enhancement is used as an extra task to get higher recognition performance on the unseen data.

4.1. RNN architecture

Figure 2 shows the block diagram for robust ASR using a multitask RNN. The RNN is supplied with noisy speech at time *t* and produces class posteriors and enhanced speech for the same time. The RNN basically has an Elman architecture [18], where there are fully connected recurrent links from the past hidden layer to the present hidden layer. The number of input units depends on the size of feature vector, i.e. the number of spectral channels (32 channels in the experiments reported). The number of hidden units is determined by experimentation (120 in our experiments). There are output

units for each pattern class and extra units for enhancement. In our case the classes are taken to be whole words, so in the isolated digit recognition experiments we report, there are eleven output units, for '1' - '9', 'zero' and 'oh', with an additional 32 output units corresponding to the length of feature vector.

RNN weights are updated using back-propagation through time [19]. The average of the RNN output error over all the time frames of a training utterance is taken after these frames had gone through a forward pass. This error is used to mediate weight change for the training example. The error for both output classes and the enhanced features is estimated as the sum squared error between the correct targets (one of n for the classification units and the clean values for the enhancement units) and the RNN output for each frame.

The recognition phase consists of a forward pass to produce RNN output for unseen data and enhancement at each time step. The highest value in the averaged output vector is taken as the recognised class.

5. Experimental setup and database

Experiments were performed using data from male speakers in the isolated digits section of the AURORA database [20]. This database contains about 1200 isolated digits from 55 male speakers, where each speaker spoke 2 examples of the 11 word vocabulary (the digits 1-9, 'oh' and 'zero'). All speech data in the Aurora database is in turn obtained from the TIDigit database after downsampling to 8 KHz and filtering with a G712 characteristic.

1000 examples were chosen for training. A validation set of 110 examples was used to control the stopping condition in training. Recognition performance was evaluated on the isolated digit section of Aurora test set A which has four types of noises (subway, babble, car and exhibition) added at SNRs from 20 dB to -5 dB at 5 dB intervals.

Acoustic vectors were obtained from a 32 channel auditory filter bank [21] with centre frequencies spaced linearly in ERB-rate from 50 to 3750 Hz. The instantaneous Hilbert envelope at the output of each filter was smoothed with a first order filter with an 8 ms time constant, and sampled at a frame rate of 10 ms. Finally a cube root compression was applied to the frame of energy values. Spectral domain acoustic vectors are used to allow comparison with our earlier work with missing data techniques [17].

5.1. Classification performance

The classification performance of our combined MTL classification+enhancement net ('MTRNN:CLASS+ENH' curve) is shown in figure 3 for subway noise in Aurora test A. We compare our results with:

1. CDHMM systems trained both on clean isolated digits ('HMM:NP' curve) and noisy isolated digits ('MCHMM' curve). These systems consisted of eleven whole word HMMs ('1' - '9', 'oh', 'zero'), each with 16 states and 2 mixtures per state.
2. A CDHMM system using marginalisation based missing data recognition ('HMM:MARG' curve) [22], [23].

We can see improvement in results for this noise and results for

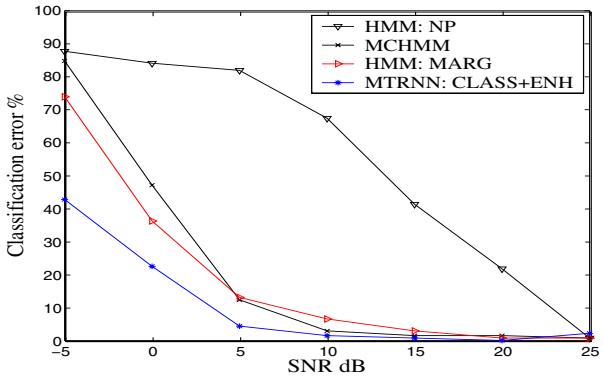

Figure 3: RNN performance on classification

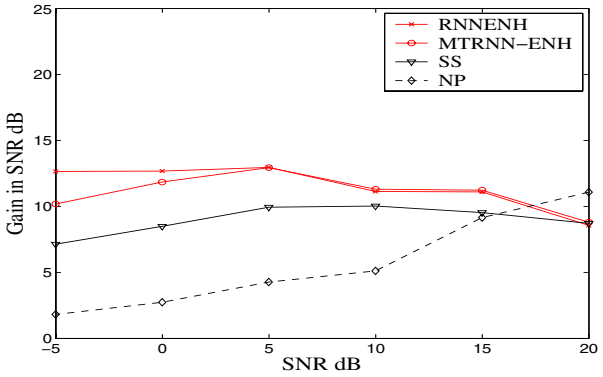

Figure 4: RNN performance on pattern completion

other noises in set A were similar. The overall error from 'MCHMM' (i.e. for all conditions including clean speech and noisy speech at SNRs from -5 dB to 20 dB) for subway, babble, car and exhibition noises was 21.53%, 23.91%, 28.88% and 28.29% respectively. The incremental reduction in word error rate (WER) compared to 'MCHMM' for subway, babble, car and exhibition noises was 11.49%, 5.26%, 10.69% and 12.49% (53.37%, 21.99%, 37.01% and 44.13% relative WER) respectively. It seems that enhancement as a hint to classification for babble noise does not give as good an improvement as for other noises. The reason may be the non-stationarity of babble noise which prevents the RNN from searching the optimal solution for the enhancement task.

5.2. Pattern completion performance

Figure 4 compares speech enhancement results obtained form a separate single task learning (STL) RNN ('RNNENH' curve) and MTL RNN ('MTRNN-ENH' curve) as a measure of gain in SNR for speech added with subway noise in Aurora test set A. The results from both nets were superior to spectral subtraction ('SS' curve) and noisy condition ('NP' curve). We can see the effect of a slightly lower enhancement performance from the MTL net than the STL enhancement net. Similar results were obtained for babble, car and exhibition noises in Aurora test A. However the difference in speech enhancement performance is not significant. The enhancement task itself in STL needed about 30 hidden units. In the MTL net the problem might lie in selecting a suitable hidden layer representation

from a huge number of hidden layer representations provided by a large hidden layer. Enhancement performance may be improved by defining schedules in the training to emphasise on the minimisation of enhancement error.

6. Conclusion and future work

In this paper we have demonstrated the effectiveness of MTL in robust ASR to get improved recognition accuracy. We used the idea of MTL as it is more suitable for our problem and has also been proven efficient than sequential learning with prior knowledge.

Enhancement is not the only hint which can be used to achieve higher classification performance. We believe that speech contains a lot of information apart from the class labels for training e.g. gender, speaker identity, accent, word boundary and off course the information about sources of distortions. These extra sources of information can be used as hints for classification task in MTL framework.

We are currently working on extending this recognition system for the connected digits recognition task, following the Aurora standard for robust ASR. In this case we introduce 'silence' as an additional recognition class, and the training targets are obtained by forced-alignment on clean speech with an existing recogniser. We also plan to investigate integration of other useful sources of prior knowledge in robust ASR, which may result in better recognition performance.

7. Acknowledgement

This work is being supported by Nokia Mobile Phones, Denmark and the UK Overseas Research Studentship scheme.

8. References

[1] Lippmann, R. P. (1997). "Speech recognition by machines and humans". *Speech Communication*, vol. 22, no. 1, p. 1-15.

[2] Furui, S. (1997). "Recent advances in robust speech recognition". *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, France, p. 11-20.

[3] Thrun, S. (1996). "Is Learning The n-th Thing Any Easier Than Learning The First?" *Advances in Neural Information Processing Systems*.

[4] Giraud-Carrier, C.G. (1994). "On Integrating Inductive Learning with Prior Knowledge and Reasoning". *PhD Thesis, Dept. of Computer Science. Brigham Young University.*

[5] Ben-David, S. and Schuller, R. (2002). "Exploiting Task Relatedness for Multiple Task Learning". *NIPS 2002*.

[6] Omlin, C. W. and Giles, C. L. (1992). "Training Second-Order Recurrent Neural Networks using Hints". *Proceedings of the Ninth International Conference on Machine Learning*.

[7] Caruana, R. (1993). "Multitask Learning: A Knowledge-Based Source of Inductive Bias". *Proceedings of the 10th International Conference on Machine Learning*, p. 41-48.

[8] Abu-Mostafa, Y. S. (1995). "Hints". *Neural Computation*, 7:639-671, July 1995.

[9] Frasconi, P. et al. (1999). "Insertion of Prior Knowledge". *in J.F. Kolen and S.C. Kremer (Editors), A Field Guide to Recurrent Neural Networks*, IEEE Press.

[10] Niyogi, P. et al. (1998). "Incorporating Prior Information in Machine Learning by Creating Virtual Examples". *Proceedings of the IEEE*, vol. 86, no. 11, p. 2196-2209.

[11] Allen, J. et al. (1996). "Integrating multiple cues in word segmentation: A connectionist model using hints". *In Proceedings of the 18th Annual Cognitive Science Society Conference*, p. 370-375. Mahwah, NJ: Lawrence Erlbaum.

[12] Al-Mashouq, K. A. et al. (1991). "Including hints in training neural nets". p. 418-427. *Neural Computation*, vol. 3, no. 3.

[13] Hinton, G. E. (1986). "Learning distributed representations of concepts". *Proc. of the 8th International Conference of the Cognitive Science Society*, p. 1-12.

[14] Caruana, R. (1997). "Multitask Learning". *Machine Learning, PhD Thesis, CMU*.

[15] Silver D.L et al., (1996). "The parallel transfer of task knowledge using dynamic learning rates based on a measure of relatedness". *Connection Science Special Issue*. Vol. 8, No. 2, p. 277-294, Cambridge, MA.

[16] Bourlard, H. and N. Morgan (1998). "Hybrid HMM/ANN systems for speech recognition: Overview and new research directions". *In C. L.Giles and M. Gori (Eds.), Adaptive Processing of Sequences and Data Structures*.

[17] Parveen, S. and Green, P. (2001). "Speech Recognition with Missing data techniques using Recurrent Neural Networks". *Advances in Neural Information Processing Systems 14*, (T.G.Dietterich, S. Becker and Z. Ghahramani eds.), MIT Press.

[18] Elman, J.L. (1990). "Finding structure in time". *Cognitive Science*, vol. 14, p. 179-211.

[19] Werbos. P. J. (1990). "Backpropagation Through Time: What it does and how to do it". *Proceedings of the IEEE*, vol. 78, no. 10, p. 1550-1560.

[20] Pearce, D. and Hirsch, H.G. (2000). "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions". *In Proc. ICSLP 2000*, vol. IV, p. 29-32, Beijing, China.

[21] Cooke, M.P. (1991). "Modelling Auditory Processing and organisation". *PhD thesis, Department of Computer Science, University of Sheffield.*

[22] Barker, J. et al. (2001). "Linking auditory scene analysis and robust ASR by missing data techniques". *Workshop on Innovation in Speech Processing 2001*, Stratford-upon-Avon, UK.

[23] Cooke, M. et al. (2001). "Robust automatic speech recognition with missing and unreliable acoustic data". *Speech Communication*, vol. 34, no. 3, p.267-285.