WEIGHTED DISTANCE MEASURES FOR EFFICIENT REDUCTION OF GAUSSIAN MIXTURE COMPONENTS IN HMM-BASED ACOUSTIC MODEL

Atsunori Ogawa[1] *and Satoshi Takahashi*[2]

[1]NTT Communication Science Laboratories, [2]NTT Cyber Space Laboratories, NTT Corporation

ABSTRACT

In this paper, two weighted distance measures; the weighted K-L divergence and the Bayesian criterion-based distance measure are proposed to efficiently reduce the Gaussian mixture components in the HMM-based acoustic model. Conventional distance measures such as the K-L divergence and the Bhattacharyya distance consider only distribution parameters (i.e. mean and variance vectors of Gaussian pdfs). Another example considers only mixture weights. In contrast to them, the two proposed distance measures consider both distribution parameters and mixture weights. Experimental results showed that the component-reduced acoustic models created using the proposed distance measures were more compact and computationally efficient than those created using conventional distance measures.

Index Terms— speech recognition, acoustic model, Gaussian mixture component reduction, distance measure, mixture weight

1. INTRODUCTION

It is well known that acoustic likelihood calculation is the most computationally expensive processing in Hidden Markov Model (HMM)-based speech recognition. Generally speaking, in total speech recognition processing, more than 60% of computational time is spent on acoustic likelihood calculation. Thus, to speed up the speech recognition process, acoustic likelihood computation should be reduced. A lot of research has been done to solve this problem [1, 2, 3, 4].

To reduce the amount of acoustic likelihood computation, without degrading recognition accuracy, several researches have been made on reducing Gaussian mixture components in HMM-based acoustic models [3, 4]. Their reduction techniques work by merging Gaussian mixture components based on distance measures that are defined between two Gaussian probability density functions (pdfs). In Gaussian mixture component-based HMM, the mixture weight for each component is an important factor, as are the distribution parameters (i.e. the mean and variance vectors of Gaussian pdfs). However, the conventional distance measures, such as the Kullback-Leibler (K-L) divergence [3, 5, 6] and the Bhattacharyya distance [5, 6] consider only distribution parameters. On the other hand, in [4], the distance measure considers only mixture weights.

In this paper, we propose two weighted distance measures; the weighted K-L divergence and the Bayesian criterion-based distance measure that consider both distribution parameters and mixture weights. The experimental results showed that the component-reduced acoustic models created using the weighted distance measures were more compact and computationally efficient than those created using conventional distance measures.

2. CONVENTIONAL DISTANCE MEASURES

In this section, we classify the conventional distance measures that consider either distribution parameters (i.e. mean and variance vectors of Gaussian pdfs) or mixture weights, and explain their characteristics. We denote the k-th mixture component in an HMM state as $c_k(\boldsymbol{x})$, which consists of an I-dimensional diagonal covariance Gaussian pdf $g_k(\boldsymbol{x})$ and its weight w_k (i.e., $c_k(\boldsymbol{x}) = w_k g_k(\boldsymbol{x})$). In $g_k(\boldsymbol{x})$, we denote the i-th element of the mean and variance vector as μ_{ki} and σ_{ki}, respectively.

2.1. Distance Measures Which Consider Only Distribution Parameters

The most representative examples are the K-L divergence (d_D) [3, 5, 6] and the Bhattacharyya distance (d_B) [5, 6]. d_D is a distance measure based on the difference area (or log-likelihood ratio) of two Gaussian pdfs. d_B, in contrast, is based on their overlap area.

$$\begin{aligned} & d_D(c_1(\boldsymbol{x}), c_2(\boldsymbol{x})) \\ &= \int g_1(\boldsymbol{x}) \log \frac{g_1(\boldsymbol{x})}{g_2(\boldsymbol{x})} d\boldsymbol{x} + \int g_2(\boldsymbol{x}) \log \frac{g_2(\boldsymbol{x})}{g_1(\boldsymbol{x})} d\boldsymbol{x} \\ &= \frac{1}{2} \sum_{i=1}^{I} \left\{ \frac{\sigma_{1i}^2 + (\mu_{1i} - \mu_{2i})^2}{\sigma_{2i}^2} + \frac{\sigma_{2i}^2 + (\mu_{2i} - \mu_{1i})^2}{\sigma_{1i}^2} \right\} \\ & \quad -I \end{aligned} \tag{1}$$

$$\begin{aligned} & d_B(c_1(\boldsymbol{x}), c_2(\boldsymbol{x})) = -\log \int \sqrt{g_1(\boldsymbol{x}) g_2(\boldsymbol{x})} d\boldsymbol{x} \\ &= \frac{1}{4} \sum_{i=1}^{I} \left\{ \frac{(\mu_{1i} - \mu_{2i})^2}{\sigma_{1i}^2 + \sigma_{2i}^2} + \log \frac{\left(\sigma_{1i}^2 + \sigma_{2i}^2\right)^2}{4\sigma_{1i}^2 \sigma_{2i}^2} \right\} \end{aligned} \tag{2}$$

More properly, d_D should be referred as "symmetrized" K-L divergence. But, in this paper, we omit it for shortening.

We explain the characteristics of d_D and d_B using Fig.1. The figure shows four mixture components in an HMM state.

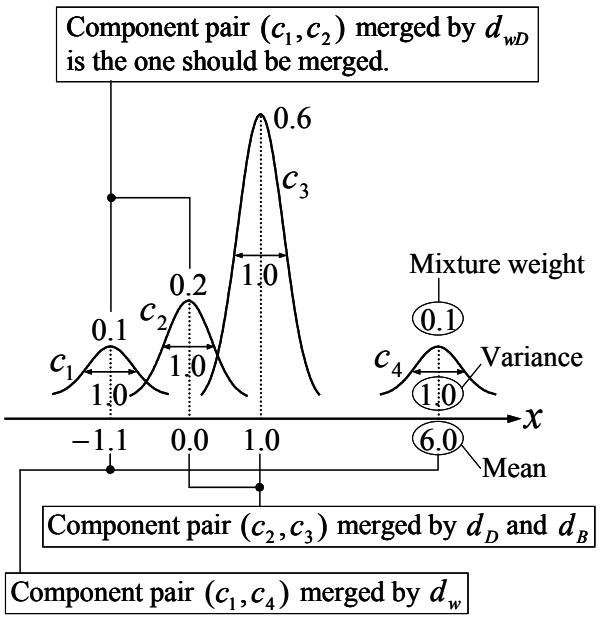

Fig. 1. Component pairs merged by each distance measure.

Each of them consists of a one-dimensional diagonal covariance Gaussian pdf. To simplify the explanation, we assume that all Gaussian pdfs have a common variance of 1.0. We try to reduce the number of mixture components to three by merging one pair of mixture components. We should choose the pair of mixture components that gives the minimum difference in likelihood of the state before and after merging. By this criterion, $c_3(\boldsymbol{x})$ should not be chosen because its weight is largest in the state. Moreover, $c_4(\boldsymbol{x})$ also should not be chosen because it is located far away from the other three components. The remaining components, $c_1(\boldsymbol{x})$ and $c_2(\boldsymbol{x})$, are thus the pair that should be merged because they are located close to each other and their weights are not too large. However, when using d_D or d_B, mixture weights are ignored, $c_2(\boldsymbol{x})$ and $c_3(\boldsymbol{x})$ are merged because they are the closest pair among all of the pairs.

2.2. Distance Measure Which Considers Only Mixture Weights

In [4], a sum of mixture weights of two components is used as the distance measure as follows (d_w).

$$d_w(c_1(\boldsymbol{x}), c_2(\boldsymbol{x})) = w_1 + w_2 \tag{3}$$

Using this measure, the larger the mixture weight is, the less likely it is that components such as $c_3(\boldsymbol{x})$ in Fig.1 will not be chosen for merging with other components. This characteristic is reasonable in merging components, as described in Section 2.1. Nonetheless, the distribution parameters are ignored; sometimes even widely separated pairs such as $c_1(\boldsymbol{x})$ and $c_4(\boldsymbol{x})$ are chosen to be merged.

3. PROPOSED DISTANCE MEASURES

We propose two distance measures that consider both distribution parameters and mixture weights. These measures are suitable for merging mixture components.

3.1. Weighted Kullback-Leibler Divergence

As a reasonable extension, by multiplying the mixture weight to each Gaussian pdf in the definition of the conventional K-L divergence described by Eq.(1), we get the definition of the weighted K-L divergence (d_{wD}) as follows [6].

$$\begin{aligned}
&d_{wD}(c_1(\boldsymbol{x}), c_2(\boldsymbol{x})) \\
&= \int w_1 g_1(\boldsymbol{x}) \log \frac{w_1 g_1(\boldsymbol{x})}{w_2 g_2(\boldsymbol{x})} d\boldsymbol{x} \\
&\quad + \int w_2 g_2(\boldsymbol{x}) \log \frac{w_2 g_2(\boldsymbol{x})}{w_1 g_1(\boldsymbol{x})} d\boldsymbol{x} \\
&= \frac{1}{2} \sum_{i=1}^{I} \Bigg[(w_2 - w_1) \log 2\pi\sigma_{1i}^2 + (w_1 - w_2) \log 2\pi\sigma_{2i}^2 \\
&\quad + \frac{w_1 \left\{ \sigma_{1i}^2 + (\mu_{1i} - \mu_{2i})^2 \right\}}{\sigma_{2i}^2} \\
&\quad + \frac{w_2 \left\{ \sigma_{2i}^2 + (\mu_{2i} - \mu_{1i})^2 \right\}}{\sigma_{1i}^2} \Bigg] \\
&\quad + I \Big\{ (w_1 - w_2) \log w_1 + (w_2 - w_1) \log w_2 \\
&\quad - \frac{1}{2} (w_1 + w_2) \Big\}
\end{aligned} \tag{4}$$

Here, there are the two mixture components, $c_1(\boldsymbol{x})$ and $c_2(\boldsymbol{x})$. They each have a one-dimensional ($I = 1$) diagonal covariance Gaussian pdf with a common variance of 1.0 and respective means of $\mu_{11} = -1.0$ and $\mu_{21} = +1.0$. The mixture weights are denoted as w_1 and w_2. Fig.2 shows the values of d_D, d_B, and d_{wD} as functions of $w_1 (= w_2)$, continuously ranging from 0.01 to 1.00. As can be seen this figure, d_{wD} varies with mixture weight, while d_D and d_B do not vary. This characteristic of d_{wD} is the same as with d_w

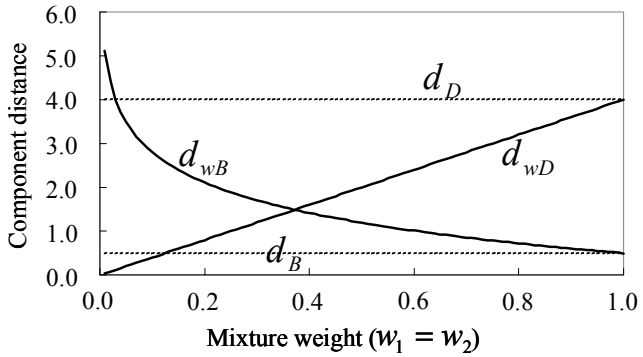

Fig. 2. Component distances as functions of mixture weights.

and is reasonable for component merging. Having both the characteristics of d_D and d_w, d_{wD} enables optimal component merging. Thus, by using d_{wD}, $c_1(\boldsymbol{x})$ and $c_2(\boldsymbol{x})$ can be merged, as shown in Fig.1.

On the other hand, similar to d_{wD}, we can also consider placing mixture weights on the definition of the conventional Bhattacharyya distance described by Eq.(4) (d_{wB}) [6].

$$\begin{aligned} & d_{wB}(c_1(\boldsymbol{x}), c_2(\boldsymbol{x})) \\ &= -\log \int \sqrt{w_1 g_1(\boldsymbol{x}) w_2 g_2(\boldsymbol{x})} d\boldsymbol{x} \\ &= -\frac{I}{2} \log w_1 w_2 + d_B(c_1(\boldsymbol{x}), c_2(\boldsymbol{x})) \end{aligned} \quad (5)$$

However, as shown in Fig.2, by using d_{wB}, the larger the mixture weight is, the easier it will be for the component to be merged with other components. This characteristic is against the criterion of optimal component merging described in Section 2.1; thus we can expect that d_{wB} can not be a good measure for component merging.

3.2. Distance Measure Based on the Bayesian Criterion

Recently, the Bayesian criterion has been actively applied to the problems in speech recognition. We propose a distance measure based on the Bayesian criterion in addition to d_{wD}. Concretely, we used the state split stopping criteria that works on the tree-based state clustering described in [7]. Based on the Bayesian criterion, by using prior distribution, robust model parameter estimation is possible even with small training data. Thus, we can expect that the Bayesian criterion-based distance measure will be effective for component merging in an HMM state. In an HMM state s, by denoting $c_0(\boldsymbol{x})$ as the merged component of $c_1(\boldsymbol{x})$ and $c_2(\boldsymbol{x})$, $g_s(\boldsymbol{x})$ as the prior component which is obtained by merging all of the components (its weight is 1.0), and $\mathcal{F}(c_k(\boldsymbol{x}), g_s(\boldsymbol{x}))$ as the Bayesian evaluation function of $c_k(\boldsymbol{x})$, the Bayesian criterion-based distance measure between $c_1(x)$ and $c_2(x)$ is obtained as follows.

$$\begin{aligned} & d_{BC}(c_1(\boldsymbol{x}), c_2(\boldsymbol{x}), g_s(\boldsymbol{x})) \\ &= \mathcal{F}(w_1 g_1(\boldsymbol{x}), g_s(\boldsymbol{x})) + \mathcal{F}(w_2 g_2(\boldsymbol{x}), g_s(\boldsymbol{x})) \\ &\quad -\mathcal{F}(w_0 g_0(\boldsymbol{x}), g_s(\boldsymbol{x})) \end{aligned} \quad (6)$$

4. MIXUTRE COMPONENT REDUCTION ALGORITHM

We used the mixture component algorithm as follows. This algorithm is based on the one described in [3].

(a) For the baseline acoustic model to be reduced, train the one that has a fixed number of Gaussian mixture components in every HMM state.

(b) In each HMM state of the acoustic model, calculate the distance between all of the Gaussian mixture component pairs and merge the closest one. By repeating this procedure, construct the Gaussian component binary tree in the bottom-up direction.

(c) By tracing the Gaussian component binary tree constructed in step (b) in the top-down direction, repeat the splitting of the Gaussian components. By stopping this procedure based on the MDL criterion, a reduced number of Gaussian mixture components can be obtained in each HMM state.

(d) Re-train the reduced acoustic model that is obtained by the above procedure.

We can put a penalty factor into the MDL criterion in step (c), and, by adjusting this factor we can get the desired size of the reduced model. In [3], the Gaussian component binary trees in each HMM state are constructed in the top-down direction using the k-means clustering technique and the conventional K-L divergence described by Eq.(1). Our algorithm, in contrast, works in bottom-up direction as described in step (b). This is because, when using distance measures that consider mixture weights such as d_w, d_{wD}, d_{wB}, and d_{BC}, there is no reasonable criterion for how to weight the centroid components that are generated in the k-means clustering procedure.

5. SPEECH RECOGNITION EXPERIMENTS

We evaluated the six distance measures described above in the BNF-grammar-based four-figure digits utterance recognition experiments.

5.1. Experimental Setup

Using a 45.5-hour speech database, which consisted of 131,603 word utterances by 93 male and 88 female speakers (not including digit utterances), we trained a baseline acoustic model that had 2,000-states with 16-mixture components in each HMM state (thus, the total number of mixture components was 32,000). Using the six distance measures (d_D, d_B, d_w, d_{wD}, d_{wB}, and d_{BC}) described in Sections 2 and 3, we reduced the size of the baseline model to get a reduced model that had 8,000 mixture components (corresponding to 4-mixture components in each HMM state, but the number of mixture components in each state was not fixed).

Evaluation experiments were done for the six reduced models just after the reduction procedure (i.e. after step (c) in the reduction algorithm of Section 4) and their re-trained versions (i.e. after step (d)). The reason why we evaluated the models just after the reduction procedure was that we wanted to purely compare the aptitudes of the six distance measures in component merging. The models were re-trained using the conventional maximum likelihood parameter estimation method with five-time iterations. For comparison, the 4, 10, and 14-mixture fixed component models that were generated while constructing the baseline 16-mixture models were also evaluated (denoted as b_4, b_{10}, b_{14}, and b_{16}).

Eight male and eight female speakers uttered the evaluation speech data. Each speaker uttered 40 four-figure digit sequences. Thus, the total number of utterances was 640. Each utterance was recognized by the BNF-grammar that accepts any-figure digit sequences. We used the speech recognition engine VoiceRex [8], which was developed at NTT Cyber Space Labs.

5.2. Experimental Results

The experimental results are shown in Fig.3. The vertical axis shows the digit accuracy. Real Time Factors (RTFs) of b_4, b_{10}, b_{14}, and b_{16} are shown in the parentheses. They are normalized by the RTF of b_{16}. The RTFs of the 12 reduced acoustic models are almost equal to the RTF of b_4 (0.45), which has the same model size as the reduced models.

Before the re-trainings (just after reduction), it can be seen that there are great differences in the digit accuracies of the six reduced models. The two proposed distance measure (d_{wD} and d_{BC}) based models give better accuracies than conventional ones (d_D, d_B, and d_w-based models). The d_{BC}-based model shows the best result. It gives a 14% improvement in accuracy (corresponding to 45% error reduction) compared with b_4, which has the same size with the d_{BC}-based model (from 68.59% of b_4 to 82.65% of d_{BC}). The accuracy of the d_{BC}-based model is almost the same as that of b_{10} (82.93%), and comparing with this model, the d_{BC}-based model gives a 39% RTF reduction (from 0.74 of b_{10} to 0.45 of d_{BC}) and a 60% model size reduction (from 20,000-components of b_{10} to 8,000-components of d_{BC}). The accuracy of the d_{wD}-based model (80.82%) is 7% higher than that of the d_D-based model (73.59%) (corresponding to 27% error reduction). This improvement is obtained by the effect of considering mixture weights in combination with the conventional K-L divergence. However, the d_{wB}-based model does not show good performance. Consequently, we can confirm that d_{wB} is not a reasonable distance measure for component merging, as expected in Section 3.1.

After the re-trainings, all reduced models show performance improvements. The improvements are large especially for the models that had poor performances before the re-trainings (such as the d_w and d_{wB}-based models). However, the d_{wD} and d_{BC}-based models still has the best performances. The d_{wD} and d_{BC}-based models give 18% improvements in digit accuracy (corresponding to 58% error reduction) compared with b_4 of the same size with the d_{wD} and d_{BC}-based models (from 68.59% of b_4 to 86.80% of d_{BC}). Moreover, these accuracies are close to that of b_{14} (87.46%), and compared with b_{14}, the d_{wD} and d_{BC}-based models give a 50% RTF reduction (from 0.90 of b_{14} to 0.45 of d_{BC}) and a 71% model size reduction (from 28,000-components of b_{14} to about 8,000-components of d_{BC}).

6. CONCLUSION

We proposed two weighted distance measures; the weighted K-L divergence and the Bayesian criterion-based distance measure for efficient reduction of the Gaussian mixture components in the HMM-based acoustic model. They consider both distribution parameters (i.e., mean and variance vectors of Gaussian pdfs) and mixture weights. Using these distance measures, in the BNF-grammar-based four-figure digits utterance recognition experiments, we could reduce the model size by 71% and the RTF by 50% from the baseline while maintaining higher accuracies than were possible with conventional distance measures.

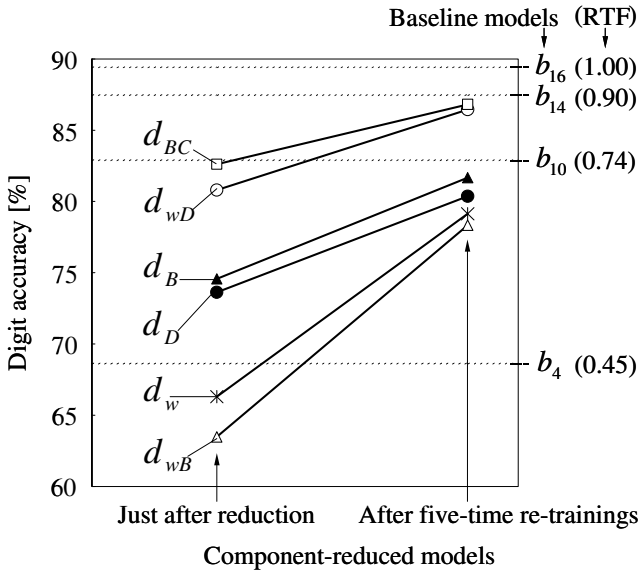

Fig. 3. Speech recognition results using component-reduced models created by each distance measure.

7. ACKNOWLEDGEMENTS

The authors thank Dr. Shinji Watanabe at NTT Communication Science Labs. He taught us the Bayesian criterion-based speech recognition techniques. The authors also thank Prof. Kazuya Takeda at the Graduate School of Information Science, Nagoya University, for his helpful comments and insights.

8. REFERENCES

[1] E. Bocchieri, "Vector quantization for the efficient computation of continuous density likelihoods," Proc. ICASSP93, vol.2, pp.692–695, Apr. 1993.

[2] S. Kanthak, K. Schutz, and H. Ney, "Using SIMD instructions for fast likelihood calculation in LVCSR," Proc. ICASSP00, vol.3, pp.1531–1534, June 2000.

[3] K. Shinoda and K. Iso, "Efficient reduction of Gaussian components using MDL criterion for HMM-Based speech recognition," Proc. ICASSP02, vol.1, pp.869–872, May 2002.

[4] V. Fischer and T. Rob, "Reduced Gaussian mixture models in a large vocabulary continuous speech recognizer," Proc. EUROSPEECH99, vol.3, pp.1099–1102, Sept. 1999.

[5] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification (2nd ed.), Wiley Interscience, 2000.

[6] T.Y. Young and K.S. Fu, Handbook of Pattern Recognition and Image Processing, Academic Press, 1986.

[7] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clutering for speech recognition," IEEE Trans. SAP, vol.12, no.4, July 2004.

[8] A. Ogawa, Y. Noda, and S. Matsunaga, "Novel two-pass search strategy using time-asynchronous shortest-first second-pass beam search," Proc. ICSLP00, vol.4, pp.290-293, Oct. 2000.