MULTI-RATE HMM QUANTIZATION FOR SPEECH RECOGNITION

Marcel Vasilache

Nokia
P.O. Box 100, 33721 Tampere, Finland
E-mail: marcel.vasilache@nokia.com

ABSTRACT

This paper refines the idea of scalar quantization for hidden Markov model (HMM) parameters which was introduced in an earlier contribution. With the proposed multi-rate approach it is shown that an increased model compression can be achieved with a significant computational complexity reduction while also closely preserving the recognition performance of the original models.

Index Terms— Hidden Markov models, Quantization, Speech recognition

1. INTRODUCTION

As a consequence of the explosive market penetration of portable computing platforms and the attractivity of speech recognition for such personal and input challenged devices, substantial efforts have been made to adapt speech recognition solutions to the limited computational capabilities of the mobile devices.

Of key importance are solutions which focus on the size reduction of the acoustic models [1, 2, 3, 4, 5, 6, 7] since, as direct consequence of this target, substantial computational complexity reductions are also seen. The interest in effective compression schemes for acoustic models is not likely to fade even though the continuously increasing storage and computational capabilities of mobile devices are now challenging the personal computers of the not so distant past. A persistent driver for even better compression performance is represented by the intense competition for resources of the multitude of applications which now find their way into the mobile world. Another equally important factor consists in the increased demands placed on speech recognition engines. These have evolved from simple speaker dependent systems into large multi-lingual speaker independent systems which have to cope with increasingly difficult tasks. High recognition capabilities and the inherently difficult mobile environment often require acoustic models of increased precision with the added option of creating speaker specific model sets. All of these result in storage requirements for large numbers of parameters.

One of the dominant methods of parameters compression for HMMs is described in [2], [3] and [5]. These models, subspace distribution clustered HMMs, have the capability to effectively approximate the original models with decreasing rates of quantization. However, when either the speaker or his environment significantly differs from the training conditions this accurate modelling comes at the cost of reduced flexibility for parameter changes in the process of adaptation. The scalar quantized models, as introduced in [1] and further illustrated in [8] and [6], trade off a somewhat reduced compression capability for an increased adaptation flexibility and a much simpler implementation design.

In the following, the paper will briefly revise the design of scalar quantized HMMs with a stronger focus on the multi-rate construction. In the final sections a set of experiments are performed and conclusions are drawn.

2. CDHMM WITH SCALAR QUANTIZED PARAMETERS

2.1. Basic structure

The scalar quantization for continuous density HMMs (CDHMMs) targets each mean and variance component individually. For a given set of models, once state densities have been trained, the estimated mean and variance values are replaced by using a small subset of values as given by dedicated scalar quantizers. In a direct approach, each density dimension would require a pair of quantizers to more accurately reflect the distribution of the mean and variance parameters associated to it. Since for most used state emission likelihood functions the parameter estimation is invariant to the energies in the individual streams, as design simplification, it is beneficial to bring the parameters within the same bounds. This can be done either as part of front-end design or by the use of two global mean and scale vectors during training or, finally, by a global affine transformation of the feature vectors and model parameters before quantization and during recognition.

With the range of values under strict control it is possible to share the same scalar quantizers for all the model dimensions. With the shared quantizer we consider that the normalizing mean and scale vectors are part of quantization parameters. The values in these two vectors are such chosen as to maximize the overlap of the individual component distributions with a provision of discarding outliers.

Finally, with the set of mean and inverse standard deviation parameters from all density components, two Lloyd-Max scalar quantizers are trained. By using these quantizers the model parameters are then replaced with joint indexes (for each mean and inverse standard deviation pair) and the two quantizers and global mean and scale vectors are stored with the models.

2.2. Multi-rate design

In most feature extraction setups, the feature vector components are of unequal importance for the classification task. For a conventional CDHMM based speech recognizer engine, when using unit energy streams, we can illustrate the differences with the plot of the median of the inverse standard deviation across each component of the trained acoustic model set. Assuming also that the components are independent, this gives a rough initial idea of their estimation precision and their relative importance for the recognition task. For example, in the case of the $C0$ to $C12$ Mel cepstral coefficients, the decreasing importance trend is obvious, with the 1st component, the energy-like $C0$, being by far the most important (Figure 1).

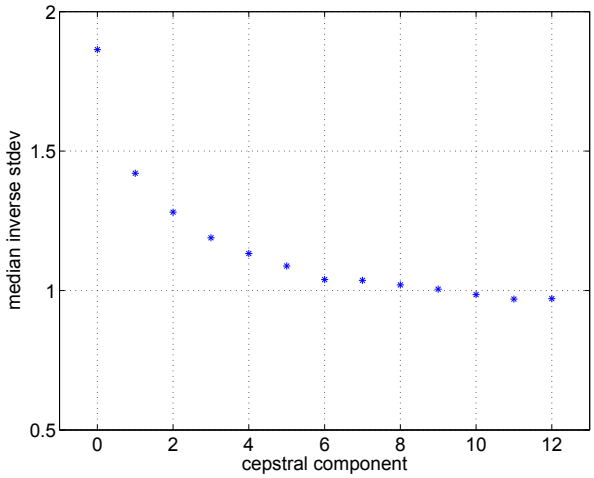

Fig. 1. Inverse standard deviation for $C0$-$C12$ components.

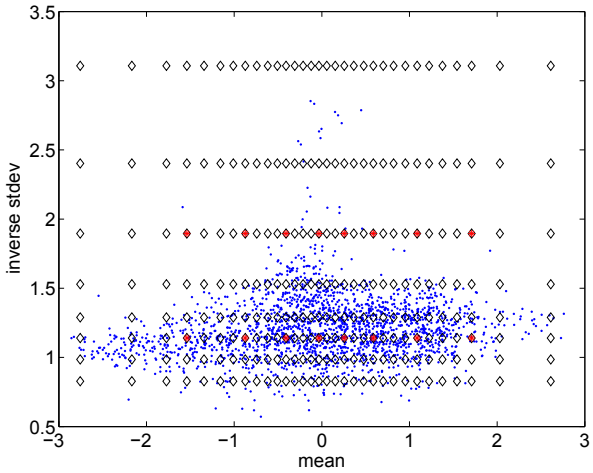

Fig. 2. Typical mapping of quantizers over the joint space.

In parameter quantization an optimized bit allocation scheme could take this into account with the aim of evenly balancing the errors induced by the parameter quantization into the evaluation of the state likelihood measure for each component (e.g. as in Equation 1). In a very simplistic model when considering Gaussian densities this results in balancing the variances for the random variables $\delta\mu_n(X_n-\mu_{nd})\sigma_{nd}^{-2}$ and $\delta\sigma_n^{-1}(X_n-\mu_{nd})^2\sigma_{nd}^{-1}$ which are obtained by differentiating the Mahalanobis distance. Here $\delta\mu_n$ is the quantization distortion for the mean quantizer in the n^{th} component and $\delta\sigma_n^{-1}$ the corresponding one for the inverse standard deviation. The distributions have a continuous part for the quantizer errors and the feature space X_n and a discrete one for the model parameters indexed by the density index d. There is a strong dependency relation since we focus on preserving the best performance for the cases where "accurate" likelihood scores matter most (e.g. high occupation probability for a density when given the feature vectors). This simple evaluation could be further extended. However, the effort of deriving optimal quantization rates can have little practical value due to the inherent limitations in the theoretical models used and due to design and implementation constraints (i.e. fractional rates are not feasible and a simple byte packing for the joint indexes would be desirable).

The design which is presented here selects a small subset of components to be high rate quantized, removes some of the empirically found ineffective components [1] and uses a half rate quantization for the remaining components. For a 39 dimensional feature space consisting of 13 normalized MFCC's and 1st and 2nd order time derivatives 8 components, namely; $C0$-$C3$, $\Delta C0$-$\Delta C2$ and $\Delta\Delta C0$ are selected for high rate and 8 components are removed (high order statics and most of the high order 2nd time order derivatives).

As in [1] sharing the same two scalar quantizers (for μ and σ^{-1}) across all dimensions is feasible and also desirable. With the multi-rate approach quantizers for each rate need to be stored. In the simplified case of only high and low rates with such large difference in allocated rates it may also be possible to pick the values for the low rate quantizer from the high rate quantizer, while storing them at the starting index positions. With this, again, only two quantizers are used in indexing and table evaluations. As will be shown in the experimental section, for the "strategic" rates of 5m3v (i.e. 5 bit means, 3 bit variances) and 3m1v for low rate, this approach worked satisfactory.

In addition, since the rate is halved for the low rate components, they can be grouped into pairs of dimensions for a combined indexing. With this, the effective working dimension for the state densities is halved resulting in good computational savings, as further described in the next section.

In Figure 2 a typical situation for such design is illustrated. The scatter plot of the mean and inverse standard deviation pairs is superposed on the high rate quantizer (black diamond centroids) from which the low rate quantizer values are selected (red stars).

2.3. Computational benefits

The computational benefits of quantization follows directly from the possibility of tabulating the most expensive computational part, the evaluation of state emission likelihoods. For instance, for states with mixtures of Gaussian densities, the state emission log likelihood formula is:

$$\log b(\mathbf{x}) = \log \sum_{k=1}^{K} \exp\left\{ \log\left(w_k \frac{1}{\prod_{i=1}^{N}\sqrt{2\pi\sigma_{ki}^2}} \right) - \sum_{i=1}^{N} \frac{(x_i-\mu_{ki})^2}{2\sigma_{ki}^2} \right\} \quad (1)$$

where K represents the number of densities in the mixture and N is the dimension of the feature vector space.

The Gaussian normalization factor is a constant with respect to the observed features therefore the most costly operation is the computation of the Mahalanobis distance. With quantization, for any given feature vector, each term of the summation can take only a limited range of values. For a typical rate of 5 bits for a mean component and 3 bits for a variance there are only $256 = 2^{5+3}$ distinct values which, when computed and tabulated in advance for each frame, will reduce the distance evaluation costs to an indexed summation from these tables.

[1] Which, in fact, is equivalent to a zero rate quantization.

With even lower rates the number of terms in the sum can be reduced by combining adjacent tables into a single one. E.g. with half the previous rate, combining two such tables results in the same number of distinct values but reduces the summation costs to half.

Computing the tables at each frame can be avoided if the feature vectors are also quantized [6]. In this case the entire state likelihood evaluation is reduced to table look-up and summation with no other overhead costs per frame.

2.4. Speaker adaptation

For a high performance speech recognition system for mobile devices an adaptation procedure is an integral part of the design. For mobile devices the effects of the uncertain environment can be mitigated against the fact that such devices are usually highly personal hence leading to good speaker adaptation potential.

As described in [8] a simple Bayesian adaptation approach provides both reduced complexity and good performance. The same procedure can be directly applied with the multi-rate quantization. In essence, model parameters are updated after each utterance by computing new values and quantizing them with the existing scalar quantizers which are left unchanged. In this context, the rather inefficient scalar quantization is, in fact, helpful by leaving more room for parameter adaptation.

3. EXPERIMENTS

3.1. Experimental setup

The scalar quantization algorithms were evaluated on a multi-lingual name dialling application. The recognizer is based on monophone HMMs with feature vectors consisting of recursively normalized Mel cepstral coefficients with appended 1st and 2nd order time derivatives [9]. The monophones cover 27 of the European languages. However, for a more focused presentation, only the figures for German and English are shown. To scale the difficulty of the task, three recognition grammars were created with sizes of about 100, 500 and 1000 name entries. The entries were a mixture of first names or both first and last names. Only slightly over 100 distinct names were present in the test sentences with a total of 11000 utterances for both languages combined [2]. The original recordings are denoted below with "clean" environment. A mismatched test condition was created (denoted in the following as "noise") by mixing various noise types with a randomly selected SNR value for each utterance. The SNR values were uniformly distributed from 5 to 20 dB. The original HMMs contain about 2000 Gaussian densities, trained from a multitude of databases, all recorded in noise free conditions. Except for the noise robust feature extraction procedure, there is no other noise compensation algorithm applied for these tests.

3.2. Experimental results

The results are presented in the Tables 1-4. Each line contains a grammar setup for a given language with bold lines representing the rates for the given grammar and the test set as a whole. The columns denote the test models; *orig* - the original models, *5m3v* - high rate quantized models, *3m1v* - low rate quantized models and *multi* - the multi-rate ones as described in the previous section.

[2]The actual counts are 4325 and 6698 for English and German, respectively.

Models / Grammar	*orig*	*5m3v*	*3m1v*	*multi*
uk100	1.39	1.55	2.36	1.97
ger100	0.67	0.73	0.94	0.78
Avg100	**0.95**	**1.05**	**1.50**	**1.25**
uk500	2.47	2.54	3.91	3.19
ger500	1.76	1.85	2.57	1.94
Avg500	**2.04**	**2.12**	**3.10**	**2.43**
uk1000	3.05	3.03	4.62	3.77
ger1000	2.18	2.22	3.15	2.54
Avg1000	**2.52**	**2.54**	**3.73**	**3.02**

Table 1. Error rates for speaker independent models.

3.2.1. Clean case

In "clean", which is a matched case to training conditions, we observe the expected result for both speaker independent (Table 1) and speaker adapted tests (Table 2). In spite of the masked components, the multi-rate models outperform significantly the low rate ones while having the same complexity requirements. The high rate models, at nearly double the costs, do provide better performance levels and these are very close to the ones of the originally trained models.

With speaker adaptation, the practical differences in performance diminish, while a similar performance ordering as before is observed.

Models / Grammar	*orig*	*5m3v*	*3m1v*	*multi*
uk100	0.39	0.39	0.65	0.51
ger100	0.28	0.27	0.46	0.33
Avg100	**0.32**	**0.32**	**0.53**	**0.40**
uk500	0.53	0.55	1.02	0.74
ger500	0.40	0.39	0.66	0.43
Avg500	**0.45**	**0.45**	**0.80**	**0.55**
uk1000	0.69	0.65	1.13	0.81
ger1000	0.43	0.43	0.73	0.46
Avg1000	**0.53**	**0.52**	**0.89**	**0.60**

Table 2. Error rates for speaker adapted models.

3.2.2. Noisy case

In the noisy case we observe an interesting phenomenon. Likely helped by the more constrained range of values for the low order cepstral coefficients, the *5m3v* and *3m1v* quantized models are outperforming even the original ones for nearly all the unadapted tests. The multi-rate models, perhaps slightly handicapped by the missing components, are a bit behind. Overall, in relative terms, the error differences are not very large making all the model sets about similar in practical use. This also illustrates that the acoustic mismatch due to the unseen testing environment and the inherent problems induced by noise dominate with respect to the other "noise" induced in the model parameters by quantization.

With speaker adaptation, the performance ordering reverts to the expected one with the exception of the smallest sized grammar where the multi-rate models are still slightly behind, perhaps also as result of the proportionally lower initial performance point. For

Models Grammar	*orig*	*5m3v*	*3m1v*	*multi*
uk100	10.96	10.31	10.61	11.79
ger100	8.24	7.78	7.81	8.29
Avg100	**9.31**	**8.77**	**8.91**	**9.66**
uk500	15.12	14.64	14.50	15.38
ger500	13.05	12.48	13.15	13.81
Avg500	**13.86**	**13.33**	**13.68**	**14.43**
uk1000	16.92	16.32	16.39	17.20
ger1000	14.94	14.17	15.39	15.77
Avg1000	**15.72**	**15.01**	**15.78**	**16.33**

Table 3. Noise condition error rates for speaker independent models.

the larger grammars, in comparison with *3m1v* we can observe a more precise adaptation for the multi-rate models which end up with slightly better figures in spite of starting from a worse position.

Models Grammar	*orig*	*5m3v*	*3m1v*	*multi*
uk100	4.83	4.83	5.62	6.24
ger100	3.14	3.51	4.03	4.03
Avg100	**3.80**	**4.03**	**4.65**	**4.90**
uk500	5.62	5.78	7.24	7.03
ger500	4.08	4.45	6.23	5.55
Avg500	**4.68**	**4.97**	**6.63**	**6.13**
uk1000	6.36	6.71	8.23	7.70
ger1000	4.88	5.02	7.20	6.54
Avg1000	**5.46**	**5.68**	**7.60**	**7.00**

Table 4. Noise condition error rates for speaker adapted models.

3.2.3. Final observations

In an overall comparison it is readily apparent the significant boost that adaptation can bring in performance. The relative quantization performances are visible for the clean environment with a clear superiority of the high rate models, followed by the multi-rate ones. In noisy environments the performance differences are no longer very important, with an exception for speaker adaptation where the multi-rate models have a slight edge.

For quantization the compression factor can be quickly estimated as original bit-rate for a parameter pair [3] over the 8 bits or 4 bits rates required by the quantization methods proposed. However, there are no specific experimental numbers presented in support of the computational complexity savings since these are highly implementation dependent. A theoretical evaluation of the number of arithmetic operations required is not difficult and an example is given in [1].

[3] this could be 32 bits or even 16 bits for the usual fixed point implementations

4. CONCLUSION

For CDHMMs based speech recognizers scalar quantization provides a simple yet valuable approach towards memory and computational complexity reductions. If the design imposes strong memory limitations and, therefore, the quantization must be done at very low bit-rates, it is more effective to consider a multi-rate quantization approach. Although the performance differences among various quantization options are rather marginal for mismatched conditions, in better matched cases, the multi-rate approach can offer improved performance and better adaptation potential for same costs as a single rate quantization solution.

5. REFERENCES

[1] Vasilache M., "Speech recognition using HMMs with quantized parameters," in *ICSSLP'2000*, 2000, pp. 871–874.

[2] Mak B. K.-W. and Bocchieri E., "Subspace distribution clustering hidden Markov model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 264–275, March 2001.

[3] Mak B. K.-W. and Bocchieri E., "Direct training of subspace distribution clustering hidden Markov model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 378–387, May 2001.

[4] Kiss I. and Vasilache M., "Low complexity techniques for embedded ASR systems," in *Eurospeech-Scandinavia*, 2002, pp. II.1265–1268.

[5] Leppanen J. and Kiss I., "Comparison of low footprint acoustic modeling techniques for embedded ASR systems," in *Interspeech-Eurospeech*, 2005, pp. 2965–2968.

[6] Vasilache M., Iso-Sipila J., and Viikki O., "On a practical design of a low complexity speech recognition engine," in *ICASSP*, 2004, pp. V.113–116.

[7] Zheng-Hua Tan and Børge Lindberg, Eds., *Automatic Speech Recognition on Mobile Devices and over Communication Networks*, chapter 11, Springer-Verlag, February 2008.

[8] Vasilache M. and Viikki O., "Speaker adaptation of quantized parameter HMMs," in *Eurospeech-Scandinavia*, 2001, pp. II.1265–1268.

[9] Viikki O., Bye D., and Laurila K., "A recursive feature vector normalization approach for robust speech recognition in noise," in *ICASSP'98*, Seattle, USA, May 1998, pp. 733–736.