論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Speaker identification using products of experts
著者(和文)	ディクソン ポール, 古井 貞熙
Authors(English)	Paul Dixon, Sadaoki Furui, Martin Russell
出典(和文)	日本音響学会2005年春季講演論文集, Vol. , No. 1-5-5, pp. 9-10
Citation(English)	, Vol. , No. 1-5-5, pp. 9-10
発行日 / Pub. date	2005, 3

1—5—5

Speaker Identification using Products of Experts*

©Paul Dixon[†‡] Sadaoki Furui[†] Martin Russell[‡]
Tokyo Institute of Technology[†] University of Birmingham [‡]

1 Introduction

Products of Experts (PoEs) are a type of generative model in which the outputs of many simpler probabilistic models are multiplied together and normalised to specify an overall density. In this paper we investigate the application of PoEs to a speaker identification task and compare the results with those obtained using a conventional Gaussian mixture based system.

Speaker recognition is concerned with identifying the characteristics of a speaker's voice as opposed to actually transcribing what has been said. There are two tasks speaker recognition can be divided into. Firstly *speaker verification* is the process of confirming a claimants identity from their voice. Secondly *speaker identification* is the task of determining the speaker from a pool of known speakers. In addition both of these tasks can be either text-dependent where the system is restricted to certain words or phrases or text-independent where there are no such constraints. State-of-the-art speaker recognition systems are statistical classifiers based on Gaussian mixture models (GMMs) [1]. 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 be obtained from Bayes' theorem.

Hinton's *Product of Experts* paradigm is an alternate method of combining many simpler experts together to build more complex models. PoEs have the advantage that they can have an overall distribution that is sharper than any of their component distributions. For a point to be given a high probability under the model it needs a majority vote among the experts, in other words individual experts can assign high probabilities to invalid regions of the space which are then rejected by the other experts. This leads to a model that can perform well in high dimensional spaces with fewer components and much vaguer experts [2].

Others[2, 3] have reported good results using PoEs on tasks such as optical character recognition and sentence modelling. Because of the similarities PoEs have with GMMs we would like to know how they can perform on speech processing tasks, and in particular speaker recognition. One possible advantage of the PoE is that it may discover some underlying factorisation of the data.

2 Products of Experts

In a PoE model with n experts the probability of data vector $\boldsymbol{d}$ is defined as:

$$p(\boldsymbol{d} \mid \theta_1, \ldots, \theta_n) = \frac{\prod_m p_m(\boldsymbol{d} \mid \theta_m)}{\sum_c \prod_m p_m(\mathbf{c} \mid \theta_m)}, \quad (1)$$

where θ_m are the parameters of the m^{th} expert, and $p_m(\boldsymbol{d}|\theta_m)$ is the probability of data vector $\boldsymbol{d}$ under the m^{th} expert [2]. The denominator term is the sum over all possible data vectors and makes training difficult. Estimating parameters by maximising the likelihood of the data requires direct evaluation of the denominator and therefore cannot be applied. When training the PoE model we would like to minimise the Kullback-Liebler (KL) divergence between the training data's distribution $P^0$and the equilibrium distribution P_θ^∞ of data generated by the model through exhaustive Gibbs sampling[2].

$$KL(P^0 \| P_\theta^\infty) = \sum_d P^0(\boldsymbol{d}) \log \frac{P^0(\boldsymbol{d})}{P_\theta^\infty(\hat{\boldsymbol{d}})} \quad (2)$$

Fitting data this way is still problematic as sampling from the model's equilibrium distribution is computationally expensive and there is a tendency for the sampler to become trapped in a certain mode of the distribution. To avoid these problems we use the *contrastive divergence* based learning algorithm[2] and minimise the following criterion.

$$CD = KL\left(P^0 \| P_\theta^\infty\right) - KL\left(P_\theta^1 \| P_\theta^\infty\right) \quad (3)$$

Where P_θ^1 denotes the distribution of the Gibb's one-step data reconstructions. After taking derivatives the equilibrium terms involving P_θ^∞ will cancel out. Ignoring a term which Hinton[2] shows is safe to discard, yields the following approximate learning rule:

$$\Delta\theta_k \propto \left\langle \frac{\partial log p_k(\boldsymbol{d}|\theta_k)}{\partial\theta_k} \right\rangle_{P^0} - \left\langle \frac{\partial log p_k(\hat{\boldsymbol{d}}|\theta_k)}{\partial\theta_k} \right\rangle_{P_\theta^1} \quad (4)$$

The justification of this update rule is that in a perfect model there will be no drift in a Gibbs sampling chain started at the data distribution. A real imperfect model will move away from P^0 and the one-step reconstruction gives sufficient information about P_θ^∞ to adjust the model parameters to fit the data.

---

* プロダクト・オブ・エキスパートを用いた話者識別
ディクソン ポール[†‡], 古井 貞熙[†], ラッサル マーティン[‡] (東工大[†], バーミンガム大[‡])

3 Experimental Setup

We evaluated the performance of the PoEs by performing a text-independent speaker identification task on the YOHO corpus. The YOHO database consists of clean speech recorded from 138 speakers. There are 96 training utterances and 40 test utterances per speaker. Each utterance is a combination lock phrase e.g "fifty two - eighty one - twenty seven"[4]. These experiments were conducted on a 118 speaker subset.

The speech waveforms were converted into a sequence of 12 dimensional Mel Frequency Cepstral vectors (25ms analysis window, 10ms frame rate).

For the speaker models we need to be able to build arbitrary shaped densities. Selecting a Gaussian distribution as a PoE expert will not allow for this, because a product of Gaussians is a Gaussian. Instead the *unigauss* expert[2] was used. This is a simple mixture model composed of a Gaussian and uniform distribution. The probability density function (pdf) of the unigauss expert is given by,

$$p_k(\boldsymbol{d}|\theta_k) = w_k \mathcal{N}(\boldsymbol{d} \mid \mu_k, \Sigma_k) + \frac{1 - w_k}{r}, \quad (5)$$

where r is the volume of the support of the uniform distribution, and w_k is the weighting between the two components[2].

The PoE models were randomly initialised with a wide variance and trained using contrastive divergence. It has already been noted that computation of the denominator term in (1) is a significant difficulty. For classification tasks this is a particular problem as the resulting class-conditional pdfs may have unequal areas, resulting in a bias. To deal with this, the un-normalised log PoE scores were fed onto a linear network with logistic activations. The weights of this network were learnt using a gradient descent algorithm [2, 3].

For a baseline comparison each PoE system was compared to a GMM based system with the same number of components. The GMMs were initialised using K-Means clustering and trained using the expectation maximisation algorithm. The Gaussians in both systems had diagonal covariance matrices. To ensure a fair comparison between the systems, the GMM system was also connected to a linear network in the same way. This guaranteed both systems had the same number of parameters and were exposed to an identical amount of training.

To evaluate the speaker identification performance with respect to the number of experts per speaker model, experiments were conducted with expert counts of 10, 25, 50 and 100. In the limit, as the number of GMM components is increased the PoE cannot do better than the GMM. One might hope for performance improvement when the number of experts is small.

4 Results

Figure 1 shows the performances of both of the systems. For the GMM system ID rates are given with, and without, the linear network in place. When the linear network was removed from the PoE system the identification rate was 0.875% in all cases, which is the performance of a random classifier. This corresponds to all data being classified as one speaker, due to the failure to compute the denominator term in (1). Even with the linear network the PoE system is not able to perform as well as the GMM system. An interesting result is the reduction in error rates seen when the GMM classifier is used in combination with the linear network.

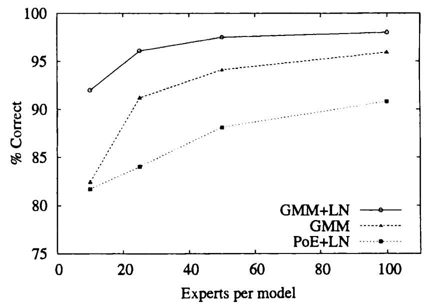

Figure 1. Speaker identification performance

5 Discussion

We have applied PoEs to a text-independent speaker identification task. The results show that on the YOHO database PoEs cannot perform as well as GMMs. When using PoEs in a classification system it is essential to deal with the denominator term in (1) to achieve satisfactory performance.

Future work 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 contrastive divergence learning and the initialisation of the experts.

References

[1] Reynolds, D.A. "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, vol.17, pp. 91-108, 1995.

[2] Hinton, G.E. "Training Products of Experts by Minimizing Contrastive Divergence", Neural Computation, vol. 14, pp. 1771-1800, 2002.

[3] Hinton, G.E. and Mayraz, G. "Recognizing Handwritten Digits using Hierarchical Products of Experts", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, pp. 189-197, 2002.

[4] Campbell, J.P., Jr. "Testing with the YOHO CD-ROM voice verification corpus". IEEE ICASSP, vol.1, pp. 341-344, 1995.