Speaker Adaptation of Quantized Parameter HMMs

Marcel Vasilache and Olli Viikki

Speech and Audio Systems Laboratory
Nokia Research Center, Tampere, Finland
{marcel.vasilache, olli.viikki}@nokia.com

Abstract

This paper extends the evaluation of Hidden Markov Models with quantized parameters (qHMM) presented in [5] to the case of speaker adaptive training. In speaker-independent speech recognition tasks, qHMMs were found to provide a similar performance as the original continuous density HMMs (CDHMM) with substantially reduced memory requirements. In this paper, we propose a Bayesian type of adaptation framework for qHMMs to improve the speaker-specific acoustic modeling accuracy. Experimental results indicate that the proposed qHMM adaptation scheme provides a comparable performance as obtained with the Bayesian adaptation of CDHMMs in a noise-free test environment. In the presence of noise, on the other hand, the performance improvement due to qHMM adaptation is lower than obtained in the CDHMM case. In general, the adaptation gains are on a similar scale fact that confers to qHMMs a great practical value.

1. Introduction

Over the recent few years, Automatic Speech Recognition (ASR) technology has widely been applied to various platforms equipped with different implementation resources. For portable devices, memory and computational resources may set severe performance limitations on the speech recognition engine. The complexity limit dictates the accuracy of acoustic modeling and the size of the recognition vocabulary. Due to the large popularity of small hand-held devices, it is increasingly important to design and modify speech recognition algorithms such that they can be run on resource-scarce embedded platforms.

Since the memory consumption has a big effect on the factory price of all mass-produced devices, it is important to attempt to minimize the memory usage of the algorithms running on these systems. In speech recognition, acoustic modeling requires a substantial amount of memory and algorithm simplifications can hence result in considerable savings in the manufacturing price. Recently, the realization of compact acoustic models has turned out to be an active research topic in the ASR community. In [3][4], good recognition results with very compact acoustic models were obtained by replacing Gaussian distributions in HMMs with neural networks. In [7], the authors evaluated a vector quantization scheme with a modified *k*-means clustering algorithm in the context of large vocabulary speech recognition. In [5], we proposed a quantization scheme for the parameters of continuous density Hidden Markov Models. The proposed novel approach, denoted as qHMMs, was found to reduce substantially the memory consumption while simultaneously maintaining the high recognition performance obtained with CDHMMs.

It is apparent that the low complexity should not be realized at the expense of the performance, but the same recognition accuracy should be achieved as done with the conventional CDHMMs. In speaker-independent speech recognition, it is customary to use acoustic model adaptation techniques to maximize the speaker-specific recognition accuracy [1][2]. In this paper, we extend the qHMM framework to include the possibility to carry out parameter adaptation. The proposed new adaptation algorithm for qHMMs achieves similar adaptation performance gains as obtained when applying the widely used Bayesian adaptation algorithm to the parameters of continuous density HMMs.

The remainder of the paper is organized as follows. Chapter 2 outlines the Bayesian adaptation principle. In Chapter 3, the framework for qHMMs is presented. Experimental results are given in Chapter 4 and the main conclusions are formulated in the end.

2. Bayesian Adaptation

A large number of adaptation methods for CDHMMs are now well established and can be used in compensating training/testing mismatches induced both by speaker characteristics and/or by the operating environment of the recognizer. Up to now, Bayesian [1] and Maximum Likelihood Linear Regression (MLLR) [2] adaptation have widely been investigated and applied to HMM based ASR by many researchers. In the following, we concentrate on Bayesian adaptation due to its effectiveness, nice asymptotic convergence properties, and flexibility of implementation. One potential problem of Bayesian adaptation, namely a reduced performance gain with a limited amount of observed data, was considerably diminished in our case as acoustic modeling was based on a small number of context-independent monophone HMMs.

Bayesian adaptation of the mean and variance coefficients of CDHMMs is described by the following two equations.

$$\mu_k' = \frac{\tau\mu_k + \sum_{t=1}^{T}\gamma(t,k)\mathbf{x}_t}{\tau + \sum_{t=1}^{T}\gamma(t,k)} \quad (1)$$

$$\sigma_k'^{\,2} = \frac{\tau(\sigma_k^2 + \mu_k^2) + \sum_{t=1}^{T}\gamma(t,k)\mathbf{x}_t^2}{\tau + \sum_{t=1}^{T}\gamma(t,k)} - \mu_k'^{\,2} \quad (2)$$

We use the following notations: μ and σ^2 for the means and variances of the initial speaker-independent Gaussian densities, $\mathbf{x}$ for the observation vectors, $\gamma(t,k)$ for the occupation probability function with parameters: t the frame

number and *k* a generic density index. With μ' and σ'^2, we denote the adapted means and variances.

The parameter τ weighting the speaker-independent parameters is used to control the adaptation speed. In practice, the value for this parameter must be carefully handled. Too small values of τ can degrade the recognition performance, especially in the cases when the adaptation procedure is not supervised.

3. qHMM Framework

3.1. Structure

For representing a CDHMM, real valued parameters are necessary. These parameters usually have a fixed or floating point representation with the size a multiple of bytes. Due to its statistical nature, the error level in the parameter estimation procedure can be significantly higher than the error of parameter representation. For memory efficiency, this fact can be exploited with a more constrained parameter space representation. In addition, if the classifier has a wide margin (i.e. separation power between the correct and the best incorrect hypothesis), this can also be utilized for even higher parameter space compression.

As presented in [5], qHMMs are built starting from continuous density HMMs (CDHMM) by quantizing the mean and variance parameters. We focused on CDHMMs with Gaussian densities having diagonal covariance matrices. In parameter quantization, only mean and variances were of our primary interest because their number exceeds with orders of magnitude the remaining CDHMM parameters. Only two scalar quantizers, one for the means and the other for the variances were used. This is possible by applying a global normalization of the feature space such that zero mean and unity variance feature components are presented to the recognizer. For maximal efficiency, using a non-linear Lloyd-Max type of quantization, is essential.

3.2. Advantages of qHMM Representation

Besides the obvious memory savings, by quantization it is also possible to reduce significantly the number of floating-point operations required in computing the observation probability values. This saving is done by precomputing for every frame all the possible outcomes of the terms $(x_i - \mu_j)^2 / 2\sigma_k^2$, where x_i denotes the *i*th component of the feature vector and μ_j, σ_k^2 are the mean and variance corresponding to quantizer indexes *j* and *k*, respectively. We store these values in a table for all level combinations of mean and variance quantizers. In the second stage, the observation probabilities are obtained with an indexed summation of the values from the precomputed tables. Since means and variances are grouped in pairs, a single, joint index can be used. It can be estimated that if the product of mean and variance quantization levels is small compared to the number of Gaussian densities to be evaluated, the number of floating point operations is reduced by almost 75%. However, in practice, the run time gain is heavily dependent on the hardware platform to be used. For such architectures, where multiplications and additions can be executed in parallel and/or integer indexing is expensive, the gains are significantly more modest.

3.3. Training Algorithm

The training of qHMMs must produce the optimal density means and variances for the given classification task. Due to the quantization, the limited selection of quantizer levels gives a strong optimization constraint.

A simple and straightforward training method consists of the following steps: First, the CDHMM models are trained with a conventional training procedure. Second, based on the parameters obtained for the models, the two Lloyd-Max scalar quantizers are created according to the targeted quantization rates. In the final training step, the model parameters are quantized. When training the quantizers the Euclidean distance was selected as distortion measure for mean values. Variance quantization was done as well with the Euclidean distortion measure but applied to the inversed standard deviation values.

In practice, we found that a considerably more complex joint optimization of model parameters and quantizers during CDHMM training resulted in similar performance as with the simple procedure. Since in evaluating the required compression level for the parameter space it is useful to explore a wider range of quantization rates, the simple procedure enables this at minimal costs.

3.4. Speaker Adaptation Algorithm of qHMMs

In speaker adaptation, the parameter update was done in a similar, straightforward fashion as in training. Due to the qHMM structure, the scalar quantizers have the advantage of a more uniform coverage of the parameter space which allows for less constrained adaptation updates compared with the case of applying vector quantization to the multi-dimensional Gaussian densities. For updating a density, the new means and variances are first computed according to Equations (1) and (2) and then quantized with the corresponding mean and variance quantizers.

In Equations (1) and (2), it is visible that the adaptation algorithm requires storing the terms $\sum_{t=1}^{T}\gamma(t,k)$, $\sum_{t=1}^{T}\gamma(t,k)\mathbf{x}_t$ and $\sum_{t=1}^{T}\gamma(t,k)\mathbf{x}_t^2$. The last two terms require having, essentially, a new set of parameters with the size of the original models. This can be prohibitive in terms of memory usage, considerably reducing the memory gains of quantizing the speaker-independent models. The solution is to update the original models after each recognized utterance. In this case, the adaptation can be done using, for the parameter space, a temporary storage with the size requirements of a single density. Therefore, the main memory overhead consists only of storing the features to be used in adaptation.

4. Experiments

4.1. Setup

For the experiments, we used a speaker-independent, phoneme based, isolated word recognition engine. We had a conventional front-end based on FFT derived Mel cepstral coefficients and their first and second order time derivatives. On top of this, to improve the noise robustness of the

computed parameters, we employed a normalization procedure similar to the one presented in [6].

The recognition vocabulary consisted of 120 names. In the test database, there were recordings from 50 male and 32 female speakers, 11,991 utterances for males and 7,676 for females. Each speaker pronounced the set of 120 names twice. The recognition models were built using monophone HMMs that were trained according to the Maximum Likelihood (ML) criterion on a different database with phonetically rich sentences. For training, only noise free utterances were used. The final HMMs had 3 states with mixtures of 10 Gaussian densities each. Based on these models, we created a collection of qHMMs using the simple procedure outlined in Section 3.3.

For testing the recognition performance in a noisy environment, the clean test data was artificially corrupted with various types of noises at random SNR values. The noise types were: car noise, cafeteria noise, music and car noise with ambient speech. The SNR values were set using a uniform random variable that ranged from 5 dB to 20 dB.

4.2. Speaker-Independent Performance

The first two tables show the speaker-independent recognition performance for various quantization rates[1]. The row "orig" shows the performance obtained with the original IEEE 32-bit floating-point representation of the model parameters.

The first column lists the quantization rate parameters. Table 2 summarizes the recognition performance with the same speakers in the presence of noisy environment. It is seen from Tables 1 and 2 that a good performance can be obtained with the small numbers of bits. Even with as low as 2 bits (2m0v), the recognition engine provides a reasonable recognition performance. With more than 6 bits, there are practically no differences to the original models. It is also worth noting that only a few quantization levels are needed for variance parameters.

Model Set	Men	Women	Average
1m1v	30.84	28.87	30.07
2m0v	4.64	3.49	4.19
2m1v	3.26	2.16	2.83
2m2v	2.18	1.20	1.80
3m1v	2.83	1.60	2.35
3m2v	1.88	0.86	1.48
4m2v	1.71	0.78	1.35
5m3v	1.65	0.74	1.29
6m4v	1.63	0.70	1.27
7m5v	1.59	0.76	1.27
orig	1.64	0.77	1.30

Table 1: Speaker-independent word error rates, noise-free environment.

[1] The notation "4m2v" denotes that 4 bits were assigned for the mean quantizer and 2 bits for the variance quantizer, respectively.

Models	Men	Women	Average
1m1v	73.76	72.49	73.26
2m0v	18.08	14.96	16.86
2m1v	19.19	16.64	18.19
2m2v	19.79	18.00	19.09
3m1v	16.47	13.99	15.50
3m2v	16.55	14.66	15.81
4m2v	16.23	14.10	15.40
5m3v	15.82	13.90	15.07
6m4v	15.79	14.32	15.22
7m5v	15.86	14.16	15.20
orig	15.89	14.13	15.20

Table 2: Speaker-independent word error rates, noisy environments.

4.3. Speaker Adaptive Performance

For speaker adaptation, only the mean parameters were updated. With our preliminary experiments, we found that variance adaptation had a negligible effect on the recognition performance, but it often caused problems when aggressive adaptation parameters were used (i.e. small adaptation batch and a small value for the τ parameter). Adapting only the means, on the other hand, resulted in a stable performance improvement even for very aggressive settings. However, to succeed in fast adaptation, it was mandatory to have a supervised adaptation process.

In order to reduce the computation costs, a segmental adaptation procedure was used (i.e. zero/one state occupation function given by the Viterbi “best-path” algorithm). As mentioned in Section 3.4, for reducing the memory requirements, we avoid storing long-term accumulators by adapting the models after every utterance. The value of the parameter τ was set to 10.

The performance after speaker adaptation is illustrated by Tables 3 and 4. The speaker-independent (SI) results are again visible in the second column for an easier performance comparison. Under heading "SA" are presented the word error rates after adaptation. The last column gives the relative error rate reductions obtained thanks to adaptation.

Models	SI	SA	Gain [%]
1m1v	30.07	25.49	15
2m0v	4.19	1.45	65
2m1v	2.83	1.43	49
2m2v	1.80	1.02	43
3m1v	2.35	0.57	76
3m2v	1.48	0.54	63
4m2v	1.35	0.38	72
5m3v	1.29	0.39	70
6m4v	1.27	0.36	72
7m5v	1.27	0.34	73
orig	1.30	0.23	83

Table 3: Speaker adaptation performance, noise-free environment.

Models	SI	SA	Gain [%]
1m1v	73.26	70.18	4
2m0v	16.86	9.72	42
2m1v	18.19	12.55	31
2m2v	19.09	14.59	24
3m1v	15.50	7.64	51
3m2v	15.81	9.68	39
4m2v	15.40	7.71	50
5m3v	15.07	8.24	45
6m4v	15.22	8.55	44
7m5v	15.20	8.71	43
orig	15.20	6.47	57

Table 4: Speaker adaptation performance, noisy environments.

While qHMM adaptation in the clean conditions provides in the absolute scale a similar performance as obtained with adapting CDHMMs, the adaptation gain in the noisy conditions is not as high.

4.4. Log-Likelihood Ratio Measurements

Since recognition rates are more vulnerable to statistical fluctuations, we have derived an additional performance test. In this, we measured for all recognitions the normalized log-likelihood ratio (NLLR) of the correct recognition hypothesis versus the best scoring incorrect hypothesis. The normalization was done according to the number of speech frames contained in the utterance. The results for this test are summarized in Table 5.

In the light of these figures, it is now visible that NLLR increases as a function of the total quantization rate. Although we have not explored all possible combinations, one can assert that optimal results at low rates require a higher resolution for the mean quantizer.

When comparing against the figures for the original models, one can observe a saturation zone starting near the 5m3v point. For the SI models, the saturation value corresponds to the performance of the original models and we can conclude that this level of precision is enough for this particular recognition task. However, for the SA models, a similar saturation is observed but at a level significantly lower. These findings correlate well with the recognition rates.

	CLEAN		NOISE	
Models	**SI**	**SA**	**SI**	**SA**
1m1v	0.332	0.394	0.022	0.052
2m0v	1.126	1.467	0.610	0.813
2m1v	1.278	1.561	0.675	0.838
2m2v	1.440	1.650	0.748	0.867
3m1v	1.499	2.096	0.798	1.152
3m2v	1.683	2.145	0.880	1.148
4m2v	1.739	2.385	0.904	1.286
5m3v	1.804	2.391	0.938	1.285
6m4v	1.835	2.383	0.951	1.273
7m5v	1.838	2.376	0.952	1.270
orig	1.839	2.862	0.952	1.596

Table 5: Average normalized log-likelihood ratios.

5. Conclusions

In this paper, we have proposed a Bayesian-like adaptation scheme for quantized parameter HMMs. This important extension to the qHMM framework substantially improves the applicability of this novel acoustic modeling approach in speaker-independent ASR applications. Compared with conventional continuous density HMMs, qHMMs provide considerable memory reductions while still maintaining the same recognition rate. The compactness of the acoustic modeling module is particularly important in embedded systems where one needs to cope with limited implementation resources. The adaptive framework of qHMMs also produces a comparable high performance to that of adapting CDHMM parameters in clean conditions. In the presence of noise, despite the impressive performance gains with respect to the speaker-independent qHMM system, the improvements are not as high as when adapting the parameters of CDHMMs.

6. References

[1] Gauvain, J.-L, Lee, C-H., "Maximum a posteriori estimation of multivariate Gaussian mixture observations of Markov chains", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298, 1994.

[2] Leggetter, C. J., Woodland, P., C., "Speaker adaptation of HMMs using linear regression", *Proc. of International Conference on Spoken Language Processing*, Yokohama, Japan, 1994.

[3] Riis, S. K., "Hidden neural networks: application to speech recognition", *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, USA, 1998.

[4] Riis, S. K., Viikki, O., "Low complexity speaker-independent command word recognition in car environments", *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000.

[5] Vasilache, M., "Speech recognition using HMMs with quantized parameters", *Proc. of International Conference on Spoken Language Processing*, Beijing, China, 2000.

[6] Viikki O., Bye D., and Laurila K., "A recursive feature vector normalization approach for robust speech recognition in noise", *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, USA, 1998.

[7] Pan J., Yuan B. and Yan Y., "Effective vector quantization for a highly compact acoustic model for LVCSR" , *Proc. of International Conference on Spoken Language Processing*, Beijing, China, 2000.