TAIL DISTRIBUTION MODELLING USING THE RICHTER AND POWER EXPONENTIAL DISTRIBUTIONS

M.J.F. Gales and P. A. Olsen

IBM T.J. Watson Research Center, P.O. Box 218
Yorktown Heights, NY 10598, USA
Email: {mjfg,pederao}@watson.ibm.com

ABSTRACT

The vast majority of HMM–based speech recognition systems use Gaussian mixture models as the state distribution model. The use of these distributions is motivated more by ease of training, decoding and the fact that a sufficient number of Gaussian components may be used to approximate any distribution, than some underlying aspect of the data being modelled. If distributions were selected that better modelled the observed data, fewer components should be required and recognition accuracy should improve. This paper examines two distributions for improving the modelling of the tails of the densities. The first distribution, the Richter distribution, fits within the general framework of Gaussian component tying, but has some attractive attributes for decoding. The second distribution, the power exponential, does not fit within a tying framework. Despite gains in likelihood, indicating that the Gaussian components are sub–optimal in a likelihood sense, only small gains in recognition performance were observed on a large vocabulary speech recognition task.

1. INTRODUCTION

Modelling of the statistical distribution of acoustic data is commonly done using Gaussian mixtures. A notable exception is Phillips' speech recognizer whose densities are mixtures of Laplacians [1]. However, examining the histograms of a single dimension of the acoustic data assigned to a particular state of a speech recognition system, there are three notable features that are distinctly non–Gaussian. Some histograms are skew–symmetric, peakier than typical Gaussians and have tails that taper off at a slower rate than for a Gaussian tail. Despite these limitations, mixtures of Gaussians perform well in speech recognition experiments. If the modelling of the tail, peak or skewness of the distribution is improved the performance of the recognizer may be expected to improve. One option to improve modelling is to increase the number of Gaussian components in the mixtures. However, this dramatically increases the total number of parameters in the speech recognizer. It would be desirable to improve the density model without drastically changing the number of parameters. This paper proposes two different distributions to improve modelling of the tail of the distributions, the Richter distribution and the power exponential distribution. Both the proposed distributions are symmetric, so they do not address the skew–symmetric problem.

First, the Richter distribution is examined. This class of distributions was first suggested by Alan Richter in [7], and was referred to as the Richter distribution in [3]. The Richter distribution is a mixture of Gaussians where all the means are equal and the covariance matrices are multiples of each other. This may be considered as a particular form of Gaussian mixture parameter tying. A Richter distribution consisting of R Gaussians will only have $2R-2$ parameters in addition to parameters describing a single Gaussian. Despite the small increase in memory and computational load Richter distributions have fallen out of favour compared with more standard tying schemes. Second, the power exponential distributions is considered. A power exponential distribution is a distribution for which the exponent of a Gaussian is raised to a power possibly different from that of the Gaussian. For large powers the power exponential distributions become increasingly more like a uniform distribution whereas for small powers the distributions have sharp peaks and heavy tails. In the case of small powers special care must be taken when estimating the means, variances and mixture weights [2]. The power changes the behavior of the tails drastically, but adds only one parameter to that of a single Gaussian. The increase in memory requirements is therefore small, whereas the computational load is somewhat larger.

This paper details re–estimation formulae for training HMM–based speech recognition systems with both Richter and power exponential components. In addition, equations for adapting Richter distributions using linear transformations are described. The performance of the two systems are then compared with appropriate Gaussian component systems on the 1997 Hub4 partitioned evaluation test set.

2. RICHTER DISTRIBUTIONS

One scheme for improving the tail distribution modelling is to use the class of distribution described by

$$f(\mathbf{o};\mu,\mathbf{\Sigma},p(v)) = \int \mathcal{N}(\mathbf{o};\mu,v^2\mathbf{\Sigma})p(v)dv \ , \qquad (1)$$

where $p(v)$ is a probability density function, i.e. $p(v) \geq 0$ and $\int p(v)dv = 1$. It is simple to see that this form of distribution is a generalisation of the standard Gaussian distribution where $p(v) = \delta_1(v)$, $\delta_{v_r}(v)$ is the Kronecker delta function. This class also includes the Cauchy distribution as another standard case. By appropriately modifying the distribution of v it is possible to alter both the tails and the 'peakiness' of the distribution. In [6] an EM scheme is described for ML estimates of μ and $\mathbf{\Sigma}$ that does not require explicitly obtaining the distribution $p(v)$, which remains unaltered during training. The discrete version of (1) was described in [7]. Here $p(v) = \sum_r w_r \delta_{v_r}(v)$ with $w_r > 0$, $\sum_r w_r = 1$. Then

$$f(\mathbf{o};\mu,\mathbf{\Sigma},p(v)) = \sum_r w_r \mathcal{N}(\mathbf{o};\mu,v_r^2\mathbf{\Sigma}) \ . \qquad (2)$$

In addition to giving the formulae for calculating the μ and $\mathbf{\Sigma}$, formulae are given for ML estimates for the discrete distribution

of v are described. This form of distribution was used in [3] for discrete speech modelling, though in the experiments described the discrete distribution of v was determined a priori rather than trained from the data.

For large vocabulary speech recognition systems multiple Gaussian components are typically used to model each state. This paper therefore considers the Richter mixture case where each state is modelled by a mixture of Richter components

$$\mathcal{L}(\mathbf{o}) = \sum_m \sum_r w_r^{(m)} \mathcal{N}(\mathbf{o}; \mu^{(m)}, v_r^{(m)2} \mathbf{\Sigma}^{(m)}) \ . \tag{3}$$

Furthermore it has become very common to *tie* parameters together, thus reducing the number of parameters to be stored and increasing the robustness of the parameters estimated. In the same fashion it is possible to tie the Richter distribution parameters $\mathbf{w}^{(m)}$ and $\mathbf{v}^{(m)}$ over many Richter distributions.

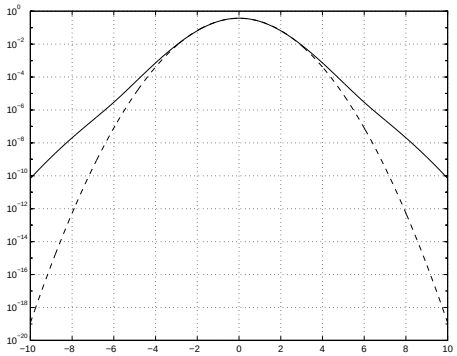

Figure 1: The log–likelihood of a Gaussian distribution and a Richter distribution using global Richter weights and scaling factors obtained from the Hub4 training data, '—' indicates the Richter distribution, '- -' indicates the Gaussian distribution.

Figure 1 shows a comparison of the log–likelihood of a four Richter component distribution and the equivalent Gaussian distribution. Globally tied Richter distribution parameters were obtained using the Hub4 training data. The tails of the Richter distribution are longer than those of the Gaussian distribution. This indicates that, at least in a likelihood sense, the Gaussian components are sub–optimal.

2.1. Parameter Estimation

The following re–estimation formulae, which are slightly modified versions of those presented in [3], are used

$$\hat{\mu}^{(m)} = \frac{\sum_{r,\tau} \frac{\gamma_r^{(m)}(\tau)}{\hat{v}_r^{(p_m)2}} \mathbf{o}(\tau)}{\sum_{r,\tau} \frac{\gamma_r^{(m)}(\tau)}{\hat{v}_r^{(p_m)2}}} \ , \tag{4}$$

$$\hat{\mathbf{\Sigma}}^{(m)} = \frac{\sum_{r,\tau} \frac{\gamma_r^{(m)}(\tau)}{\hat{v}_r^{(p_m)2}} \hat{\mathbf{W}}^{(m)}(\tau)}{\sum_{r,\tau} \gamma_r^{(m)}(\tau)} \ , \tag{5}$$

$$\hat{v}_r^{(p)2} = \frac{\sum_{M^{(p)},\tau} \gamma_r^{(m)}(\tau) \hat{q}^{(m)}(\tau)}{d \sum_{M^{(p)},\tau} \gamma_r^{(m)}(\tau)} \ , \tag{6}$$

and

$$\hat{w}_r^{(m)} = \hat{c}^{(m)} \frac{\sum_\tau \gamma_r^{(m)}(\tau)}{\sum_{M^{(p)},r,\tau} \gamma_r^{(m)}(\tau)} \tag{7}$$

where

$$\hat{c}^{(m)} = \frac{\sum_{r,\tau} \gamma_r^{(m)}(\tau)}{\sum_{S^{(m)},r,\tau} \gamma_r^{(m)}(\tau)} \tag{8}$$

$$\hat{q}^{(m)}(\tau) = \left(\mathbf{o}(\tau) - \hat{\mu}^{(m)}\right)^T \hat{\mathbf{\Sigma}}^{(m)-1} \left(\mathbf{o}(\tau) - \hat{\mu}^{(m)}\right)$$

$$\hat{\mathbf{W}}^{(m)}(\tau) = (\mathbf{o}(\tau) - \hat{\mu}^{(m)})(\mathbf{o}(\tau) - \hat{\mu}^{(m)})^T \ . \tag{9}$$

$M^{(p)}$ is the set of components sharing the same Richter parameters, p_m is the Richter class of component m, d is the dimensionality of the observation vector $\mathbf{o}(\tau)$ and $\gamma_r^{(m)}(\tau)$ is the posterior probability of being in Richter component r of component m at time τ and $S^{(m)}$ is the set of components in the same state as m. Formulae (4)–(7) yield an iterative estimation scheme since the mean and the variance are a function of $\hat{\mathbf{v}}^{(r)}$, which itself is a function of the estimates of the mean and variance. The sufficient statistics for this operation are the occupancy, sum and sum squared of the feature vector for each Richter distribution of each component. Thus if there are M components and R Richter distributions per component, the equivalent of $M \times R$ components must be stored. An alternative to this and the one used in this paper is to either update the Richter distribution parameters or the means and variances. In this case it is only necessary to store parameters at the Richter tying level or the component level.

2.2. Likelihood Calculation

One of the reasons for using Richter distributions rather than additional Gaussian components is the efficiency of the likelihood calculation. The likelihood of an observation coming from a particular component is given by

$$\mathcal{L}(\mathbf{o}(\tau); m) = \sum_{m,r} b_r^{(m)} \exp\left(-\frac{q^{(m)}(\tau)}{2v_r^{(m)2}}\right) \ ,$$

where $q^{(m)}(\tau)$ is a function of the component, m, and observation

$$q^{(m)}(\tau) = \left(\mathbf{o}(\tau) - \mu^{(m)}\right)^T \mathbf{\Sigma}^{(m)-1} \left(\mathbf{o}(\tau) - \mu^{(m)}\right) \tag{10}$$

and $b_r^{(m)}$ is a function of the Richter component r, but independent of the observation

$$b_r^{(m)} = \frac{1}{\sqrt{2^d \pi^d |\det \mathbf{\Sigma}^{(m)}|}} \frac{w_r^{(m)}}{\sqrt{v_r^{(m)2d}}} \ ,$$

The main additional cost is therefore in the log–add over the Richter components. This may be ignored if a max of the components is taken, rather than the sum.

2.3. Adapting Richter Distributions

It is also common to use linear transformations to adapt model parameters to be more representative of a particular speaker, or acoustic environment. A variety of linear transformations and re–estimation formulae are described in [5]. Modifying these formulae to handle Richter distributions is trivial. The main modification is to deal with $\frac{\gamma_r^{(m)}(\tau)}{v_r^{(m)2}}$ rather than the standard posterior component probability. As an example the estimation formulae for the transform $\hat{\mathbf{A}}$ in maximum likelihood linear regression, where Richter components are used, is

$$\hat{\mathbf{a}}_i = \mathbf{k}^{(i)}\mathbf{G}^{(i)-1} \ , \tag{11}$$

where

$$\mathbf{G}^{(r)} = \sum_{M,\tau,r} \frac{\gamma_r^{(m)}(\tau)}{v_r^{(m)2}\sigma_i^{(m)2}} \xi^{(m)}\xi^{(m)T} \tag{12}$$

and

$$\mathbf{k}^{(i)} = \sum_{M,\tau,r} \frac{\gamma_r^{(m)}(\tau)}{v_r^{(m)2}\sigma_i^{(m)2}} \xi^{(m)T} o_i(\tau) \ . \tag{13}$$

Similarly modifications to the variance adaptation formulae are possible.

3. POWER EXPONENTIALS

Consider the class of densities

$$f(\mathbf{o};\mu,\mathbf{\Sigma},\alpha) = \rho_\alpha |\det\mathbf{\Sigma}|^{-1/2} \exp(-(\gamma_\alpha q)^{\alpha/2}), \tag{14}$$

where

$$q = (\mathbf{o}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{o}-\mu) \ , \tag{15}$$

$$\gamma_\alpha = \frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)} \tag{16}$$

and

$$\rho_\alpha = \frac{\alpha\Gamma^{\frac{1}{2}}(3/\alpha)}{2\Gamma^{\frac{3}{2}}(1/\alpha)}. \tag{17}$$

This class was recently suggested and studied in [2]. The one dimensional case appears to have first been suggested by Subbotin, [8]. The class (14) will be referred to as the the power exponential distribution. It is also known as the error function, p–Gaussians or as α–Gaussians.

Following (3) a model is considered where each state in the system is modelled by a mixture of power exponential distribution, i.e.

$$\mathcal{L}(\mathbf{o}) = \sum_m w^{(m)} f(\mathbf{o};\mu^{(m)},\mathbf{\Sigma}^{(m)},\alpha^{(m)}). \tag{18}$$

It is worth noticing that the class of functions described in (14) is not a subset of the class described in (1). Power exponential distributions can not in general be modelled with Richter distributions. This fact can be verified by noticing that functions in the class (1) are all log concave, whereas the power exponentials are not log concave for $0 < \alpha < 1$. This makes the framework of [6] unsuitable for parameter update for $0 < \alpha < 1$.

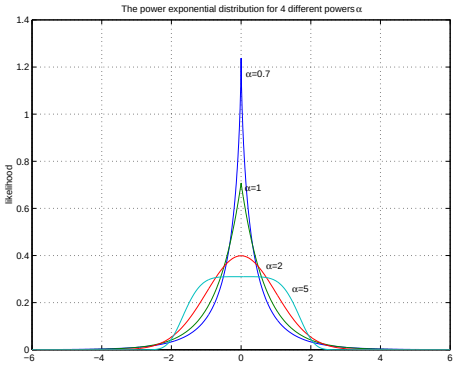

Figure 2: The power exponential function for various values of α.

3.1. Parameter Estimation

The estimation formula for $w^{(m)}$ is identical to the standard HMM re–estimation formulae. Update formulae for $\mu^{(m)}$ and $\mathbf{\Sigma}^{(m)}$ are suggested in [2]:

$$\hat{\mu}^{(m)} = \frac{\sum_\tau \gamma^{(m)}(\tau)\left(q^{(m)}(\tau)\right)^{\alpha^{(m)}/2-1}\mathbf{o}(\tau)}{\sum_\tau \gamma^{(m)}(\tau)\left(q^{(m)}(\tau)\right)^{\alpha^{(m)}/2-1}} \ , \tag{19}$$

and

$$\hat{\mathbf{\Sigma}}^{(m)} = \frac{\sum_\tau \gamma^{(m)}(\tau)\left(q^{(m)}(\tau)\right)^{\alpha^{(m)}/2-1}\hat{\mathbf{W}}^{(m)}(\tau)}{\sum_\tau \gamma^{(m)}(\tau)} \ , \tag{20}$$

where $q^{(m)}(\tau)$ is defined in equation (10), $\hat{\mathbf{W}}^{(m)}(\tau)$ in equation (9) and $\gamma^{(m)}(\tau)$ is defined to be the posterior probability of being in the power exponential component m at time τ. It is not known that the overall likelihood is guaranteed to increase with the update given by (19)–(20), but numerical evidence suggests that this is so.

Special consideration for $0 < \alpha < 1$ is suggested in [2]. The powers $\alpha^{(m)}$ can either be fixed on a global level or they can be updated according to the formula given in [2]:

$$\hat{\alpha}^{(m)} = \tag{21}$$
$$\text{argmax}_\alpha \sum_\tau \gamma^{(m)}(\tau)\log\left(f(\mathbf{o}(\tau);\hat{\mu}^{(m)},\hat{\mathbf{\Sigma}}^{(m)},\alpha)\right)$$

With this update of $\alpha^{(m)}$ the likelihood is guaranteed to increase. Figure 3 shows the distribution of α estimated on a per component case. The mean of the values of α is approximately one. It is interesting to note that the Gaussian component equivalent of power exponential components, $\alpha = 2$, occurs infrequently. Again, this indicates that Gaussian components are sub–optimal in a likelihood sense.

Currently adaptation of power exponentials have not been investigated.

4. RESULTS

The two forms of modified tail distribution modelling were investigated on the 1998 Hub4 partitioned evaluation test set. A 60–dimensional LDA based front end was used. The LDA was based

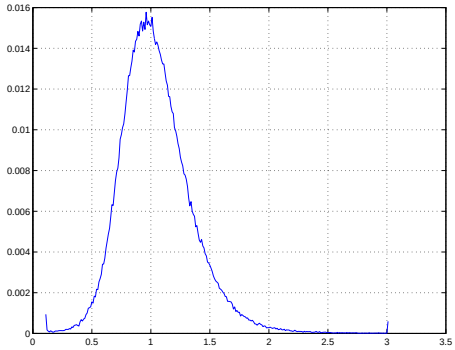

Figure 3: The distribution of powers, α, after training using (21) on the Hub4 1997 data.

on splicing 9 time frames of 24 dimensional Cepstra, including c_0. A context dependent state–clustered allophone system was built on the broadcast news training data. More details of the test and language model setups are given in [4].

The baseline system for the Richter components had a total of about 135,000 components. A 4 distribution Richter component system (R=4) was initialised using the means and variances of the baseline system. In preliminary experiments the best, though not significantly so, system was found to be one with the Richter parameters tied at the state level. This is the one considered in these experiments. Table 1 shows the comparison of a Richter system

System	Error Rate (%)		
	F0	F1.	Avg
base	11.6	18.5	18.7
base+adapt	10.1	17.0	16.4
Richter	11.3	18.1	18.4
Richter+adapt	10.1	16.9	16.3

Table 1: Results on the Hub4 1997 partitioned evaluation test set

and the equivalent baseline system. The adaptation scheme used in both was a global mean and full variance transform described in [5]. This was applied in an unsupervised batch adaptation mode. Using Richter components showed a small gain in performance over the standard Gaussian components. After adaptation the performance of the two systems was almost indistinguishable.

The experiments using power exponential components used a modified baseline system consisting of approximately 120,000 Gaussians. The test was performed on a subset of the 1997 partitioned evaluation that was used for development [4]. Finally a smaller language model than for that of the tests with the Richter distribution where used, thus degrading the performance for the spontaneous speech category, F1, and for some of the more difficult conditions, F2–FX. Two power exponential systems were built. The first used a fixed value of $\alpha^{(m)} = 1$ for all components, motivated by figure 3. The second system used a per–component value of $\alpha^{(m)}$ obtained using equation (21).

Table 2 shows the performance of the various power exponential systems. Again only small reductions in word error rate were observed using the improved tail modelling.

System	Error Rate (%)		
	F0	F1.	Avg
base ($\alpha = 2$)	11.8	22.9	26.1
$\alpha = 1$	11.5	23.0	25.5
EM update for α	11.9	22.6	25.4

Table 2: Results for the power exponential distribution on a subset of the Hub4 1997 partitioned evaluation test set

5. CONCLUSIONS

This paper has described two schemes for improving the tail distribution modelling in an HMM–based speech recognition scheme. Though both schemes indicate that Gaussian components are sub–optimal in a likelihood sense, they yielded only small reductions in word error rate. Though disappointing in terms of reductions in word error rate, the results indicate that using alternatives to Gaussian components for speech modelling may be useful. Investigating other distributions may give reductions in the word error rate. In particular both distributions investigated in this paper are symmetric, still requiring multiple components to model any non–symmetric attributes of the data. Explicit non–symmetric distributions may be an interesting avenue of investigation.

6. REFERENCES

[1] X Aubert, H Bourland, Y Kamp, and C J Wellekens. Improved hidden Markov models for speech recognition. *Phillips Journal of Research*, 43:254–245, 1988.

[2] S Basu, C A Micchelli, and P A Olsen. Power exponential densities for the training and classification of acoustic feature vectors in speech recognition. *Research report, T.J.Watson Research Center*, 1999.

[3] P Brown. *The Acoustic-Modelling Problem in Automatic Speech Recognition*. PhD thesis, IBM T.J. Watson Research Center, 1987.

[4] S S Chen, E M Eide, M J F Gales, R A Gopinath, D Kanevsky, and P Olsen. Recent improvements to ibm's speech recognition system for automatic transcription of broadcast news. In *Prodeedings of the Broadcast News Transcription and Understanding Workshop*, 1999.

[5] M J F Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.

[6] L A Liporace. Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Transactions Information Theory*, 28:729–734, 1982.

[7] A G Richter. Modelling of continuous speech observations. In *Advances in Speech Processing Conference*. IBM Europe Institute, 1986.

[8] M Subbotin. On the law of frequency of errors. *Mathematicheskii Sbornik*, 31:296–301, 1923.