MULTITASK LEARNING IN CONNECTIONIST SPEECH RECOGNITION

Youyi Lu, Fei Lu, Siddharth Sehgal, Swati Gupta, Jingsheng Du, Chee Hong Tham, Phil Green and Vincent Wan
Speech and Hearing Research Group
Department of Computer Science, University of Sheffield, UK
pdg@dcs.shef.ac.uk

Abstract

Humans do not learn new tasks in isolation: for instance speech recognition interacts with speaker recognition. Multitask Learning (MTL) attempts to take advantage of these interactions by adding additional, related tasks to the main task. In this work, we present applications of MTL to Automatic Speech Recognition in noise, incorporating speech enhancement and gender recognition as additional tasks. A Recurrent Neural Net architecture is used to integrate multiple tasks. Results show that related extra tasks can indeed produce significant improvements to performance on the main task, i.e. ASR.

1. Introduction

ASR performance, for current systems, degrades dramatically when there is a mismatch between the training and testing conditions, for instance due to the presence of other sound sources (Lippmann 1997). However, many potential applications will remain unacceptable to users until ASR is sufficiently robust to cope with sources of variation such as background noise, reverberation and speaker change.

This work explores MultiTask Learning as a technique for improving recogniser robustness. An MTL device is asked to learn several related tasks together rather than in isolation. Parveen and Green (Parveen 2003; Parveen & Green 2003) used MTL (Caruana 1997), implemented by RNNs, to recognise isolated digits in the presence of added noise. Their networks were given an additional task – speech enhancement. ASR performance improved dramatically: the error rate was reduced by almost 50%. In this paper the effectiveness of this technique is further explored. We review the argument for multitask learning and its implementation by RNNs. Different additional tasks are then introduced and experiments are reported with both Spectral and Cepstral domain acoustic vectors.

2. Multi-task Learning

One of the key aspects of human knowledge acquisition is that people encounter a spectrum of learning problems in their lifetime, and that these problems do not appear one at a time. Humans are able to categorise and relate the problems and to take advantage of this organisation when faced with a new problem. In contrast, most work in machine learning approaches each new problem separately, without the benefit of experience (Thrun 1996).

Neural networks and other Machine Learning (ML) techniques have difficulty in learning if given a single, isolated and difficult task. Hinton (Hinton 1986) proposed that if networks are trained to represent underlying regularities of the domain, their generalisation will be better. Use of extra targets associated with additional tasks, also known as adding *catalyst* output units, is an interesting way to incorporate prior knowledge (Allen et al. 1996). In a persuasive thesis Caruana developed this idea, coining the term *MultiTask Learning* (Caruana 1997). The argument is that sharing the information among the tasks learned can help to perform those tasks together more efficiently and more easily than in isolation.

Implementing MTL requires an architecture which involves a shared representation between tasks. This can be provided by the hidden layer of a neural network. In this arrangement, shown in Figure 1, extra units associated with extra tasks are added to the output layer. The extra tasks force the network to learn internal complex representations in order to approximate both the main and extra tasks. In addition, it may be more efficient for the network to learn two or more tasks with this integrated learning rather than trying to learn each function separately (Caruana 1997).

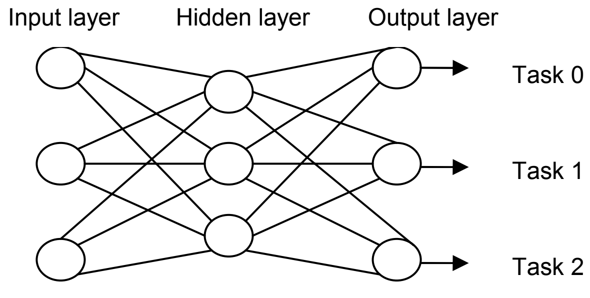

Figure 1: *An MTL network with one main task and two extra task., one output unit for each task.*

Proceedings of the 10th Australian International Conference on Speech Science & Technology

Accepted after full paper review

Figure 1 shows a MTL network with one hidden layer and outputs with each specific task. Three input units are fully connected with the hidden layer. This represents the central idea in MTL: to share the learned information while the tasks are learned in parallel.

3. Recurrent Neural Networks for ASR using MTL

As mentioned above an architecture is required which involves the shared representation between tasks to implement MTL. The RNN is one such approach. In comparison with conventional ASR techniques which are based on Hidden Markov Models, RNNs have the following advantages:

- RNNs are discriminative models which do give direct estimates of posterior probabilities (Thrun 1996),
- RNNs have the potential to capture long-term contextual effects over time (Parveen & Green 2003),

In this paper applications of MTL in robust ASR are presented which apply the RNN architecture to the problem of integrating digit classification, speech enhancement and gender identification. As a sanity check, we also define a 'random task', whose target values are random during training: an unrelated task should not effect results.

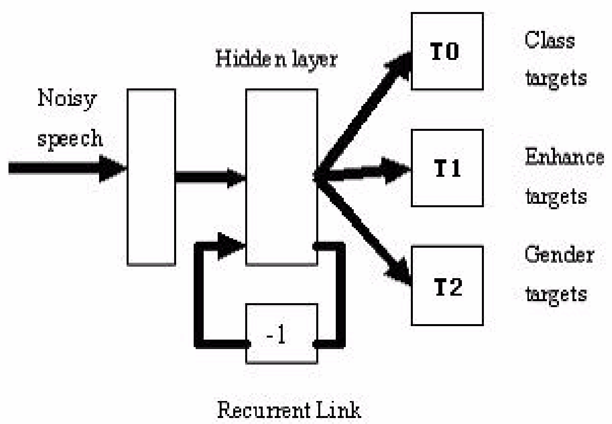

Figure 2: *MTL RNN with one main task and two extra tasks: enhancement plus gender identification*

4. Experimental setup and database

Figure 2 shows the basic architecture of our experiments. It is an Elman RNN (Elman 1990). Noisy speech data are input and class posteriors are produced as well as enhanced speech and gender classification. RNN weights are updated using the back-propagation through time algorithm (Werbos 1990). There are different output unit sets for the different tasks. There are fully connected recurrent links from the past hidden layer to the present hidden layer. The recognition phase uses forward propagation to produce RNN outputs for unseen testing data at each time step. The recognition decision is based on the highest average RNN output over time.

Parveen's work was based on spectral feature vectors: auditory rate maps (Cooke et al. 2001). Here we make comparisons with more conventional MFCC features. The number of input units was 32 for spectral features and 12 for cepstral coefficients. The number of hidden units was 120 for gender dependent experiments and 80 for gender independent experiments.

In Single Task Learning (STL) experiments, there were 11 output units for '1'-'9', 'zero' and 'oh'. In the enhancement task, there were additional output units corresponding to the length of the input vector. In gender recognition, there were 2 additional output units for the two genders. In the random control task, there were also 2 additional outputs, for which the target values are random.

The main task in all the experiments reported was isolated digit recognition ('1'-'9', 'zero' and 'oh'). In keeping with the Aurora methodology for assessing recogniser robustness (Pearce & Hirsch 2000), speech from the isolated word section of the TIDIGITS database was mixed with four types of noises (subway, babble, car and exhibition) taken from the NOISEX database at several SNRs (20dB, 15dB, 10dB, 5dB and clean). The performance of the digit recognition task was assessed for added noise amounts from -5dB to +25dB SNR at 5dB intervals.

4.1. Speech enhancement as the extra task

Here we investigate using speech enhancement as the additional task for both spectral and cepstral acoustic vectors.

Experiments were performed using the same training, validation and testing set as in (Parveen 2003). 1000 examples were chosen for training. A validation set of 110 examples was used to control the stopping condition in training. Tests were performed on the isolated digit samples in Aurora test set A. Only data from male speakers was used.

Proceedings of the 10th Australian International Conference on Speech Science & Technology

Accepted after full paper review

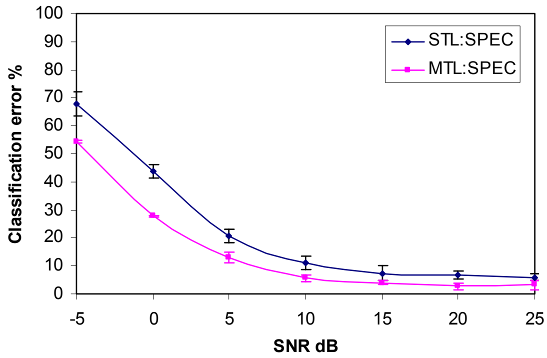

Figure 3: *Speech enhancement as the extra task, spectral domain acoustics*

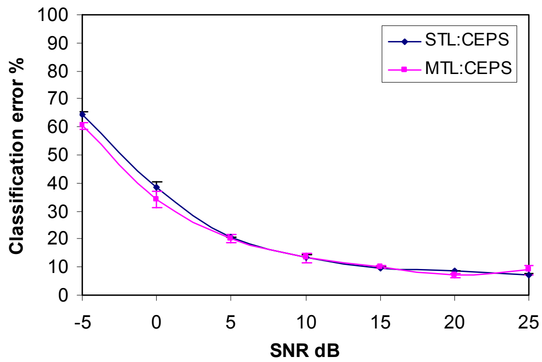

Figure 4: *Speech enhancement as the extra task, cepstral domain acoustics*

The classification performance of our STL and combined MTL digit classification plus speech enhancement net in spectral and cepstral domain is shown in Figure 3 and 4, respectively.

It can be seen that enhancement as an extra task gives significant improvement on the main task in the spectral domain (an average relative reduction in error rate of 47%) but not in the cepstral domain. A possible explanation is that with spectral data the net is able to use training errors in the enhancement task to learn that the values of corrupted channels may be inferred from their neighbours, but with orthogonalised cepstral features this information is not available. Redundancy in the spectral features can be exploited by the RNN to counter the effects of noise.

4.2. Adding gender recognition

Here we investigate gender recognition as an extra task, alone and in combination with speech enhancement. We also seek to confirm that a random task has no effect on the results. Experiments were performed using data from both male and female speakers in the isolated digits section of the AURORA database (Pearce & Hirsch 2000). 2000 examples were chosen for training. A validation set of 200 examples was used to control the stopping condition in training. Recognition performance was evaluated on the isolated digit section of Aurora test set A.

The classification performance of STL, MTL digit classification with speech enhancement, MTL digit classification with gender recognition, MTL digit classification with speech enhancement plus gender recognition and MTL digit classification with random task net in spectral and cepstral domain is shown in Figure 5 and Figure 6, respectively.

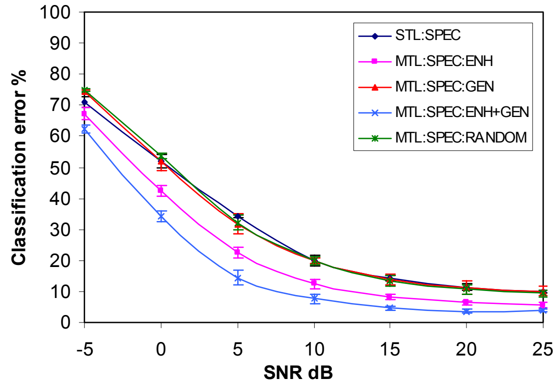

Figure 5: *Spectral features, enhancement, gender and random tasks*

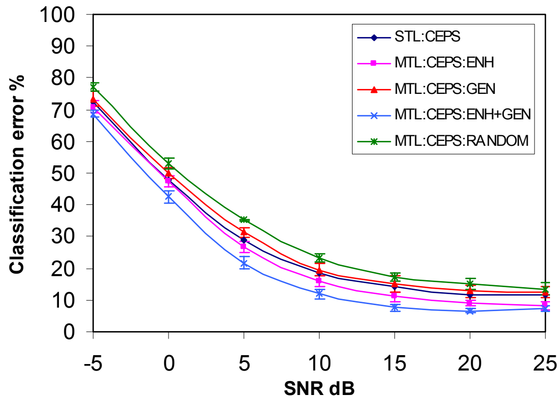

Figure 6: *Cepstral features, enhancement, gender and random tasks*

In Figure 5, STL, MTL gender and MTL random are very close to each other: gender recognition and the control task do not give any improvement for the main task. MTL with the enhancement task improves on STL by 6.846% on average in the gender-independent case.

When *both* enhancement and gender recognition are supplied as extra tasks, a further significant improvement is observed: 11.674% over STL on average. The improvement gained by incorporating two extra tasks is more than the addition of their improvements when applied separately. This illustrates the MTL principle of exploiting the relationship between

Accepted after full paper review

tasks: loosely speaking, knowing the speaker's gender helps to recognise the speech in difficult listening conditions.

Figure 6 shows a similar pattern for cepstral features. Here, the combination of enhancement and gender as extra tasks does give a significant improvement but overall, smaller improvements are obtained because there is no redundancy for the RNN to exploit.

5. Conclusion

In this paper we have demonstrated the performance of MTL incorporating different extra tasks: speech enhancement, gender recognition, and enhancement combined with gender recognition, in both spectral and cepstral domains. Experiments show that related tasks can improve the performance of the main task and a control experiment has shown no such improvement with an unrelated task. An unrelated task may even make performance worse. Furthermore, it is also found that the improvement by incorporating two extra tasks was more than the additional improvement of incorporating them in isolation. The overall performance in spectral domain was better than that in cepstral using RNNs.

6. Future Work

We have seen the success of implementing Multitask Learning over Single Task Learning in RNN for isolated digit classification. Our results suggest that related extra tasks can help to learn the main task better, thus resulting in high recognition accuracy.

Taking this into consideration we can expand our approach to implementing MTL for Large Vocabulary Continuous Speech Recognition (LVCSR) systems. MTL fits well with the hybrid approach (Bourlard & Morgan 1998), in which a network is used to estimate phone probabilities prior to statistical sequence decoding.

Acknowledgements

This work was conducted in part-fulfillment of a masters degree in Advanced Computer Science in the Department of Computer Science, University of Sheffield. The authors would like to thank Shahla Parveen for her valuable discussions and suggestions related to our work.

References

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.

Bourlard, H. & Morgan N. (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.

Caruana, R. (1997) *Multitask Learning, Machine Learning*, PhD Thesis, CMU.

Cooke M. P., Green P. D., Josifovski L. & Vizinho A. (2001), Robust Automatic Speech Recognition with Missing and Uncertain Acoustic Data, Speech Communication 34, p267-285.

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

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

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

Parveen, S.(2003) *Connectionist Approaches to the Deployment of Prior Knowledge for improving Robustness in Automatic Speech Recognition*, PhD Thesis, University of Sheffield.

Parveen, S.& Green, P.(2003) *Deployment of Prior Knowledge in Connectionist Robust ASR using Recurrent Neural Networks*. Eurospeech-03.

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.

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

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.

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after full paper review