Generating Small, Accurate Acoustic Models with a Modified Bayesian Information Criterion

Kai Yu, Rob A. Rutenbar

Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA 15213, USA
{kaiy,rutenbar}@ece.cmu.edu

Abstract

Although Gaussian mixture models are commonly used in acoustic models for speech recognition, there is no standard method for determining the number of mixture components. Most models arbitrarily assign the number of mixture components with little justification. While model selection techniques with a mathematical derivation, such as the Bayesian information criterion (BIC), have been applied, these criteria focus on properly modeling the true distribution of individual tied-states (senones) without considering the entire acoustic model; this leads to suboptimal speech recognition performance. In this paper we present a method to generate statistically-justified acoustic models that consider inter-senone effects by modifying the BIC. Experimental results in the CMU Communicator domain show that in contrast to previous strategies, the new method generates not only attractively smaller acoustic models, but also ones with lower word error rate.

Index Terms: acoustic model training, model selection, BIC, Gaussian mixture models

1. Introduction

It is common for acoustic models to represent tied-triphone states (or *senones*) with Gaussian mixture models (GMMs). For each time frame a feature vector composed of acoustic features is computed, and this is then used to calculate senone probabilities. Generating an acoustic model of a proper size is important for speech recognition performance. If too many Gaussian mixture components are used, the acoustic model will overfit the training data and require unnecessary computational time. On the other hand, if too few mixture components are used, there will not be enough resolution to model the acoustic complexity or differentiate between senones.

While the use of GMMs is widespread, there is no consensus on how to determine the optimal number of mixture components. To assign a number of mixture components to a particular senone, popular methods include setting it to an arbitrary constant or a numerical fraction of its instances in the training set [1]. However, there is no statistical justification for these methods, and the acoustic complexity of individual senones is not considered. Another approach is to use *model selection* techniques derived in statistics [2], such as the *Bayesian information criterion* (BIC) [3], which focus on modeling each individual senone accurately. As previously noted [4], this does not consider the senone within the context of the entire acoustic model and can lead to senone models that encroach onto the space of other senones. Discriminant measures to account for interactions between senones have been proposed [4] but they do not account for acoustic complexity.

In this paper we introduce a method to generate GMMs whose number of mixture components takes into account both acoustic complexity and inter-senone effects. We take the BIC and add a senone-specific prior on the complexity penalty inversely proportional to how often other senones encroach on its space. Biem [5] also proposed a change to BIC to consider the entire acoustic model, but it was used for classifying handwritten digits, a simpler problem with fewer states. We test our method in the Communicator domain [6] with CMU's Sphinx 3.0 decoder [7]. Our results show that the proposed method generates an acoustic model that has the lowest word error rate compared to other common methods and only slightly more mixture components used than BIC. We also explain why this method allocates mixture components more effectively.

The organization of the paper is as follows. In section 2 we describe current techniques to generate GMM-based acoustic models, and in section 3 we present our proposed method. Experimental results and analysis are reported in section 4, and conclusions in section 5.

2. Acoustic model generation methods

After state-tying to cluster the triphone states into senones, the free parameters of a GMM-based acoustic model are the means, variances, mixture weights, and the number of mixture components. The means, variances, and mixture weights are usually calculated using the Baum-Welch algorithm or segmental K-means algorithm [8]. Determining the number of mixture components is much more difficult, and the following are the most common methods.

2.1. Arbitrary Constant

In many acoustic models the number of mixture components per senone is set to an arbitrary constant. This is usually achieved by creating an initial model with one mixture component, then iteratively splitting the mixture component with the largest mixture weight and re-estimating the model parameters until the target number of mixture components is reached [9]. This method is the simplest, but it disregards the acoustic complexity of individual senones.

2.2. Proportional to training samples

Some systems set the number of mixture components per senone to be proportional to the number of training set frames assigned to the senone, up to a specified maximum [1]. Using this method, the number of mixture components for senone X, denoted as c_X, will be

$$c_X = \min\left(n_X / D, MCS\right), \tag{1}$$

where n_X is the number of frames assigned to senone X in the training set, D is the minimum number of training samples required to train a mixture component, and *MCS* is the maximum number of mixture components per senone. D is often related to the dimensionality of the feature vector. This method requires first using an initial model to segment the training data into the optimal sequence of senones, usually through Viterbi segmentation. If O_X is all the frames assigned to senone X, then n_X is simply the size of O_X. While this method ensures each mixture component has sufficient training data, the number of times a senone appears in the training data does *not* necessarily correlate with acoustic complexity.

A related method [4] employs a discriminant criterion to choose MCS, which is summarized here as follows. The number of mixture components is proportional to the training set instances, but there are two different values for *MCS* depending on how "aggressive" the senone is. The aggressiveness of senone X is computed as the average ratio of the likelihood of senone X divided by the sum of the likelihood of senone X and other senones over O_X. If this ratio is high, the model of senone X captures the space well. However, if the ratio is low, the model for senone X is not aggressive enough and other senones have encroached its space. Using this ratio the senones are classified into either being aggressive or non-aggressive, and the non-aggressive senones use a larger *MCS* value.

2.3. Model Selection

Model selection aims to select the GMM whose dimensionality best represents the true distribution. Although the true distribution is not known, by assuming the observations are consistent with the underlying distribution, model selection provides a methodical way to maximize the likelihood of the training data and simultaneously avoid overtraining. Many model selection criteria have been proposed in the statistics literature [3], and the most common one used in speech recognition is the Bayesian information criterion (BIC):

$$BIC(\theta_X^j) = \log p(O_X \mid \theta_X^j) - \frac{1}{2} f(\theta_X^j) \cdot \log(n_X), \qquad (2)$$

where θ_X^j is the jth model for senone X, $p(O_X \mid \theta_X^j)$ is the likelihood of the data assigned to senone X given θ_X^j, and $f(\theta_X^j)$ is the number of free parameters in θ_X^j. The BIC is the sum of the log likelihood of the data and a complexity penalty term, and the model with the largest BIC score will be selected. Other information criteria used for model selection also include the log likelihood of the data as one term but differ on how to penalize for model complexity. The BIC compares different models to represent the same senone and requires generating GMMs with different numbers of mixture components. In some implementations of BIC for acoustic model generation, the maximum number of mixture components per senone is bounded [8].

A regularization parameter λ was first introduced for speech recognition in [2], yielding the form:

$$BIC(\theta_X^j) = \log p(O_X \mid \theta_X^j) - \frac{\lambda}{2} f(\theta_X^j) \cdot \log(n_X). \qquad (3)$$

This added factor helps correct for the fact that the data used to train the model may not perfectly represent the acoustic space. The same value of λ is used for all senones. While original ($\lambda = 1$) BIC may not always produce the model with the lowest WER [4], there is no technique to decide *a priori* what value of λ will minimize the WER. A more practical goal for using the regularization term is to adjust the size of the acoustic model. Since there are no free parameters in (2), only one acoustic model can be generated. For applications that are resource-constrained, like embedded speech recognition, adding the λ term in (3) gives a mechanism to generate smaller acoustic models.

When evaluating senone models, BIC only considers an individual senone's acoustic complexity, not its encroachment effect on other senones. This means the chosen model may score both its data and other senone's data with a high likelihood, which negatively impact the word error rate. For example, consider the decoding of a test sentence by two acoustic models in Table 1. Model B has a higher correct sentence likelihood than model A because it more accurately represents each senone. However, in terms of WER, model B is worse than model A because it will decode an incorrect sentence since it has a higher likelihood than the correct sentence. This simple example illustrates that focusing on accurately modeling *individual* senones is not sufficient; the effects of *other* senones also need to be considered.

Acoustic Model	Likelihood of Correct Sentence	Maximum Likelihood of an Incorrect Sentence
A	0.20	0.10
B	0.50	0.60

Table 1. *Example models and likelihoods of a single test sentence.*

Overall, BIC models the acoustic complexity and provides a statistical justification for the number of mixture components. Also, there is at most one free parameter, λ, in BIC for the user to arbitrarily assign. However, when creating models for individual senones, the method only compares intra-senone models without considering the interaction between senones. Inter-senone effects are important because they influence speech recognition performance, so using BIC does not lead to optimal WER.

3. Proposed model selection method: mBIC

Ideally, the acoustic model generated should select numbers of mixture components that are statistically justified and also account for inter-senone effects. To do this, we propose to modify the BIC complexity penalty term to account for the effect from other senones, resulting in the following modified BIC (*mBIC*):

$$mBIC(\theta_X^j) = \log p(O_X \mid \theta_X^j) - \frac{1}{2} k_X \cdot f(\theta_X^j) \cdot \log(n_X), \qquad (4)$$

where we define k_X as the *inter-senone correction factor*. Unlike λ, a *unique* k_X is calculated for *each* senone. The factor k_X can be interpreted as a prior probability proportional to how heavily complexity should be penalized, and its value should fall between 0 and 1. When k_X is equal to 1, mBIC simplifies to BIC.

If the space of senone X in frequently encroached by other senones, k_X should be close to 0. Reducing the penalty on complexity will result in a model with higher likelihood for O_X and decrease the effects from other senones. Conversely, if senone X is rarely affected by other senones, then the original BIC formulation is sufficient and the correction factor should be close to 1. Since the correction factor can only reduce the complexity penalty and never increase it, the

acoustic model generated using mBIC will always be larger than that using BIC.

We define k_X to be the fraction of frames in O_X that originate from *incorrectly*-decoded training set sentences and where senone X has a *lower* likelihood than the senone force-aligned to the frame using the incorrectly-decoded sentence. Only frames from incorrectly-decoded sentences are used because they have the most significant impact on the word error rate. A more complex correction factor could use a ratio of the likelihoods, consider correctly-decoded training set sentences, or factor in how often a senone encroaches on the space of other senones, but we find that our simple definition was sufficient to improve performance.

To generate the acoustic model using mBIC, we follow these steps:

1. Generate an acoustic model by segmenting the training data. Create models with different number of mixture components. Select models that maximize the BIC for each senone.
2. Decode the training set with the new acoustic model.
3. Force-align all incorrectly-decoded sentences with both the correct transcript and the incorrect sentence transcript.
4. For each senone, count the number of frames assigned to it by the correct transcript where it has a lower likelihood than the senone assigned by the incorrect sentence transcript. Divide this by the total number of frames assigned to it through force-alignment with the correct transcript on the entire training set to compute the correction factor.
5. Using the models generated in step 1 and the correction factors from step 4, use mBIC to select new mixture component numbers and generate the final acoustic model.
6. Re-estimate the means, variances, and mixture weights with EM.

The idea is to augment BIC (steps 1 and 6) with four intermediate steps that include inter-senone effects. Note that the models generated in step 1 for BIC do not depend on the model selection method used, so they can be reused in step 5. This is important because creating models with different numbers of mixture components is by far the most time-consuming step. The proposed correction factor can also be used to modify the complexity penalty term of other model selection methods. Also, as in BIC a regularization factor could be added to obtain

$$mBIC(\theta_X^j) = \log p(O_X \mid \theta_X^j) - \frac{\lambda}{2} k_X \cdot f(\theta_X^j) \cdot \log(n_X) \cdot \quad (5)$$

In summary, the proposed mBIC combines the strengths of the discriminant criterion [4] and BIC, without their weaknesses. The correction factor is similar to the "aggressiveness" factor in the discriminant criterion, but our method adds mixture components to senone models in a justified, systematic way. Our method builds upon the form of BIC, but mBIC considers the entire acoustic model and inter-senone effects on speech recognition performance.

4. Experimental results

Multiple acoustic model generation methods were tested using the 2001-word Communicator corpora [6]. The training set contains 120,185 utterances totaling slightly less than 68 hours. A frame size of 10 ms and sliding window of 250 frames was used, so that the training set had a total of 24.4 million frames. All acoustic models trained had triphones clustered into 2165 senones, and the feature vector was 39 elements composed of 12 mel frequency cepstral coefficients, the log energy, and their first and second derivatives. The speech recognizer used was Sphinx 3.0 [7], a flat lexical decoder. We use WER to measure speech recognition performance.

We compare acoustic models generated by the following methods:

- Baseline method with arbitrary constant of 32 mixture components for all senones.
- BIC with no regularization and *MCS* of 32.
- mBIC with no regularization and *MCS* of 32.
- BIC with regularization such that the acoustic model generated is the same size as the mBIC one, and a *MCS* of 32. This corresponds to $\lambda = 0.9765$.
- Arbitrary constant of 16 mixture components for all senones.

We do not evaluate a model where the number of mixture components is proportional to the number of training set frames because it would have only been 1% smaller than the baseline model when using $D = 39$. Since only Gaussian mixture components with diagonal covariance matrices were considered, there are 79 free parameters per mixture component (1 mixture weight, 39 means, and 39 variances). We use a *MCS* of 32 for all the methods so none of the methods would be able to generate a model larger than the baseline model. The average number of components per senone, WER, and change in WER for the different models are listed in Table 2.

Method	Avg. Comp/Senone	WER	ΔWER
Constant (Baseline)	32	14.95	-
BIC	21.06	15.29	+ 2.3%
mBIC	**21.40**	**14.84**	**- 0.7%**
BIC, $\lambda = 0.9765$	21.40	15.10	+ 1.0%
Constant	16	18.14	+ 21 %

Table 2. *Comparison of acoustic model generation methods.*

The proposed mBIC results in a model with not only the *lowest* WER, but it also uses 33% fewer mixture components compared to the baseline model. The mBIC model is only slightly larger than the BIC model, but it has a 3% better relative WER. To show that the improvement is due to the algorithm and not the increased number of components, we generated a regularized BIC acoustic model to be the same size as our mBIC model (second to last row of Table 2); as can be seen, mBIC still performs better.

4.1. Comparison of mBIC and other methods

In this section we will compare mBIC against the other presented methods and try to explain why mBIC produces a better acoustic model.

4.1.1. Arbitrary constant

Properly representing acoustic complexity is clearly quite important. Using an *ad hoc* arbitrary constant blindly allocates the same number of mixture components to all

senones, which results in overtrained models and increases both the amount of errors and computational time. Both BIC and mBIC are able to generate much smaller models with nearly as good or better performance than using an arbitrary constant.

4.1.2. Proportional to training samples

This method assumes the acoustic complexity is proportional to the number of training set instances per senone. To check this, in Figure 1 we plot the number of mixture components selected by BIC versus the number of training set examples per senone. We use the BIC results because the number of mixture components is only dependent on the acoustic complexity. For better resolution, we only display the senones with less than 45,000 training set instances. We also plot the line representing $D = 500$ as a reference.

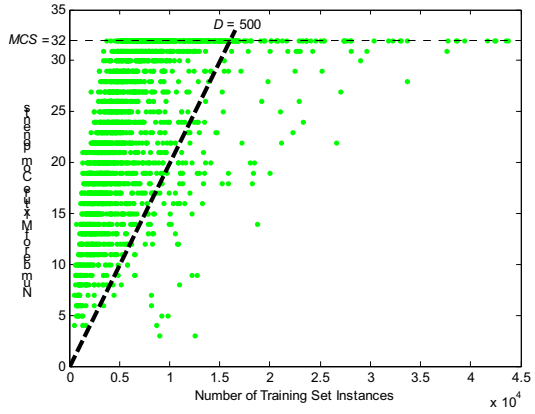

Figure 1: *Plot of number of mixture components as a function of training set instances per senones trained using BIC, and a line for D = 500.*

In an acoustic model trained with mixture components proportional to training set instances using $D = 500$ and $MCS = 32$, the number of mixture components will be the minimum of the value on the dotted line and 32. For all the senones that are above the dotted line, too few mixture components will be used. On the other hand, the senones that fall below the dotted line and use less than 32 mixture components will be assigned too many mixture components. Clearly this does *not* model the acoustic complexity well, and any line with a different D value would not be able to do so either. This method is flawed because the relationship between training set instances and acoustic complexity is more complex than a linear function. Even if multiple *MCS* values were used, as with the discriminant criterion, the acoustic complexity still would not be fully captured.

4.1.3. Model selection

The acoustic models generated with regularized BIC with $\lambda = 0.9765$ and mBIC have the same total number of mixture components. They are also similar at the senone level, with only 18% of the senones differing in number of mixture components. Both start with the same original BIC model, but they allocate extra mixture components differently. Since regularized BIC applies a global regularization parameter and mBIC uses an inter-senone correction factor unique for each senone, mBIC is expected to perform better. Our experimental results agree, and we further compare BIC and mBIC over different acoustic model sizes in Table 3. The models were generated by varying the regularization parameter in (3) and (5) until the target model size was reached. The results show that mBIC consistently outperforms BIC.

Avg. Comp./Senone	12	15	18
WER of BIC model	16.98	16.00	15.62
WER of mBIC model	16.75	15.89	15.51

Table 3. *WER of BIC and mBIC with different average number of mixture components per senone.*

5. Conclusion

We propose a modification to BIC, mBIC, to generate acoustic models whose number of mixture components is statistically well-justified. The key idea is to consider inter-senone effects by introducing an inter-senone correction factor, derived from training data, into the BIC model complexity penalty term. In our experimental results, we show that mBIC has both a lower WER and 33% fewer mixture components than the baseline model that uses 32 mixture components per senone. In addition, it also consistently outperforms the BIC method. We also explain why mBIC allocates mixture components more effectively than other methods. For future work, we plan to study whether more complex correction factors can further improve speech recognition performance.

6. Acknowledgements

This research was supported by the National Science Foundation and the FCRP Focus Center for Circuit & System Solutions (C2S2). K. Yu is supported by a National Science Foundation Graduate Research Fellowship. The authors would like to thank Ravishankar Mosur and David Huggins-Daines for their valuable suggestions.

7. References

[1] L. R. Bahl, *et al.*, "Performance of the IBM large vocabulary speech recognizer on the ARPA Wall Street Journal task," in *Proc. ICASSP*, 1995, pp. 41-44.

[2] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in *Proc. ICASSP*, 1998, pp. 645-648.

[3] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, pp. 461-464, 1978.

[4] M. Padmanabhan and L. R. Bahl, "Model complexity adaptation using a discriminant measure," *IEEE Trans. Speech and Audio Proc.*, vol. 8(2), pp. 205-208, 2000.

[5] A. Biem, "Model selection criterion for classification: Application to hmm topology optimization," in *Proc. 7th ICDAR*, 2003, pp. 104-108.

[6] C. Bennett and A. I. Rudnicky, "The Carnegie Mellon communicator corpus," in *Proc. ICSLP*, 2002, pp. 341-344.

[7] P. Placeway, et. Al., "The 1996 Hub-4 sphinx-3 system," in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 85-89.

[8] R. Hu, X. Li, and Y. Zhao, "Acoustic model training using greedy EM," in *Proc. ICASSP*, 2005, pp. 697-700.

[9] Y. Chan, M. Siu, and K. Mak, "Pruning of state-tying tree using Bayesian information criterion with multiple mixtures," in *Proc. ICSLP*, 2000, pp. 294-297.