

The Royal Academy of Engineering
International Travel Grant Report
ITG 04-783

Paul Dixon
School of Electrical, Electronic and Computer Engineering
University of Birmingham
Edgbaston
Birmingham

1 Introduction

This report sets out details of my visit to Professor Furui's laboratory at Tokyo Institute of Technology(Titech) in Japan from September 2004 to November 2004. The Royal Academy of Engineering kindly provided financial assistance towards my research visit.

The main reason for the visit to Japan was to perform experiments applying a new machine learning paradigm called *Products of Experts* to a speaker recognition task under the guidance of Professor Furui, which would allow me to receive detailed criticisms on my theoretical and practical work from a world leader in the field.

It also gave me the opportunity to study alongside Japanese researchers working in similar fields and to view at first hand any interesting new topics currently being actively researched in Japan. I also planned to visit Advanced Telecommunications Research Institute International (ATR) at Kyoto.

Finally I have been studying Japanese for approximately 6 years, and although I am fluent in the everyday use of the language I have always wanted to improve my technical Japanese. Working in a Japanese lab would give me an excellent opportunity to discuss my research in Japanese and to further improve my technical Japanese language skills.

In my PhD research work I have been investigating the use of a type of machine learning technique called a *Product of Experts* (PoE) [1] to speech processing tasks. A PoE is a type of generative model in which the outputs of many simpler probabilistic models are multiplied together and normalised to specify an overall density. Prior to the Japanese visit I had conducted research-applying PoEs to synthetic data and simple speech data. I have used this data to investigate, exhaustively, the practical issues which arise when applying PoEs to data. These include model initialisation and the stability of the training algorithms.

2 The Visit

At Tokyo Institute of Technology, a leading public university in Japan with a reputation for producing high quality research, I had the privilege of meeting with and working alongside Professor Furui who is an eminent scientist engaged in a wide range of research into spoken language processing. He has been responsible for a great deal of pioneering research in the field of speech processing, and is currently the President of the International Speech Communication Association (ISCA). In Professor Furui's laboratory there are approximately 25 members, namely, two assistant professors, post-doctorate researchers and students ranging from 4th year bachelor students to PhD students, together with a number of students from Europe and Asia.

At Titech I conducted experimental work on applying PoEs to speaker recognition. Speaker recognition is concerned with recognising the characteristics of a speaker's voice. There are two tasks speaker recognition can be divided into. Firstly speaker verification is the process of confirming a claimant's identity from their voice. Secondly speaker identification is the task of determining the speaker from a pool of known speakers [2]. Current state-of-the-art speaker recognition systems are statistical classifiers based on Gaussian mixture models (GMMs) [2]. A *speaker dependent model* is used to compute the probability $p(X|S)$ of a sequence of acoustic vectors given a particular speaker S , and a *general speaker* GMM is used to compute $p(X)$. The posterior probability $P(S|X)$ can then be obtained from Bayes' theorem.

Because both PoEs and GMMs are generative models, a speaker verification system was implemented using PoEs in place of the mixture models. The motivation is that the PoE based system may be able to uncover some underlying factorisation of the data and outperform the GMM system using fewer components. A significant problem with the PoEs is how to deal with a normalisation term that ensures their outputs define a true probability density function. Several techniques were explored to deal with this issue. Experiments were conducted on both speaker verification and identification tasks using the YOHO database [3].

Speaker Identification is a closed set task, to deal with the un-normalised scores a single layer classifier was connected to the PoE outputs. The PoE

system was able to perform comparably with the baseline system for a small amount of experts, however, as the expert count was increased the conventional GMM system performed better.

For the speaker verification experiments several score normalisation techniques were applied namely, background speaker model, T-Norms and Z-Norms [4]. Unfortunately the PoEs did not perform well on this task.

My future work in the UK will concentrate on trying to explain why the PoE system does not perform as well as the GMM based one; establishing whether this is due to the modelling capabilities of this particular PoE or an issue with the particular learning algorithm which was used.

During my visit I gave a presentation of my prior work titled “Vowel Classification using Products of Experts” to Professor Furui and his colleagues at Titech. Following the presentation I answered questions about the merits of Products of Experts and the motivations for my research work. This was a worthwhile experience, which also provided me with some important feedback regarding my presentation skills, which will be very useful when I make oral presentations at conferences in the future.

I attended the weekly group seminars where members of the group presented research papers. These sessions were particularly worthwhile because it exposed me to the terminology used in technical Japanese.

During my stay at Titech the International Workshop on Speech Summarization for Information Extraction and Machine Translation (IWSpS) was held. I was fortunate enough to be able to attend the talks, which included speakers from the USA and Germany. This was my first exposure to cutting edge work on this subject and the highlight of the talks were the presentations given by Dr Kevin Knight and Dr Stephan Vogel on machine translation.

At Titech I met Slaven Bilac, a researcher who is currently working on computational linguistics. He introduced me to a number of other researchers working at the National Institute of Informatics in Tokyo. I have kept in contact with one of these researchers and we are planning to collaborate on a future project together.

During my visit to Japan I was given access to a considerable amount of technical material (journals, books) written in Japanese. These journals described the latest areas of research that are currently being undertaken in Japanese Universities and Industry.

I travelled to Kyoto to meet Dr Nick Campbell together with some members of his staff at ATR a government funded research laboratory in Western Japan. Dr Campbell is working on speech synthesis, and after discussions about my work he took me on a tour of ATR and introduced me to members of his team who were also researching speech. In some of the other labs at ATR I observed a great amount of research being undertaken in the area of domestic robots and human robot interaction. This area of research is currently a very hot topic in Japan.

At the time of my original application to the RAE I had hoped to combine my visit to Japan with attendance at the International Conference of Spoken Language Processing (ICSLP) in Korea, however due to time and financial constraints this was not possible.

3 Conclusions

I have had the privilege to meet many other researchers including world leaders in the field of speech processing, broadening my academic horizons. I have undertaken several experiments that will contribute towards my PhD thesis. I will be finishing my PhD studies this year and the visit gave me considerable insight into my field of research together with new ideas of areas I would like to pursue upon the completion of my studies.

Finally, I would like to thank The Royal Academy of Engineering for their generous contribution towards the cost of this visit.

References

- [1] Hinton, G.E. "Training Products of Experts by Minimizing Contrastive Divergence", *Neural Computation*, 14:1771, 2002.
- [2] Reynolds, D.A. "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication*, vol.17, pp. 91-108, 1995.
- [3] Campbell, J.P., Jr. "Testing with the YOHO CD-ROM voice verification corpus". *IEEE ICASSP*, vol.1, pp. 341-344, 1995.
- [4] Auckenthaler R., Carey M. and Lloyd-Thomas H. "Score Normalization for Text-Independent Speaker Verification Systems". *Digital Signal Processing*, vol.10, no. 1, pp. 42-54, 2000.