SPEAKER ADAPTIVE CONFIDENCE SCORING USING BAYESIAN COMBINING

Tae-Yoon Kim and Hanseok Ko

Department of Electronics and Computer Engineering, Korea University
SungBuk-Gu, Anam-Dong, Seoul, 136-713, KOREA
tykim@ispl.korea.ac.kr hsko@korea.ac.kr

ABSTRACT

Bayesian combining of confidence measures is proposed for speech recognition. Bayesian combining is achieved by the estimation of joint pdf of confidence feature vector in correct and incorrect hypothesis classes. If the joint pdf in the two classes are correctly estimated, this method guarantees an optimal combining in the minimum Bayes risk sense. Investigating the distribution of confidence features, we found out that the pdfs are well estimated by Gaussian mixture model with full covariance matrix in combining small number of features. In addition, the adaptation of a confidence score by adapting the joint pdf is presented. The proposed methods reduced the classification error rate by 17% from the conventional single feature based confidence scoring method in isolated word Out-of-Vocabulary rejection test.

1. INTRODUCTION

In speech recognition, confidence measures (CMs) are used to evaluate the reliability of recognition results. ASR systems suffer severe performance degradation in real application due to environmental mismatch, noise, spontaneous speech and other factors. The capability of measuring the degree of confidence of hypothesized words enables ASR systems to detect unreliable or misrecognized results and control these erroneous outputs more actively. In most cases, the estimation of CM is followed by a decision making whether a hypothesized word is correct or incorrect. Thus, CM estimation is directly related to a binary classification problem where CM is treated as a feature vector. Previous works can be categorized into two parts in this perspective. The first approach is to find good CM features for classification. The good CM features should separate the classes of correct and incorrect hypothesis well. These CM features are obtained from acoustic information (word level likelihood ratio test [1]) or side information from a decoding process (N-best list, word posterior probability [2]). The second approach is to design a good classifier which uses various CMs as a feature vector. These works could be called CM combining. Recent efforts on CM combining include linear discriminant analysis (LDA) based CM combining [3], support vector machine (SVM) classifier [4], boosting [5], and others.

In this paper, we propose a Bayesian combining of CM features. This approach is concerned about the pdf estimation of the two classes and finds Bayes optimal decision boundary. Previous research focused on obtaining good classification results via dimensionality reduction (LDA) or maximizing generalization power (SVM). In general, Bayesian classification does not consider these factors for good classification. However, it is expected that the statistical modeling of CM features provides a solid basis for further manipulation of CM. As an application of statistical modeling of CM, we consider speaker adaptation of CM. The well known MAP adaptation method can be easily applied to statistical modeling of CM. Detailed descriptions and related experiments can be found in the subsequent sections.

This work was funded by the Korea Ministry of Commerce, Industry and Energy (No. 10011362).

2. BAYESIAN COMBINING OF CM FEATURES

CM combining can be considered as a type of binary classification problem in which individual CMs are used as features for making a decision whether the recognition result is correct or incorrect. The objective of Bayesian CM combining is to minimize the Bayes risk associated with such a decision. From the well known Bayesian classification rule in binary class cases, the following decision rule can be expressed

$$\frac{p(x_1, x_2, ..., x_N|\omega_1)}{p(x_1, x_2, ..., x_N|\omega_0)} \underset{\omega_0}{\overset{\omega_1}{\gtrless}} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} \tag{1}$$

where x_i is a feature derived from i-th individual CM, P_i is the prior probability of class ω_i, and C_{ij} is the cost associated with decision making for ω_j when ω_i is present. If the priori probabilities and the costs are given, we can easily design the optimal Bayesian CM combining which achieves the minimum Bayes risk by choosing the threshold value η_0. In the case of simple cost $C_{ij} = 1 - \delta_{ij}$, the Bayesian CM combining classifier is optimal in minimum error rate sense. If the cost values are not given explicitly, determining the ratio of cost values based on relative cost of each decision error with $C_{ii} = 0$ can be a good method.

A basic question which could be raised in Bayesian combining is, "does the combining give performance increase?" The answer is "yes." The performance of Bayesian combining of any features is better than or at least the same with those of a single feature which can be proved as follows:

Let $\vec{X}_i = (x_1, ..., x_{i-1}, x_i)$ be a given feature vector for CM combining. Associated optimal decision rule r_i can be written as

$$\frac{\int p(x_1, ..., x_i, x_{i+1}|\omega_1)dx_{i+1}}{\int p(x_1, ..., x_i, x_{i+1}|\omega_0)dx_{i+1}} \underset{\omega_0}{\overset{\omega_1}{\gtrless}} \eta_0 \tag{2}$$

where a hidden random variable x_{i+1} is revealed by marginal distribution property. When the hidden feature x_{i+1} is added to the feature vector, associated optimal decision rule r_{i+1} with extended feature vector $\vec{X}_{i+1}$ is

$$\frac{p(x_1, ..., x_i, x_{i+1}|\omega_1)}{p(x_1, ..., x_i, x_{i+1}|\omega_0)} \underset{\omega_0}{\overset{\omega_1}{\gtrless}} \eta_0 \tag{3}$$

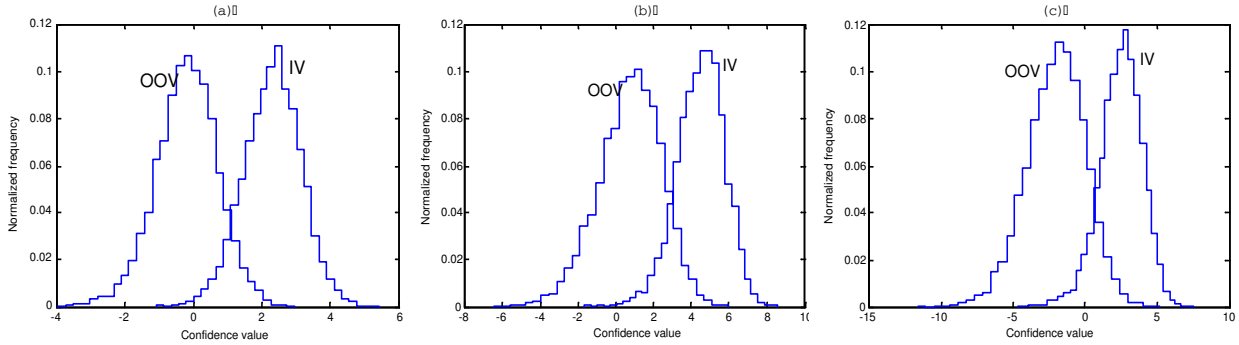

Fig. 1. Normalized histograms of confidence value in IV and OOV classes. (a) Phonetic filler, (b) 256 Gaussian mixture model, (c) Sub-word LLR (transformed)

Equation (2) tells us that the decision rule r_i is one of the possible decision rules with feature vector $\vec{X}_{i+1}$. Thus, with the extended feature vector, decision rule r_i is not better than the optimal decision rule r_{i+1}. So, the Bayes risk R_i associated with feature vector $\vec{X}_i$ has the following relationship the following relationship

$$R_1 \geq ... \geq R_i \geq R_{i+1} \geq .... \tag{4}$$

It means that the performance of Bayesian combining of any features is better than or at least the same with those of a single feature, and as more features are added, the performances shows non-decreasing property. Thus, if we can estimate accurate pdfs, Bayesian combining guarantees that performance will increase (though it's not a strict increase).

In summary, Bayesian CM combining becomes a likelihood ratio test in which the estimation of joint pdf of feature vector $\vec{X}_N = (x_1, ..., x_{N-1}, x_N)$ becomes crucial problem. However, the accurate estimation of pdf's could be another problem, especially with limited training data samples or an unknown pdf. Multivariate Gaussian and Gaussian mixture are considered for the pdf modeling. Histogram analysis of individual features reveals that some of the promising features' distributions are very similar to Gaussian distribution. Some features using the sigmoid function for limiting value range may not have the Gaussian distribution form. However, by applying the inverse transform of the sigmoid function, the distribution of those features can be made similar to Gaussian. In such a feature transform, it is easy to show that any nonsingular feature transform does not change the classification performance in Bayesian classification.

Although Gaussian distribution seems to be suitable for modeling some features or their transformed form, the others need more general form of distribution model. CMs obtained from the on-line garbage method and the word active count method are easily influenced by the decoding beam width and may have lower or upper bounded values. Since these CM features are not Gaussian, Gaussian mixture model was appropriate for estimating their pdfs. With Gaussian mixture pdf, we could model any form of distribution theoretically. But, in most cases, it requires lots of mixture components which easily lead to unreliable parameter estimation. These are examined in the experimental section.

3. MAP SPEAKER ADAPTATION OF CM

Speaker adaptation is a process of adapting statistical model parameters, mainly HMM parameters of acoustic model, to a specific speaker by using a small amount of speaker dependent speech data. Bayesian CM combining is based on statistical modeling and parameter estimation of CM features. Thus, it is easy to apply the well established various adaptation techniques to the CM adaptation. Effect of adaptation of confidence measures has not been studied well. However, since most of the CM features are derived from acoustic and language model scores that could easily vary according to the speaker, task or environmental change, we think that the adaptation of CM is also important. Here, we consider the MAP estimation for speaker adaptive CM estimation. MAP adaptation uses speaker independent statistical model for determining prior model of concerned parameters. MAP adaptation of mean parameter is expressed as

$$\hat{\mu} = \frac{N}{N+\tau}\bar{\mu} + \frac{\tau}{N+\tau}\mu_0 \tag{5}$$

where $\bar{\mu}$ is sample mean of adaptation data, μ_0 is speaker independent mean, N is the number of observation in adaptation data and τ is a weighting of prior knowledge to adaptation speech. If the Gaussian mixture is used for statistical model of CM combining, N is changed to the expected number of occupying jth mixture, N_j and $\bar{\mu}$ are changed as posterior weighted sample mean of jth mixture, $\bar{\mu}_j$ as the follows

$$N_j = \sum_{i=1}^{N_0} p(M = j|\vec{O}_i) \tag{6}$$

$$\bar{\mu}_j = \frac{\sum_{i=1}^{N_0} p(M = j|\vec{O}_i)\vec{O}_i}{\sum_{i=1}^{N_0} p(M = j|\vec{O}_i)} \tag{7}$$

4. EXPERIMENTS AND DISCUSSIONS

To evaluate the performances of the Bayesian confidence scoring methods, an isolated word Out-of-Vocabulary (OOV) rejection task was designed. About 120,000 utterances from 6,000 isolated words were used for training Korean isolated word recognition system, based on triphone units. These are recorded in an

Table 1. CM features used in OOV rejection experiments.

1	Phonetic filler	Unigram phonetic network
2	256 GMM	Likelihood normalization by 256 mixture of Gaussian
3	Sub-word LLR	Word level geometric mean of sub-word LLR
4	256 GMM_sub	Sub-word LLR with 256 GMM as anti-phone model
5	Sub-word LLR_arth	Word level arithmetic mean of sub-word LLR
6	On-line garbage	Average of likelihood difference between 1st and nth best
7	Active word count	Average number of active hypothesized word path

Table 2. Probability of error $P(e)$ of speaker independent (SI) and speaker adaptive (SA) Bayesian combining methods and FLDA combining with varying the number of CM features at equal prior condition (%). In SA Bayesian combining, 10, 20, and 100 utterances were used for each speaker adaptation.

Number of features	Single Gaussian				Gaussian mixture				FLDA
	SI	SA			SI	SA			SI
3	6.5	6.3	6.2	6.1	6.2	6.0	6.0	5.9	6.5
5	7.3	7.1	7.1	7.1	6.5	6.4	6.3	6.1	6.6
7	7.6	7.4	7.4	7.3	7.1	7.0	6.9	6.8	6.6

office environment. Another corpus recorded in a silent environment using a different microphone was used for the development and the test data set. It consisted of about 15,000 utterances of 452 isolated words collected from 30 native Korean speakers. Two hundred words among the 452 words were used for In-Vocabulary (IV) words and the other 252 words were used for OOV words. Nine thousand utterances from 20 speakers were used for estimating the statistical models for Bayesian CM combining test. And 13,000 other utterances from 10 speakers were used for the CM adaptation and OOV rejection test. Phonetic filler method, word level Gaussian mixture model (GMM) method and sub-word based log-likelihood ratio (LLR) are considered as baseline CM features (Table 1). They show good performances and have relatively low correlation coefficients with each other.

Fig.1 shows the histograms of these CM features. The CM features using phonetic filler model and GMM have distributions that are similar to the Gaussian in IV and OOV word classes. The sub-word based LLR feature which uses a sigmoid function for limiting the value range of sub-word LLR does not have Gaussian distribution. However, by applying inverse transformation of sigmoid function to the word level average of sub-word LLR, we can make sub-word LLR based CM feature have near Gaussian distribution in both IV and OOV classes. First, for the Bayesian combining of CMs, multivariate Gaussian model is used for pdf of the combined CM feature vector. Mean vector and full covariance matrix are estimated using the recognition results with the development data set. Since the CM features are highly correlated with each other, using a diagonal covariance matrix is not appropriate. The parameters of univariate Gaussian models are also estimated for statistical modeling of the individual CM features. Given the prior probability of OOV word $P(\omega_0)$, if we use the simple Bayesian cost, OOV word classification error rate $P(e)$ of the individual CMs and the Bayesian CM combiner can be obtained by the equation $P(e) = (1 - P(\omega_0))(1 - P_D) + P(\omega_0)P_F$, where P_D is the probability of detecting IV word, and P_F is the probability of falsely detecting OOV word as IV word.

In Fig.2, Bayesian CM combining consistently out performs

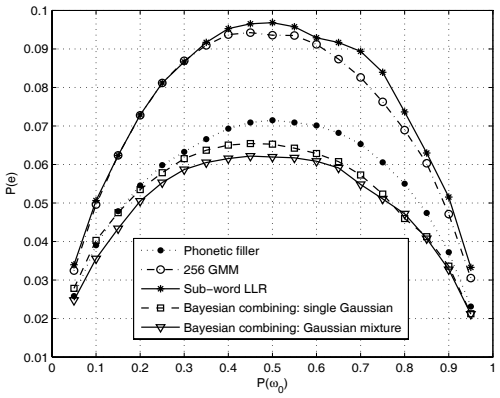

Fig. 2. Probability of error $P(e)$ of individual CM features and Bayesian combining methods in OOV rejection test

the individual CMs at all prior conditions. Compared with the individual CMs, Bayesian combining reduces the error by about 10% from the best individual feature (phonetic filler) at equal prior condition. The performance with Gaussian mixture pdf is also shown in Fig.2. When 3 Gaussian mixture components are used for pdf modeling, the error rate is slightly decreased. However, as the number of mixture Gaussian components is further increased, the error rate also consistently increase. As seen in the histograms of the three features, they have smooth and unimodal Gaussian-like distributions, thus small number of mixture components is sufficient for the modeling of the pdf. When other features are added, the error rate increase. In Table 2, we can see that the error rates of Bayesian combining are increasing as the number of combined features is increased. This is contrasted with the non-

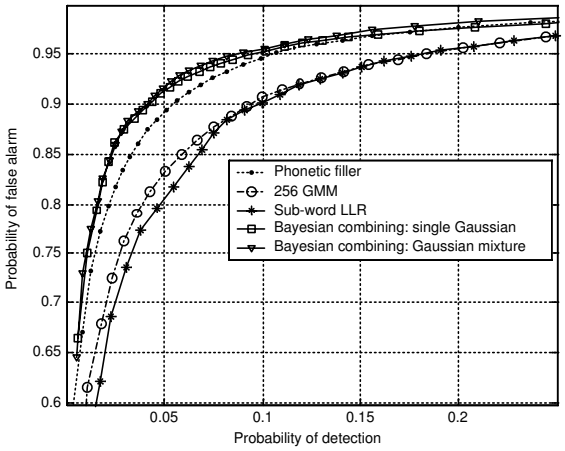

Fig. 3. ROC plot of individual CM features and Bayesian combining methods in OOV rejection test

decreasing property proved in section 2. This is due to the inaccuracy in pdf estimation. The discrimination power obtained by adding more features is not larger than the estimation error incurred by the wrong assumption of distribution family especially in single Gaussian combining and by the unreliable estimation of increased parameters in Gaussian mixture combining. Due to the full covariance matrices in mixture model, the number of parameters to be estimated easily becomes too many for reliable estimation. Thus, we should be careful in selecting good CM features and the appropriate pdf models for combining not to raise the estimation problem. Fisher's LDA (FLDA) shows a similar performance with Bayesian combining methods in the 3 CM features case. Also, when more features are added, its generalization power seems somewhat better than the Bayesian combining with Gaussian mixture model. We can see from the ROC plot (Fig.3) that Bayesian combining achieves about 92% of P_D at P_F of 5%.

MAP adaptation is performed for mean adaptation. Covariance matrix adaptation did not show significant improvement. From 20 to 100 utterances are used for the adaptation of each speaker and the amount of IV and OOV adaptation data was the same in all conditions. Prior weighting constant was set experimentally as 20. This is similar to the value usually used in acoustic adaptation in ASR systems. The error rates with speaker adapted CM is also found in Table 3. The speaker adapted Bayesian combining shows significantly lower error rates. The minimum error rate is obtained by speaker adapted Bayesian combining of 3 features that reduces the error by 18% from the best individual CM feature.

5. CONCLUSIONS

In this paper, we proposed the Bayesian method of combining confidence measures (CMs) and its application to speaker adaptation of CM. If we can estimate the pdf of the CM feature vector without error, the Bayesian CM combining gives a decision rule which achieves minimum decision error. In the OOV rejection experiments, the Bayesian CM combining showed significant error rate reduction. However, we should be careful in choosing good CM features and the appropriate pdf model for combining so as not to raise the estimation problem. Speaker adaptation of the CM feature showed significant error rate reduction. Further study on the influence of acoustic model adaptation to speaker adaptive CM is desirable in future works.

6. REFERENCES

[1] E. Lleida and R.C. Rose, "Utterance verification in continuous speech recognition: Decoding and training procedures," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 8, no. 2, pp. 126–139, 2000.

[2] Frank Wessel, Ralf Schluter, Klaus Macherey, and Hermann Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 9, no. 3, pp. 288–298, 2001.

[3] S. Kamppari and T. Hazen, "Word and phone level acoustic confidence scoring," in *Proc. ICASSP '00*, Istanbul, Turkey, 2000, pp. 5–9.

[4] R. Zhang and A. Rudnicky, "Word level confidence annotation using combinations of features," in *Proc. EUROSPEECH '01*, Aalborg, Denmark, 2001, pp. 2105–2108.

[5] P.J. Moreno, B. Logan, and B. Raj, "A boosting approach for confidence scoring," in *Proc. EUROSPEECH '01*, Aalborg, Denmark, 2001, pp. 2109–2112.