2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

VQ 방식의 화자인식 시스템 성능 향상을 위한 부쓰트랩 방식 적용

*경연정, *이진익, *,**이황수

*경기도 성남시 분당구 수내동 9-1 SK Telecom 중앙연구원

**대전광역시 유성구 구성동 373-1, KAIST 전자전산학과 전기 및 전자 전공

The bootstrap VQ model for automatic speaker recognition system

*YounJeong Kyung, *Jin-Ick Lee, *,**Hwang-Soo Lee

*SK Telecom R&D Center, 9-1 Sunae-dong, Pundang-gu, Sungnam City, Kyunggi-do 463-020, Korea

**KAIST Dept. of EECS, 373-1 Kusung-dong, Yusung-gu, Taejon, 305-701, Korea

yjkyung@sktelecom.com

요 약

VQ 모델로 구성된 화자인식 시스템의 성능 향상을 위해 Bootstrap 방식을 적용하였다. Bootstrap 및 aggregating 방식은 unstable 한 모델에서 그 성능이 유효하므로 이의 적용을 위해 먼저 VQ 모델의 bias 와 variance 를 계산하여 unstable 함을 보였다. 화자인식 실험은 TIMIT Database 를 사용하여 수행하였고 실험결과 높은 인식율 향상을 확인하였다. 또한 적은 훈련 데이터 환경에서도 좋은 인식율을 갖는 것으로 나타났다.

Abstract

A bootstrap and aggregating (bagging) vector quantization (VQ) classifier is proposed for speaker recognition. This method obtains multiple training data sets by resampling the original training data set, and then integrates the corresponding multiple classifiers into a single classifier. Experiments involving a closed set, text-independent and speaker identification system are carried out using the TIMIT database. The proposed bagging VQ classifier shows considerably improved performance over the conventional VQ classifier.

Introduction

The concept of the **bootstrap** and **aggregating** (bagging) method is as follows: A classification method is unstable if small perturbations in their training sets or in their construction can result in large changes in the constructed classifier. Unstable classifiers can have their accuracy improved by perturb and combine methods. That is, multiple versions of the classifier are generated by perturbing the training set of the construction method, then these multiple versions are

combined into a single classifier [1].

Bias and variance of classifier

Breiman introduced the notion of classifier bias and variance [2]. Let the training data set L consist of data (x,y) where y is the class label of classifier C(x,L) if the input is x, we predict y by C(x,L). Usually we have a single training data set L. Take repeated bootstrap sets L_B from L, and form classifiers C(x,L_B). Let Q(y|x) = P (C(x, L)=y) and define the aggregated classifier as

$C_A(x) = \text{argmax}_{(y)} Q(y|x)$. (Eq.1)

This is aggregation by voting.

The Bayes theorem is P(y|x) = P(x|y)P(y) / P(x) where P(y|x) is the probability of y, given x.

According to above equation, the Bayes classifier C^* is defined as $C^*(x) = \text{argmax}_{(y)} P(y|x)$.

That is, the Bayes classifier C^* chooses the class label y, which maximize the probability P(y|x).

The classifier C(x) is unbiased at x, if the predicted class label y of the aggregated classifier is not the same as the Bayes classifier. That is, C(x) is unbiased at x if $C_A(x) \neq C^*(x)$. (Eq.2)

Let U be the set of all x at which C is unbiased, and call U the unbiased set. The complement of U is called the bias set and denoted by B. Each element of the other set T is assigned to either the bias set or unbias set.

Breiman introduced the bias and variance of classifiers as follows: The bias of a classifier C is

$\text{Bias}(C) = P_{x,y}(C^*(x)=Y, x \in B) - E_T P_{x,y}(C(x,T)=Y, x \in B)$ (Eq.3)

and its variance is

$\text{Var}(C) = P_{x,y}(C^*(x)=Y, x \in U) - E_T P_{x,y}(C(x,T)=Y, x \in U)$ (Eq.4)

If the variance is large, the classifier is highly affected by construction of training data sets. On the contrary, if the variance is small, the classifier is stable on construction of training data sets. That is, the variance of the classifier measures the unstability of the classifier.

Unstability of the VQ classifier

Two experiments show that the VQ classifier is unstable. One is to compute the variance of the VQ classifier with a waveform database and show that its variance is large. The other one is to perform the ASR using the VQ classifier with minor changes to the training data set and show that the recognition results significantly vary depending on the training data set.

We compute the bias and variance with a waveform database [3]. Breiman computed variances of CART and, bagging CART in [2] with the waveform database. We use the same simulated database "waveform" to compare with the variances of the VQ classifier and bagging VQ classifier.

The variance is computed as follows: We already make the one hundred sets of 300 factors and obtained the Bayes classifier C^*, the VQ classifier C corresponding to each set, and the aggregated classifier CA. An additional set of 18,000 factors was generated from the same distribution and the aggregated classifier was computed. According to equation (Eq.2), this test set was divided into bias and unbias sets, and then, the bias and variance are computed by equations (Eq.3) and (Eq.4). The bias and variance of the VQ classifier are in Table 1.

Table 1. Bias and variance with waveform database

	CART	Bagging CART	VQ	Bagging VQ
Bias	1.7	1.4	3.54	2.97
Variance	14.1	5.3	11.95	4.35
Error	29.0	19.5	24.5	13.2

Table 1 shows that the variance of the VQ classifier is as large as that of the CART. Another experiment is also performed to show that the VQ classifier is unstable. A classifier is unstable in that minor changes in the training set could cause large changes in classifier results. We construct the bagging VQ classifier for ASR with the TIMIT database [4]. We use the 100 speakers subset TIMIT. As for the TIMIT database, one sentence between 1 and 5 is used by each speaker for training. The five sentences from 6 to 10 are used for testing. The experimental result is shown in Tables 2.

Table 2. Recognition rates using only one sentence for training

Sentence index	Number of vectors	Recognition rates (%)
1	302	20.1
2	245	24.0
3	299	18.2
4	257	16.9
5	314	19.1

The recognition rates vary from 16.9% to 24% with the TIMIT database. Therefore, according to these results, we can say that the VQ classifier is unstable. So we expect that the bagging method improves VQ classifier performance significantly.

Bagging VQ classifier for ASR

Let training set L consist of data (x_n, y_n), n=1,...,N, where N is the total amount of training data, and y is the speaker's ID number. Put equal probabilities 1/N on each sample and, using these probabilities, sample with replacement N times from the training set L, forming the resampled training set L_b.

Some samples in L may not appear in L_b; some may appear more than once. Use L_b to construct the corresponding classifier C_b. In the classification procedure, for an unknown feature vector x, the predicted speaker y of x is elected by Equation (Eq.1), where M is the number of resampled data sets.

Experiments

The ASR experiments are conducted using the bagging VQ classifier. A 100-speaker subset of the TIMIT database was used for the experiments. Using the conventional method (codebook size of 32), we obtain the recognition rate of about 85%. The results of the proposed bagging VQ classifier are shown in Figure 1. Each VQ codebook has 32 codewords.

Figure 1. Recognition rates by bagging VQ classifier

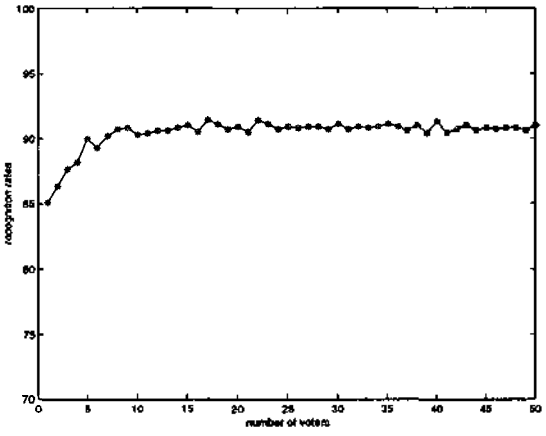

Figure 1 displays the recognition rates vs. the

number of voters. As shown in Figure 1, the recognition performance improves when at least two or three voters are used. The recognition rate of the conventional VQ is 85%. But the recognition rate of the bagging VQ with five voters is about 90%. The speaker recognition rates are improved significantly by bagging VQ. When the number of voters exceeds five, the performance improvement becomes marginal.

The second experiment concerns small training databases. The bagging VQ classifier reveals good performance even with small training databases. The experiments were performed with a database-256 [5]. Experimental results are shown in Tables 3 and 4.

Table 3. Recognition rates depending on number of training sentences (conventional VQ model)

	Use 5 sentences	Use 10 sentences
Recognition Rates	88.67 (%)	92.19 (%)
No. of training Vectors	1101	2344

Table 3 presents the results obtained from the conventional VQ classifier (codebook size of 16). This table shows the variation in recognition rates depending on the size of the training database. The recognition rate of 88.67% is obtained when we use 5 sentences for training. When we increase the number of training sentences to 10, the recognition rate improves to 92.19%. The recognition results from the bagging VQ classifier (codebook size of 16) are shown in Table 4.

Table 4. Recognition rates by bagging VQ classifier (256 database, using 5 sentences for training)

No. of voters	Recognition Rates (%)
2	90.63 (%)
3	94.53 (%)
4	95.70 (%)
5	95.70 (%)
6	96.48 (%)
7	97.27 (%)

These classifiers use only 5 sentences for training. When at least three voters are used for the bagging, the recognition rate improves to 94.53%. The bagging VQ classifier outperforms the conventional VQ classifier even with small training databases.

Conclusion

In this paper, we proposed the bootstrap and aggregating VQ model. We studied the unstability of VQ model to apply the bootstrap and aggregaing method. The recognition rates were improved significantly in this model.

References

[1] L. Breiman, "Arcing Classifiers," The Annals of Statistics, vol. 26, no. 3, pp. 801-849, 1998.

[2] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, pp. 123-140, 1996.

[3] waveform database, *ftp.ics.uci.edu/ pub/machine-earning-databases*, 1988.

[4]TIMIT, *DARPA TIMIT - Acoustic-Phonetic Continuous Speech Corpus*, NIST, 1993.

[5] Y. J. Kyung and H. S. Lee, "Text-independent speaker identification using micro-prosody," Proc. of ICSLP '98, vol. 2 pp. 157-160, 1998.