지속적으로 발성한 모음에 의한 화자인식

Automatic Speaker Identification by Sustained Vowel Phonation

배 건 성*

G. S. Bae*

요 약

지속적으로 발성한 모음에 대해 각 화자의 특징을 나타내는 벡터양자화 코드북을 만들고 이를 이용해 화자를 인식하는 방법을 제안하고 실험하였다. 특징 벡터로는 모음 /이/로 부터 각각의 피치 주기에 대해 얻어진 선형예측계수를 사용하였으며, 코드북의 크기는 4가 적절함을 실험적으로 보였다. 인식실험에서, 학습에 사용된 데이타를 이용했을 경우에는 99.4%의 인식율을 보였으며, 학습에 사용되지 않은 50개의 피치 주기를 포함하는 음성신호로 부터는 89.4%의 인식율을 보였다.

ABSTRACT

A speaker identification scheme using the speaker-based VQ codebook of a sustained vowel is proposed and tested. With the pitch synchronous LPC vector of the sustained vowel /i/ as a feature vector, a VQ codebook size of 4 was found to be suitable to characterize each speaker's feature space. For 40 normal speakers (20 males, 20 females), we achieved the correct identification rate of 99.4% with a training data set, and 89.4% with a test data set with speech samples of only 50 pitch periods.

I. Introduction

Automatic speaker identification by machine has received a great deal of attention by speech researchers. The objective of a speaker identification system is to determine the indentity of the person by his/her voice from among a known population. The usefulness of identifying a person from the characteristics of his/her voice is increasing with the growing importance of automatic information processing and telecommunications between people and computers[1].

The pitch synchronous LPC analysis provides a good parametric representation for the very short segment of vowel phonations. Therefore, we used pitch synchronous LPC analysis of a vowel to extract feature vectors. We then investigated the intraspeaker and interspeaker variabilities of the pitch synchronous analyzed LPC spectral distortion

* Dept. of Electronics Kyungpook National Univ.
접수일자 : 1991. 10. 10.

for a sustained vowel phonation. As an efficient way of characterizing the speaker-specific features, we used the speaker-based VQ codebook approach [2].

This paper is organized in the following manner. We first describe the proposed speaker identification scheme in section Ⅱ. The experimental procedure and results are described with a discussion of our findings in section Ⅲ. Finally, the summary and conclusion are given in section Ⅳ.

Ⅱ. Speaker Identification Scheme

The proposed speaker identification system is shown in Figure 1. An input vector consists of a pitch synchronous LPC vector, x, and the matrix of correlation terms of speech samples, C_x, associated with the LPC vector x. In order to measure the similarity between two feature vectors, we extended the modified form of the Itakura-Saito distortion measure[3] for the LPC vectors obtained using the covariance method. It is given by

$$d(x, y) = (x-y)^t \, C_x \, (x-y) \tag{1}$$

The similarity between the test input and each speaker's codebook of N known speakers is calculated using the distortion measure,

$$D_i = \min_{1 \le j \le L} d(x, y_j) \quad , \ 1 \le i \le N \tag{2}$$

where x is the input LPC vector, y_j is the codeword of speaker #i's codebook, L is the size of the codebook and d(x,y) is a distortion measure defined in (1).

For the given input vectors, each input vector is cpmpared with N known codebooks using (2), and then assigned to the speaker who had the least distortion. Let NR_i represent the number of

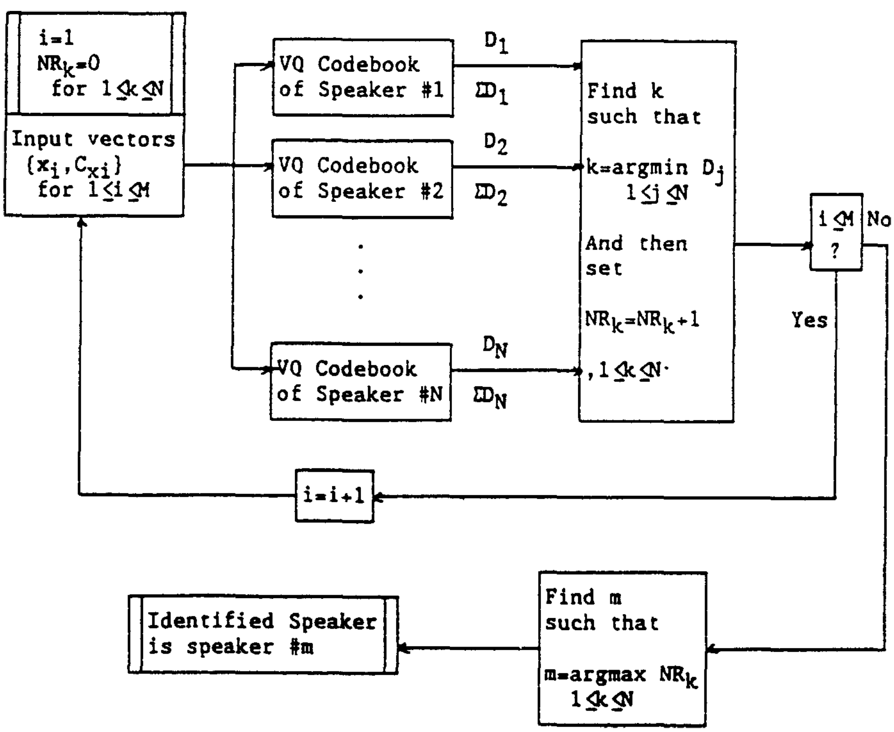

Figure 1. Block diagram of the proposed speaker identification scheme.

assignment occurrences to the speaker #i for the given input vectors, then the speaker with the largest value of NR becomes the identified speaker. That is, the identification decision is given by

$$\text{identified speaker \#} = \underset{1 \le i \le N}{\operatorname{argmax}} NR_i \qquad (3)$$

When more than one speakers have the same value of NR, the speaker who has the smallest average distortion for the total input vectors is chosen as the identified speaker.

To reduce the search time and computational burden of all the speakers' codebooks in half, we identified the input speaker's gender prior to determining the speaker's identity. The minimum distance classifier for the template of the LPC spectra for each gender, obtained from the training data set, was used to determine the input speaker's gender[4,5]. Therefore, we made two lists of known speakers having their own codebooks, one for male speakers and the other for female speakers, respectively. According to the input speaker's gender, the corresponding gender's codebooks were searched to identify the speaker using the procedure shown in Figure 1.

Ⅲ. Experimental Procedure and Results

1. Data Base

Over a period of one month, we collected speech and electroglottography(EGC) data for the sustained vowel /i/ from 40 speakers (20 males, 20 females) during four different recording sessions[4]. The 2nd, 3rd and 4th recordings were made two days after, one week after, and about one month after the first recording. Each session recorded the sustained vowel /i/ twice. Both speech and EGG signals were digitized synchronously with a sampling rate of 10kHz, respectively, and 16 bits / sample. We divided the 8 utterances from each speaker into two groups. From one group we obtained 400 sets of pitch synchronous LPC coefficients and correlation terms (100 sets from each utterance) and then they were used for training (VQ codebook generation). For the remaining group, we determinded 200 sets of pitch synchronous LPC coefficients and correlation terms (50 sets from each utterance) for testing.

2. Analysis Method

We did the pitch synchronous analysis using the covariance method for the speech samples with the following conditions :

Filter order	: 10 coefficients
Analysis frame	: 1 pitch period
Frame overlap	: none
Analysis window	: Hamming window
Speech preemphasis factor	: 0.9

For the normalization of correlation terms across subjects, the speech signal was normalized using the squared energy, after removing the mean, on

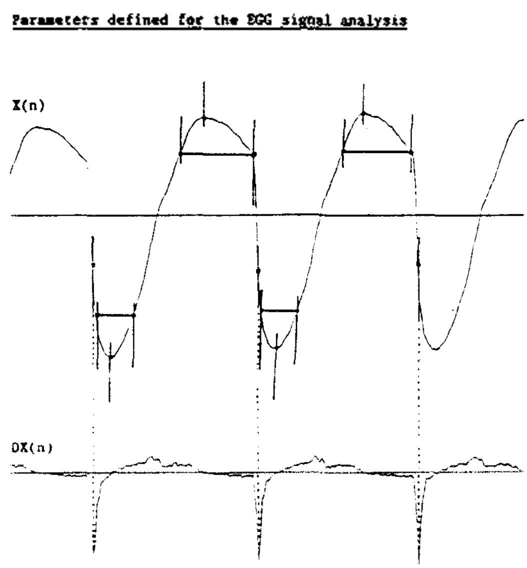

Figure 2. Typical EGG and differentiated EGG waveforms during a vowel phonation.

a pitch period basis. Pitch synchronous LPC analysis was done with an aid of differentiated EGG signal for the exact pitch period detection. It is known that the point of maximum negative value in a differentiated EGG signal agrees well with the closing time of vocal folds[6]. Typical EGG and differentiated EGG waveforms for the sustained vowel phonation are shown in Figure 2.

To evaluate the effect of different speaker identification parameters on the performance, we varied :

- The size of the VQ codebook
  We used the codebook size of 1, 2, 4 and 8.

- The number of test input vector
  We performed the speaker identification tests varying the number of test input vectors, i.e., the length of a test speech signal in time domain, from 10 to 50 pitch periods.

- The time span between the training and testing material
  By using the first two recording sessions for training, we examined the effect of intraspeaker variation to the identification performance.

3. Experimental Results

Effect of VQ codebook size. Figure 3 shows the effect of codebook size on the mean and standard deviation of the VQ distortion obtained from the training data set of 40 speakers. A good separation was shown between intraspeaker distortion and interspeaker distortion. Intraspeaker distortion decreased greatly wen codebook size increased from 1 to 4, while interspeaker distortion decreased slightly as codebook size increased. A further illustration on the effect of codebook size to the VQ distortion is given in Figure 4, which shows the averages, standard deviations, and histograms of intraspeaker and interspeaker distortions for codebook size 1 and 4. Both codebooks gave good separation between intra-and interspeaker average distortions across all the speakers. The average distortion with codebook size 4 gave better separation than that of codebook size 1.

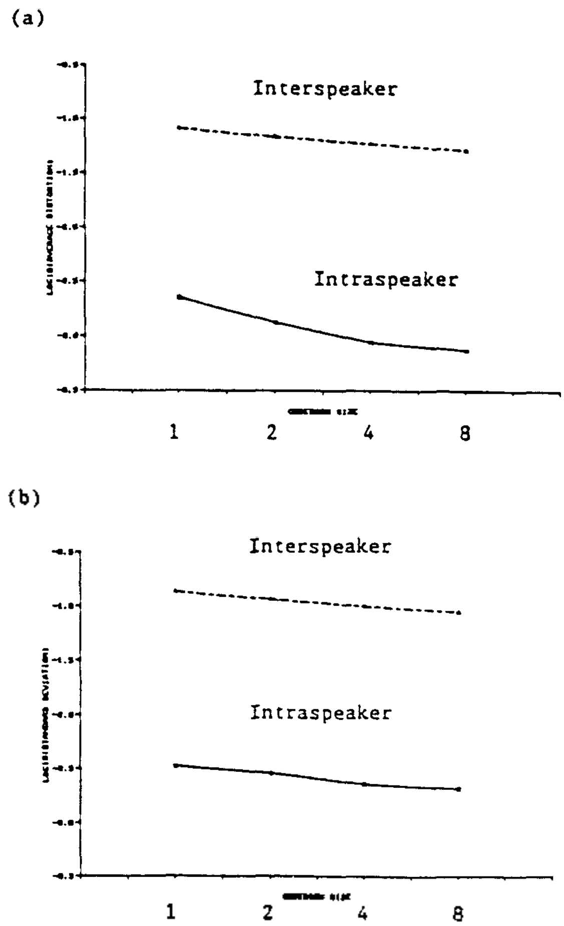

Figure 3. Effect of the codebook size.
(a) Average VQ distortion versus codebook size.
(b) Standard deviation of the distortion versus codebook size.

The speaker identification error rate is plotted as a function of codebook size in Figure 5. The identification error rate decreased when the codebook size increased from 1 to 4. However, increasing the codebook size from 4 to 8 did not reduce the identification error rate.

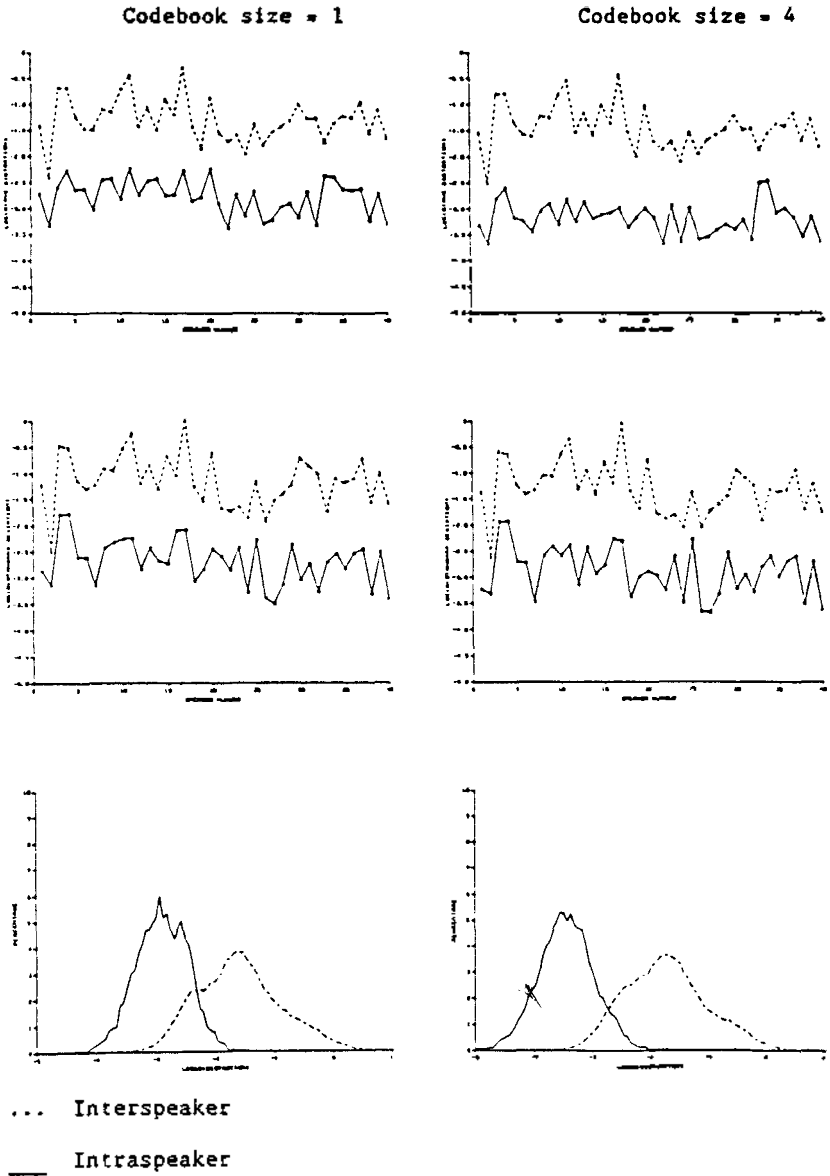

Figure 4. Average, standard deviation, and histogram of the VQ distortions for different codebook size.

Effect of test vector length. The identification error rate versus different test vector length is shown in Figure 6. The result shows that the identification error rate decreased slightly as the test vector length increased. However, at the test vector length of 30, the identification error increased.

Effect of different recording sessions. The identification error rate plotted as a function of the recording session-number is shown in Figure 7. The codebook was generated from the 200 LPC vectors, obtained from the first two recording sessions. Since the first two sets of test vectors are obtained from the utterances recorded in the first two sessions, they gave a significantly better result than other two test sets. Figure 8 shows the identification error rate versus the total number of recording sessions used for training the VQ codebook. The error rate decreased as more recording sessions were used for training.

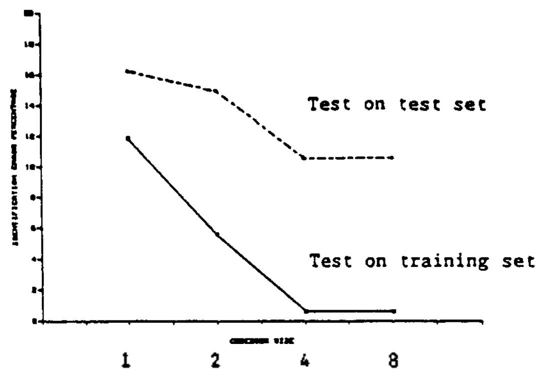

Figure 5.Speaker identification error rate versus codebook size.

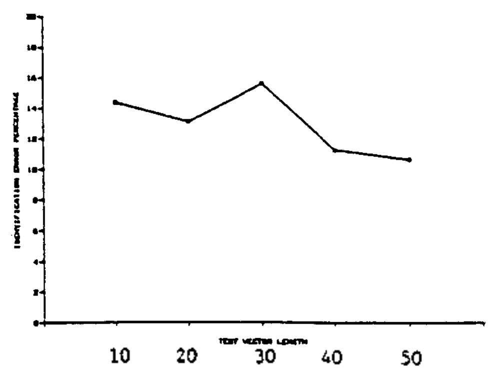

Figure 6. Speaker identification error rate versus test vector length.

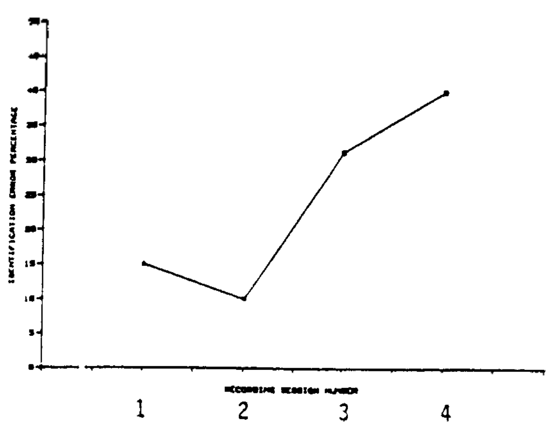

Figure 7. Speaker identification error rate versus recording session number.

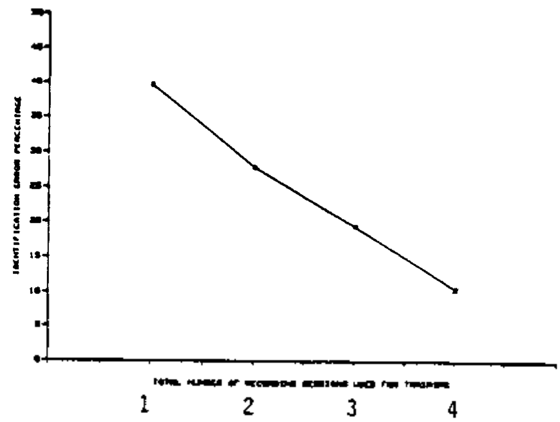

Figure 8. Speaker identification error rate versus total number of recording sessions used for training .

Ⅳ. Summary and Conclusion

We proposed a speaker identification scheme using the speaker-based VQ codebook of the sustained vowel. With the LPC vector of the sustained vowel as a feature vector, the codebook size of 4 was found to be suitable to represent each speaker's feature space. With a codebook size of 4, we achieved a correct identification rate of 99.4% for the training data set, and 89.4% for the test data set. We determined that the length of the feature vector (number of speech samples) did not greatly affect the identification performance. Therefore, our speaker identification scheme may be applicable to vowel samples extracted from the running speech. The duration of speech samples used for training the VQ codebook from each utterance was approximately 0.4 to 1.0 sec depending on each speaker's pitch period. The duration of speech samples used for testing was about 0.2 to 0.5 sec.

The experimental results for the effect of different recording sessions(i.e., the interval between recordings) indicated that even for sustained vowels, large variations occur within a speaker over time. This is in agreement with the results of that used isolated digit utterances. Thus, one needs to update the VQ codebook or to include sufficient intraspeaker variability for training the VQ codebook.

The proposed speaker identification system shows promise, especially when we consider that speaker was identified from the population of 40 speakers using only a single vowel phonation. However, identification error rate varied greatly from speaker to speaker. Further studies to reduce the effects of speaker dependence on the system performance are needed. Consideration could be given to using more than one vowel or to varying codebook size depending upon each speaker's average distortion.

Reference

1. G.R.Doddington, "Speaker recognition Identifying people by their voices," Proceedings of the IEEE, Vol. 73, pp.1651-1664, 1985.
2. F.K.Soong, A.E.Rosenberg, and B.H.Juang, "A vector quantization approach to speaker recognition", AT&T Technical Journal, Vol. 66, pp.14-26, 1987.
3. J.Makhoul, S.Roucos, and H.Gish, "Vector quantization in speech coding", Proceedings of the IEEE, Vol. 73, No. 11, pp.1551-1588, November, 1985.
4. K.S.Bae, Two channel (Speech and EGG) analysis with the application to evaluation of laryngeal function and speaker identification by voice, Ph.D. Dissertation, University of Florida, 1989.
5. D.G.Childers, K.Wu, K.S.Bae, and D.M.Hicks, "Automatic recognition of gender by voice," in Proc. IEEE International Conf. Acoust., Speech, and Signal Processing, Vol. 1, pp.603-606, 1988.
6. D.G.Childers, C.P.Moore, J.M.Naik, J.N.Larar, and A.K.Krishnamurthy, "Assessment of laryngeal function by simultaneous, synchronized measurement of speech, electroglottography, and ultra high speed film, "Edited by L.Van Lawrence, Transcripts of the Eleventh Symposium Care of the Professional Voice, New York, NY, pp.234-244, 1982.

Keun Sung Bae was born in 1953. He received the B.S. degree in electronics engineering from the Seoul National University, in 1977, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, in 1979, and the Ph.D. degree in electrical engineering from the University of Florida, Florida, S.A., in 1989.

He has been with the Kyungpook National University, Taegu, Korea, since March 1979, where he is currently an associate professor in the department of Electronics. His research interests inolude speech analysis/synthesis, speech reoognition, speech coding, varioua areas of digital signal processing, and digital communication systems.