분할 매트릭스 부호화를 이용한 문장 독립형 화자인식 시스템

Text Independent Speaker Identification Using Separate Matrix Quantization

경연정*, 이황수*

(Youn Jeong Kyung*, Hwang Soo Lee*)

※본 논문은 과학재단의 수탁과제 연구지원에 의해 수행되었습니다(과제번호:95-0100-22-01-3).

요 약

본 논문에서는 문장독립형 화자인식 시스템에 MQ(Matrix Quantization) 방법사용을 제안한다. 또한 인식율을 높이기 위해 MQ를 수정한 방법인 SMQ(Separated Matrix Quantization)를 제안한다. 기존의 VQ-distortion 방법은 대체로 좋은 성능을 가지나 화자의 동적 특성을 이용하지 못한다는 단점이 있다. MQ와 SMQ는 화자의 동적 특성을 이용할 수 있으므로 시간 변화에 대한 화자의 특징 변화까지 모델링 할 수 있는 장점이 있다. MQ는 여러 프레임을 묶어 Matrix Codebook을 가지며 SMQ는 MQ의 기본 codebook을 다시 켑스트럼의 차수에 따라 나누어 codebook을 만든다. 즉, 켑스트럼 차수를 저, 중, 고차로 나누어 각 부분별로 Matrix codebook을 만들도록 한다.

인식실험은 문장독립 음성 데이터에 대해 실행했으며 MQ모델의 경우 Matrix의 크기를 짧은 음소크기부터 음절단위까지 변화시켜 실험하였다. 아울러 SMQ 모델에서의 실험은 차수별 유용도를 보기 위하여 부분 차수를 이용하여 실험하였다. 실험결과 MQ와 SMQ방법이 VQ에 비해 좋은 성능을 가짐을 확인하였다.

ABSTRACT

In this paper, we propose separate matrix quantization(SMQ) for text-independent speaker identification. Since traditional speaker identification method using vector quantization(VQ) do not exploit the dynamic characteristics of human voice, the matrix quantization (MQ) which uses matrix codebook has been proposed to use those features such as pitch contours. To obtain the baseline system using MQ, we vary the rank of the matrix from three to ten and obtain improved recognition results compared to that using VQ. It yields the best performance for speaker recognition when the rank of the matrix is six in our simulation which is about one demi-syllable long. The recognition system with MQ uses all the cepstral coefficients for generating codebook. The proposed SMQ separates the cepstral coefficients into three parts according to the order of coefficients. We make separate codebook for each part of the cepstral coefficients. Simulation results show that the speaker identification system with SMQ yields the best performance.

I. Introduction

For text-independent speaker recognition, VQ-based methods were proposed many years ago. Burton[1] proposed the use of VQ coding Method to speaker recognition, Furui[2] and Kin Yu[3] compared the VQ model with other models. Many other researchers reported a VQ-based method that has good performance. Although VQ method has a good result, it can't make the best of speaker's dynamic features. We propose the use of MQ[1] method to text-independent speaker identification. Also we propose the SMQ method to improve the recognition rates.

II. VQ

For speaker recognition, a VQ codebook C is designed to minimize the average distortion that results from encoding a training sequence t

$$\sum_{p=1}^{P} d(\overline{t_p}, \overline{c_B})$$

where $\overline{c_B}$ is the codeword that results from encoding speech segment

* 한국과학기술원 정보통신공학부

접수일자 : 1998년 6월 1일

$$d(\overline{t_p}, \overline{c_B}) = \min_i d(\overline{t_p}, \overline{c_i})$$

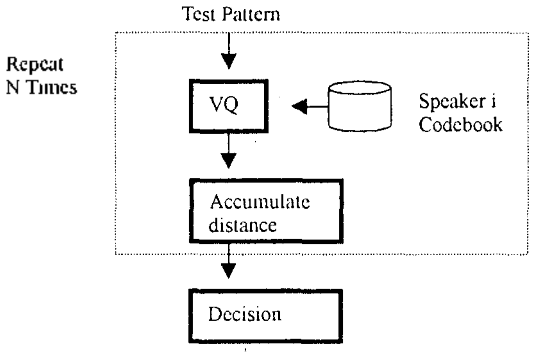
and d is an appropriate vector distortion measure. This codebook represents a speaker saying a particular word. The average distortion is defined as

$$\frac{1}{N} \sum_{i=1}^{N} d(\overline{t_i}, C_B)$$

where C_B is the codeword, t is the test utterance.

We use this average distortion in making the recognition decision.

The VQ-distortion based speaker recognition system is shown in Figure 1.



N : The number of speaker

Figure 1. ASR using VQ distortion.

III. MQ

In matrix quantization, instead of coding a single source vector in a codebook containing characteristic vectors, we code a time-ordered sequence of source vectors in a codebook containing characteristic vector sequences. Given t, we find the matrix quantization codebook C containing codeword vector sequences $c_j = [\overline{c_{j1}}, \overline{c_{j2}}, \cdots, \overline{c_{jk}}]$ that minimizes

$$\sum_{p=1}^{p-k+1} D(t_p, c_B)$$

Where c_B is the codeword matrix that results from coding the sequence of training vectors

$$t_p = [\overline{t_p}, \overline{t_{p+1}}, \cdots, \overline{t_{p+k-1}}]$$

by using the nearest neighbor rule

$$D(t, c_B) = \min_j D(t, c_j)$$

and where the distortion between a speech segment t and the jth codeword matrix is

$$D(t, c_j) = \sum_{l=1}^{K} d(\overline{t_l}, \overline{c_{jl}})$$

We call K the codeword matrix size(or rank).

To use MQ in speaker recognition, we represent each speaker by a codebook C, just as in the VQ approaches above. A recognition utterance is processed by dividing it into overlapping sequences of K frames, coding each K frame sequence in the speaker-codebook C, and computing the average quantization distortion between the utterance and the codebook. To be specific, for a recognition utterance v, the average distortion resulting from coding it with codebook C is

$$D_{avg} = \frac{1}{L-K+1} \sum_{l=1}^{L-K+1} D(\nu_l, c_B)$$

where $\nu_l = [\overline{\nu_l}, \overline{\nu_{l+1}}, \cdots, \overline{\nu_{l+k-1}}]$.

We can obtain the time-varing information using matrix quantization. In Figure 2, we show the MQ method as compared with VQ method.

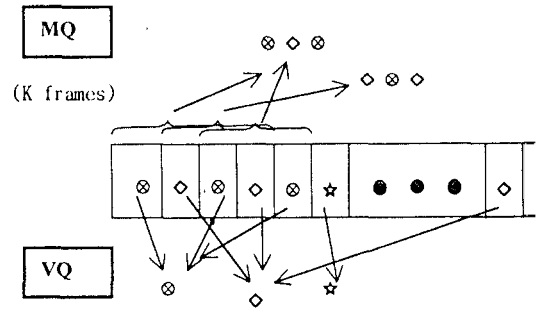


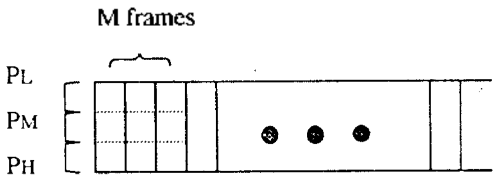
Figure 2. MQ & VQ method concepts.

IV. SMQ

To improve the recognition rate, we propose the SMQ (Separated Matrix Quantization). In SMQ, the codebook C is similar to MQ except that it is into parts of coefficients. In Figure 3, we show the SMQ method.

The codebook C is composed the three separated MQ codebooks.

The codebook C is designed to minimize



PL : the low part of cepstral coefficients'
PM : the middle part of cepstral coefficients'
PH : the high part of cepstral coefficients'

Figure 3. SMQ method

$$D_L = \sum_{i=1}^{M} \sum_{j=0}^{\frac{P}{3}} d(r_{ij}, C_{B1})$$

$$D_M = \sum_{i=1}^{M} \sum_{j=\frac{P}{3}}^{\frac{2P}{3}} d(r_{ij}, C_{B2})$$

$$D_H = \sum_{i=1}^{M} \sum_{j=\frac{2P}{3}}^{P} d(r_{ij}, C_{B3})$$

where P is the analysis cepstral dimension.

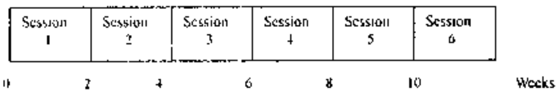
The average distortion D is defined as

$$D = D_L + D_M + D_H$$

where D_L is the average distortion of the low part's MQ, D_M is the average distortion of the middle part's MQ and D_H is the average distortion of the high part's MQ.

V. Experimental Results

The speech database consists of sentences by 20 speakers. It was recorded on six sessions every other week over twelve weeks. It is shown in Figure 4.



Training : 10 sentences in Session 1
Test : 1965 utterances
Total : 1975 utterances
The number of speakers : 20

Figure 4. Speech Database description.

Ten sentences from session 1 were used for training. 1965 utterances were used for evaluation. The duration of each sentence was very diverse. Database was recorded on common laboratory environment.

Our first tests concern to VQ and MQ. Figure 5 shows the result of experiments.

In the MQ model, we vary the matrix size K from three to ten. As noted earlier, MQ model's an utterance with a single codebook that contains an unordered set of time-ordered speech spectrum sequences. These spectrum sequences correspond to stable continuant sounds or transitions from one sound to another. Thus, for large K values, MQ method includes coarticulation and phonetic duration information. The performance of MQ model improves the recognition rate relative to the VQ.

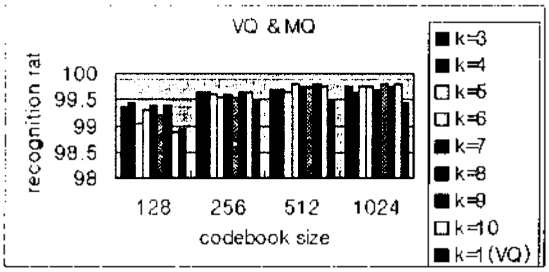


Figure 5. The comparison MQ method with VQ method.

In experimental results, 512 codebook size and 6 matrix rank are best for speaker recognition. The length of six-frame is about one demi-syllable length.

Our second tests are comparison MQ with SMQ. Table 2 shows the error rate of three models.

Table 2. The performance test of VQ, MQ, SMQ (error rates).

Model Codebook size	VQ	MQ	SMQ
256	2.04	1.53	1.53
512	3.57	2.55	2.04
1024	4.59	3.06	2.55

In this experiment, we use the partial speech database. Sixteen speakers uttered sixteen sentences. Five sentences are used to training. Through lack of training data, error rates increased in proportion to increase the codebook size.

Also we experiment the use of specific cepstral coeffic-

Table 3. The SMQ experiments (error rates).

using coeff. CB size	all coeff.	1st part	2nd part	3rd part	1st & 2nd	1st & 3rd	2nd & 3rd
256	1.53	4.08	8.67	12.24	2.55	2.55	6.12
512	2.04	3.06	7.14	11.22	1.53	2.55	4.59

ients in order to find the key part of cepstral coefficients for speaker recognition. We show the results in Table 3.

As results, the low part of cepstral coefficients is best for speaker recognition.

VI. Conclusions

This paper proposed the use of the MQ method and the SMQ method for speaker recognition.

The MQ and the SMQ can use the speaker's dynamic features. The suggested method could improve the recognition rate.

Acknowledgement

This study was supported in part by the KOREA SCIENCE AND ENGINEERING FOUNDATION. The contract number is 95-0100-22-01-3.

References

1. D.K.Burton, "Text-Dependent Speaker Verification Using Vector Quantization Source Coding," IEEE Tran. on Acoustics, Speech, and Signal Processing, Vol. ASSP-35, No. 2, pp. 133-143, February 1987.
2. T.Matsui and S.Furui, "Comparison of Text-Independent Speaker recognition methods using VQ-distortion and discrete/continuous HMMs," Proc. of ICASSP'92, pp.II-157-II-160, 1992.
3. Kin Yu et. al., "Speaker Recognition Models," Proc. of EUROSPEECH'95, 1995.

▲Youn Jeong Kyung

The Journal of the Acoustical Society of Korea(Vol. 15, No. 2E, 1996)

▲Hwang Soo Lee

The Journal of the Acoustical Society of Korea(Vol. 15, No. 2E, 1996)