# A DISCRIMINANT MEASURE FOR MODEL COMPLEXITY ADAPTATION

L. R. Bahl, M. Padmanabhan
IBM T. J. Watson Research Center
P. O. Box 218, Yorktown Heights, NY 10598

## 1 ABSTRACT

We present a discriminant measure that can be used to determine the model complexity in a speech recognition system. In the speech recogition process, given a test feature vector the conditional probability of the feature vector has to be obtained for several allophone (sub-phonetic units) classes using a gaussian-mixture density model for each class. The gaussian-mixture models are constructed from the training data belonging to the allophone classes, and the number of mixture components that are required to adequately model the pdf of each class is determined by using some simple rule of thumb – for instance the number of components has to be sufficient to model the data reasonably well but not so many as to overmodel the data. A typical example of the choice of the number is to make it proportional to the number of data samples. However, such methods may result in models that are sub-optimal as far as classification accuracy is concerned. In this paper we present a new discriminant measure that can be used to determine in an objective fashion, the number of gaussians required to best model the pdf of an allophone class. We also present the results of experiments showing the improvement in recogntion performance when the number of mixture components is chosen based on the discriminant measure as opposed to the rule of thumb. These results are presented both for the speaker-independent and speaker-adapted case.

## 2 INTRODUCTION

We present a discriminant measure that can be used to determine the model complexity in a speech recognition system. In the speech recognition problem, feature vectors are extracted periodically from the input speech and are matched to different sequences of phones, that represent words in the vocabulary. In the statistical approach to speech recognition, this is done by estimating the probability density of each phone in the feature space from the training data, and using these pdf's to assign a probability to a test feature vector. The most common case is where a parametric model is used to model the pdf, with the parametric model generally being a mixture of gaussian distributions.

Hence, in the speech recogition process, given a test feature vector the conditional probability of the feature vector has to be obtained for several allophone (sub-phonetic units) classes using the gaussian-mixture density model for each class. It is not unusual to use tens or even hundreds of thousands of different gaussians in such models. The number of gaussians used to model each class is generally chosen based on simple rule of thumb - for instance the IBM system [8] chooses the number of mixture components to be proportional to the number of data samples, subject to the constraint that their parameters can be robustly estimated. Subsequently, the gaussians are initially constructed by randomly sampling the training data to construct the initial seed, and then K-means clustering the data to refine the means and variances of these gaussians. More recently, some alternative criteria that are more objective have been introduced to choose the size of the model [2] that estimate the number of gaussians required based on the classification accuracy that has been obtained with the current model.

In this paper we present a new discriminant measure that can be used to determine in an objective fashion, the number of gaussians required to best model the pdf of an allophone class. Once the number of gaussians has been determined, the discriminant measure can also be used subsequently to re-estimate the parameters of the gaussians. We also present results showing the improvement in performance obtained by using the proposed model selection criterion as opposed to the rule of thumb.

## 3 DESCRIPTION OF DISCRIMINANT MEASURE

When modelling data samples corresponding to a class with a mixture of gaussians, the parameters of the model are the number of mixture components and the means, variances and prior distributions of these components. In general, the number of mixture components is chosen using some simple *ad-hoc* rule subject to very loose constraints; for instance the number of components has to be sufficient to model the data reasonably well but not so many as to overmodel the data. A typical example of the choice of the number is to make it proportional to the number of data samples. However, such methods may result in models that are sub-optimal as far as classification accuracy is concerned. For instance, if the number of gaussians modelling a class is inadequate, it may result in the class being mis-classified often, and if too many gaussians are chosen to model a class, it may result in the model encroaching upon the space of other classes as well. We will refer to these two conditions as "non-aggressive " and "invasive" models respectively, and describe a measure that can be used to determine the number of mixture components to avoid these two classes of models.

### 3.1 Notation

The $t^{th}$ training data sample will be denoted $x_t$, and the class it belongs to (this is assumed to be known *a-priori*) will be denoted $C(x_t)$. This is obtained by viterbi-aligning the training data against the correct transcription. The model for class $l$ will be denoted $M_l$; hence, we denote the probability assigned to data sample $x_t$ by model $M_l$ as $p(x_t/M_l)$. Further, as the classifier is generally far from perfect, for any given data sample $x_t$, in general there will be several models in addition to $M_{C(x_t)}$ that give a reasonably high (for instance greater than a specified threshold) probability to $x_t$. All classes other than the correct class whose models give a reasonably high probability to $x_t$ will be designated as the "confusable" classes for data sample $x_t$ and will be denoted $F(x_t)$. This list of confusable classes is obtained by decoding the training data and producing N-best lists of hypotheses. Subsequently, a viterbi-alignment is done against each of the hypotheses, and the list of all allophones, $l$, that align to $x_t$ and are different from $C(x_t)$, are assigned to $F(x_t)$.

The discriminant measure that we propose is a 2-dimensional vector, $d_l$,

$$d_l = \begin{bmatrix} P_c^l & P_i^l \end{bmatrix} \quad (1)$$

characterizing every class $l$. The two components of the vector which we will refer to as the "correct probability of the class", $P_c^l$, and the "incorrect probability of the class", $P_i^l$ are computed as follows: The correct probability for class $l$ is computed from the training data samples that belong to the class $l$

$$P_c^l = \sum_{t \ni C(x_t)=l} \frac{p(x_t/M_l)}{p(x_t/M_l) + \sum_{j \in F(x_t)} p(x_t/M_j)} \quad (2)$$

The incorrect probability for class $l$ is computed from training data samples that belong to other classes, but that include $l$ in the confusable list of classes for that data sample

$$P_i^l = \sum_{t \ni l \in F(x_t)} \frac{p(x_t/M_l)}{p(x_t/M_{C(x_t)}) + \sum_{j \in F(x_t)} p(x_t/M_j)} \quad (3)$$

### 3.2 Model complexity estimation

Clearly the ideal situation would correspond to the case $d_l = [1 \quad 0]$ for all $l$; in this case the model for class $l$ would always give a probability of 1 to data samples that belong to class $l$, and a probability of 0 to data samples from all other classes. However, this ideal situation is rarely achieved, and in general the classes are characterized by $P_c^l$ and $P_i^l$ values lying between 0 and 1. For instance, a two-dimensional plot of the $d_l$ values are shown in Fig. 1, for a system that has 3042 classes, each modelled by atmost 8 gaussians, with the total number of gaussians being 24183. We can draw the following conclusions based on these values :

- If $P_c^l < threshold \quad (say \quad 0.6)$, this implies that the model for class $l$ gives data samples belonging to

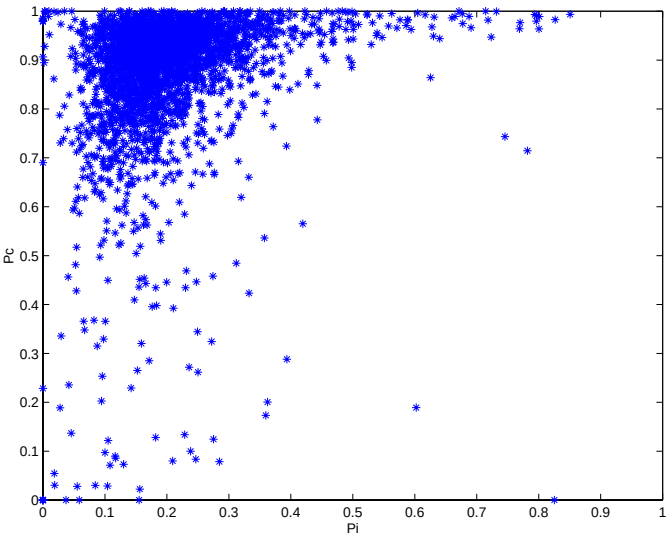


Figure 1: Two-dimensional plot of $P_i^l$ and $P_c^l$ values

the same class an average probability of less than 0.6 (non-aggressive model). Consequently, it must be the case that the model does not match the data very well, hence the resolution of the model for the class $l$ must be increased by adding components to its model.

• If $P_i^l > threshold$ $(say\ \ 0.6)$, this implies that the model for class $l$ gives a high probability to data samples from other classes, (invasive model). Hence, in order to improve the overall performance of the classifier, the number of components in the model for this leaf has to be reduced.

The above two observations form the basis for adapting the size of the model for selected allophones using the discriminant measure. The first step in the procedure involves constructing a number of systems using the conventional rule of thumb to pick the number of gaussians for each class. All the systems that we experimented with had the same number of allophones, 3042, and we constructed 7 systems using the rule of thumb for picking the number of gaussians, and assuming a maximum of 4, 5, 6, 7, 8, 20 and 56 gaussians per mixture. The total number of gausians in these systems were respectively 12156, 15187, 18205, 21204, 24183, 57129 and 117016; we will refer to these systems in the following sections as M4, M5, M6, M7, M8, M20 and M56 respectively.

In the next step of the procedure, the $d_l$ values are obtained for one of these systems , say M4. Subsequently, those allophones that had $P_c^l \leq 0.6$ were identified as non-aggressive allophones, and the mixture models for these allophones were replaced the corresponding mixture models from a larger system (M20). This resulted in a system that had 13800 gaussians will be referred to as M4xM20-0.6. A similar procedure can be carried out for the other systems. In the results presented in this paper, we only experimented with improving the models of non-aggressive allophones, and not the invasive allophones.

**3.2.1 Speaker Adaptation**—We also investigated the use of this technique to change the complexity of speaker-adapted models based on adaptation data provided by the test speaker. We constructed speaker-adapted gaussians for the M8 and M56 systems using MAP re-estimation [9] based on 50 sentences of adaptation data provided by each test speaker. Subsequently, we evaluated the discriminant measure on the adaptation data for each speaker using the M8 speaker-adapted models, and replaced the mixture models for the non-aggressive allophones with the models from the M56 speaker-adapted system.

# 4 EXPERIMENTAL RESULTS

## 4.1 Speaker-independent

We carried out recognition experiments using the M4-M56 systems, as well as the complexity adapted systems M4xM20-0.6, M6xM20-0.6, M8xM20-0.6, and M8xM56-0.6. The test data comprised of continuous speech from 10 speakers reading approximately 60 sentences each of business-related and office correspondence-related items. The results are tabulated in Table I and plotted in Fig. 2. In Fig. 2, the error rate is plotted as a function of the total number of gaussians in the system. The lower curve corresponds to the complexity-adapted systems and can be seen to be uniformly better than the conventional systems.

Table I

| Model type | # of gaussians | Error rate |
|---|---|---|
| M4 | 12.2k | 19.65 |
| M5 | 15.2k | 18.78 |
| M6 | 18.2k | 18.57 |
| M7 | 21.2k | 17.9 |
| M8 | 24.2k | 17.42 |
| M20 | 57.1k | 16.37 |
| M56 | 117.0k | 16.09 |
| M4xM20-0.6 | 13.8k | 19.0 |
| M6xM20-0.6 | 19.6k | 17.58 |
| M8xM20-0.6 | 25.4k | 16.94 |
| M8xM56-0.6 | 27.0k | 16.73 |

## 4.2 Speaker-adapted

The results for the speaker-adapted case are summarized in Table II. The test data comprised of the same 10 test speakers as above, and the adaptation data comprised of 50 sentences from each speaker. In Table II, M8sa indicates that the parameters of the gaussians of the M8 system were re-estimated using Bayesian adaptation [9] based on the adaptation data from the test speakers. Similarly, M56sa refers to speaker-adapted gaussians in the M56 system, and M8saxM56sa-0.6 refers to the complexity adapted systems that were constructed by putting together the M8sa and M56sa systems. Finally, M8xM56-0.6sa indicates that the parameters of the M8xM56-0.6 system were re-estimated using Bayesian adaptation based on the adaptation data from the test speakers. Further, as the number of gaussians differs for each speaker in the M8saxM56sa-0.6 case, the number of gaussians indicated in Table II refers to the average number of gaussians over all speakers. The results in Table II indicate that most of the advantage to be gained by adapting the complexity of the models is obtained for the speaker-independent

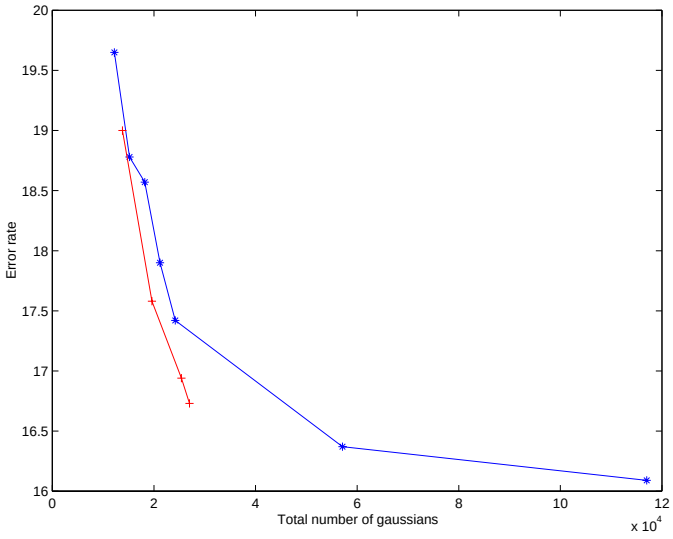


Figure 2: Plot of error rate as function of total number of gaussians

case itself, and there is not much more to be gained by adapting the complexity of the models specifically for each speaker.

Table II

| Model type | Av. # of gaussians | Error rate |
|---|---|---|
| M8sa | 24.2k | 13.52 |
| M56sa | 117.0k | 12.97 |
| M8saxM56sa-0.6 | 26.4k | 13.30 |
| M8xM56-0.6sa | 27.0k | 13.11 |

## 5 CONCLUSION

We presented a discriminant measure that can be used to determine the model complexity (i.e., number of gaussians used) in a speech recognition system. In general, the gaussian-mixture models that are used to represent the allophones in speech recognition systems are constructed from the training data belonging to the allophone classes, and the number of mixture components that are required to adequately model the pdf of each class is determined by using some simple rule of thumb – for instance making the number of components proportional to the amount of data. However, such methods may result in models that are suboptimal as far as classification accuracy is concerned. In this paper we presented a new discriminant measure that can be used to determine in an objective fashion, the number of gaussians required to best model the pdf of an allophone class. Experimental results show that the systems constructed in this manner generally provide better performance than systems constructed in the conventional manner.

## REFERENCES

[1] P. O. Duda and P. E Hart, "Pattern classification and scene analysis", Wiley, New York, 1973.

[2] Y. Normandin, "Optimal splitting of HMM gaussian mixture components with MMIE training", Proceedings of the ICASSP, pp 449-452, 1995.

[3] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum likelihood estimation from incomplete data", Journal of the Royal Statistical Society (B), vol. 39, no. 1, pp 1-38, 1979.

[4] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer, "Maximum mutual information estimation of hidden Markon model parameters for speech recognition", Proceedings of the ICASSP, pp 49-52, 1986.

[5] B. H. Juang, W. Chou, C. H. Lee, "Minimum classification error rate methods for speech recognition", IEEE Trans. Speech and Audio Processing, vol. 5, pp 257-265, May 1997.

[6] D. Luenberger, "Linear and Nonlinear Programming", Addison-Wesley Publishing Company, 1984.

[7] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm", IEEE Trans. Information theory, vol. IT-13, pp 260-269, Apr 1967.

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

[9] J. L. Gauvain and C. H. Lee, "Maximum-a-Posteriori estimation for multivariate Gaussian observations of Markov chains", IEEE Trans. Speech and Audio Processing, vol. 2, no. 2, pp 291-298, Apr 1994.