Noisy Speech Recognition Based on Spectral Mapping Techniques

스펙트럼사상기법을 기초로 한 잡음음성인식

Ki Young Lee*
이 기 영*

ABSTRACT

This paper presents noisy speech recognition method based on spectral mapping techniques of speaker adaptation method. In the presented method, the spectral mapping training makes the spectral distortion of noisy speech reduced, and for the more correctively spectral mapping, let the adjustment window's slope be adaptive to several word lengths. As a result of recognition experiment, the recognition rate is higher than that of the conventional method using VQ and DTW without noise processing. Even when SNR level is 0 dB, the recognition rate is 10 times more than that using the conventional method. It is confirmed that the speaker adaptation technique using the spectral mapping training has an ability to improve the recognition performance for noisy speech.

요 약

본논문에서는 화자적응방법에서의 스펙트럼사상기법을 기초로한 잡음인식방법을 제시하였다. 제시한 방법에서는 스펙트럼사상에 의하여 잡음음성의 스펙트럼왜곡을 감소시키며, 스펙트럼을 더욱 정확히 사상하기 위하여 정합창의 기울기로 하여금 여러 단어의 길이에 적응하도록 하였다. 인식실험의 결과, 잡음처리를 하지 않는 VQ와 DTW 를 이용한 기존의 방법보다 높은 인식율을 얻었으며, 0 dB 의 SNR 레벨에서도 기존방법의 인식율을 10배 이상으로 향상시키므로써 스펙트럼사상을 이용한 화자적응기법이 잡음음성의 인식성능을 향상시킬 수 있음을 확인하였다.

I. INTRODUCTION

Recently, the wider the application range of the speech recognizer, the more the necessity of the one robust to noise, because of the speech overlapped with noise from the external machine or the outside environment, and distorted through microphone or other channel with noise. According to the results of Acero's study[1], it is found that if speaker-independent speech recognizer was in noise environment, the recognition performance was very lowered. The conventional noisy speech recognizer has mainly utilized the noise reduction method using the statistical characteristics of speech and noise.

Boll[2] attempted to cancel noise from speech using the spectrum reduction method with DFT coefficients. Stockham[3] used the spectrum equivalent circuit as the noise compensation method. Van Compernolle[4] utilized both the spectrum re-

*Dept. Electronic Comm. Eng., Kwan Dong Univ.
관동대학교 전자통신공학과
접수일자 : 1994년 9월 13일

duction method and the spectrum equivalent circuit and won the good results. However, these methods had a defect that the independence for noise spectrum estimates must be assumed. Besides, the adaptive noise cancelling [5] method has been studying, but this method requests many computations or multi-sensors [6].

In this paper, we noticed the point of improving the recognition performance for unknown speaker's speech by the spectral mapping training based on speaker adaptation techniques [7,8], and present the noisy speech recognition method to make better performance through the spectral mapping training from one space of noisy speech spectrum to another space of speech spectrum without noise. So that spectral distortion of noisy speech is reduced and the reognition rate is better than that of the conventional mehtod using VQ and DTW without noise processing. In the spectral mapping procedure, we use DTW [9] which is adaptive to several word lengths so that the correspondent relationship is more correctly. This recognition method has two merits of applying the recognizer without changing system under any noise environment, and having a less influence on the noise characteristics in the recognition procedure because the mapped codevectors are linearly combinated with clean speech spectra.

Ⅱ. NOISY SPEECH RECOGNITION

When speech is not distorted by noise, speech recognition system can achieve very high performance. However, to recognize noisy speech is very difficult, because noise damages speech spectra which are mainly used for important feature parameters of recognition system. In speaker adaptation technique, one spectrum space of test speaker's speech is mapped onto another spectrum space of standard speaker's. This technique, in this paper, is applied to make noisy speech clean.

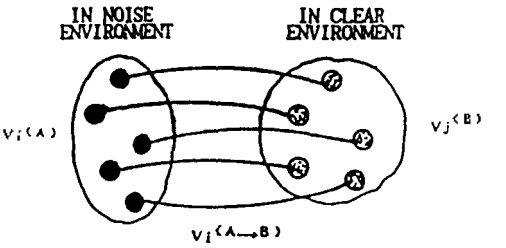

Fig.1 Basic concept of spectral mapping

1. SPCTRAL MAPPING TRAINING

For spectral mapping training, a conversion work firstly needs obtaining the mapping relationship from one space of noisy speech spectrum to another space of speech spectrum without noise. This mapping is to find correspondent probabilities between two codebooks from clean speech and noisy ones. The correspondence is represented by histogram which is obtained from a training procedure in the way of the optimal path by DTW. Speech have an infinite spectrum space, so VQ is used to make an infinite spectrum space to a finite one. Basic concept described above is shown in Fig.1.

The mapped codebook is made by linear-combination of histogram and a clean codevectors. The training procedure for the spectral mapping training is as follows.

(a) Distortion matrix is made between clean codebook and noisy codebook which are generated from training words.

(b) Training noisy speech and clean one are vector-quantized with noisy codebook and clean one, respectively.

(c) Histogram,hij by eq(1), is obtained from correspondent codevectors of the optimal path of DTW between noisy and clean words.

$$h_{ij} = h_{ij} + 1 \tag{1}$$

(d) Mapped codebook is obtained by eq. (2).

$$V_i^{(A\to B)} = \sum_{j=1}^{L} b_{ij} V_j^{(B)} \Big/ \sum_{j=1}^{L} b_{ij} \qquad (2)$$

where, $V_j^{(B)}$ is the j-th codevector of clean codebook,
$V_i^{(A\to B)}$ is the i-th codevector of mapped codebook,
L is the codebook size.

(e) Replace $\{V_i^{(A)}\}$ by $\{V_i^{(A\to B)}\}$.

(f) If average distortion is converged, end this procedure. Otherwise, then go to step (b).

2. RECOGNITION PROCEDURE

As described above, mapped codebook is made through the spectral mapping training, and in the recognition procedure, noisy codevectors of input word are replaced by mapped codevectors. And DTW is used for recognition technique with a distortion matrix between clean codebooks and mapped codebook, and the decision rule is kNN which is a rule to select a pattern of minimum distortion for multi-template of each word. The recognition procedure is as follows.

(a) Input noisy speech is vector-quantized with noisy codebook.

(b) Noisy codevectors are replaced by mapped codevectors.

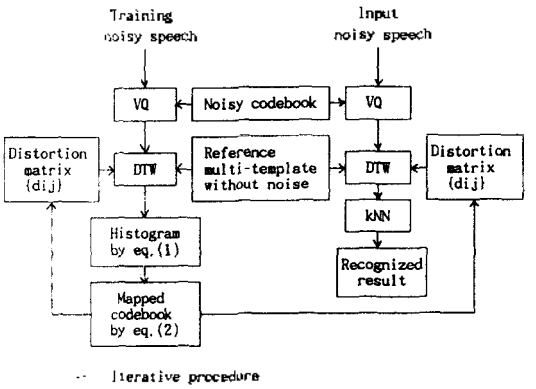

Fig.2 Block diagram of noisy speech recognition system

(c) Distortion matrix is made between clean and mapped codebook.

(d) Input noisy speech is recognized through DTW and kNN.

Fig.2 shows a block diagram of noisy speech recognition system.

Ⅲ. EXPERIMENTS AND RESULTS.

1. Database

Speech data is Korean digits(10) and spoken by 2 males 10 times ($10\times2\times10=200$) in computer laboratory room without soundproofing. Noisy speech is made by adding white noise whose SNR levels are 20, 10 and 0 dB. Training words were 10 digits first-time spoken.

Words in this experiments are low-pass-filtered with 4.5kHz cutoff frequency, and A/D converted with 10kHz sampling rate and 12-bit resolution. Sampled speech signals are passed through the preemphasis filter whose transfer function is $(1-0.95z^{-1})$. Frame length is 200 samples and interval is 100 samples, so that analysis is superpositioned upon 50%, repeatedly. The order of LPC analysis is 14, and the generating method of codebook is LBG algorithm[11]. The distortion measure in this paper is likelihood ratio[12].

For the more correctly correspondent relationship, the adjustment window's slope needs to be

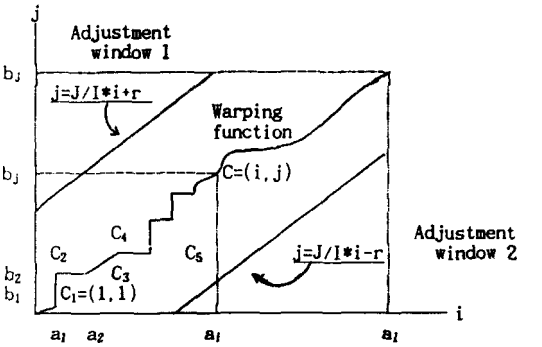

Fig.3 Adaptive adjustment window of DTW

adaptive to several word lengths. Fig.3. shows a adaptive adjustment window to lessen errors [9] on the path searched by DTW. Slope constraint con dition is p=1, and symmetric path is used. Reference multi-templates are constructed with each digit 3 times spoken, and the rests are used for test patterns of input noisy speech.

There is a special case in this experiment that a Koean digit is consisted of only one syllable. All digits are made by 5 vowels and 8 consonants. That is, total phonemes of digits is 13. To select codebook size is the problem because each phoneme features are represented by codebook. If codebook size is less than phoneme's number, each phoneme is not properly vector-quantized. To process Korean single digits, it is proper to select codebook size more than 13.

2. Examination of spectral distortion

Spectral distortions between clean and noisy speech in the training procedure are examined according to iterative numbers. If the spectral mapping training is very effective, the spectral distortion is well or quickly converged. So spectral distortion versus iterative numbers are compared, as shown in Fig.4. This figure shows spectral distortiion versus iterative numbers in training procedure when codebook size is 8, 16 and 32, respectively. In this figure, spectral distortion is similar, when codebook size is 16 and 32, and the iterative number till to convergence is averaged to 4. However, spectral distortion reaches toward minimum value, when iterative number is 1. Recognition performance will be examined when spectral distortion is minimum in the training procedure.

3. Experimental results and discussion

In this section, recognition results are compared and examined when noisy speech is input to the recognition system presented in this paper. Fig.5, 6 and 7 show the recognition rates when SNR level is 20, 10 and 0 dB, and compares them

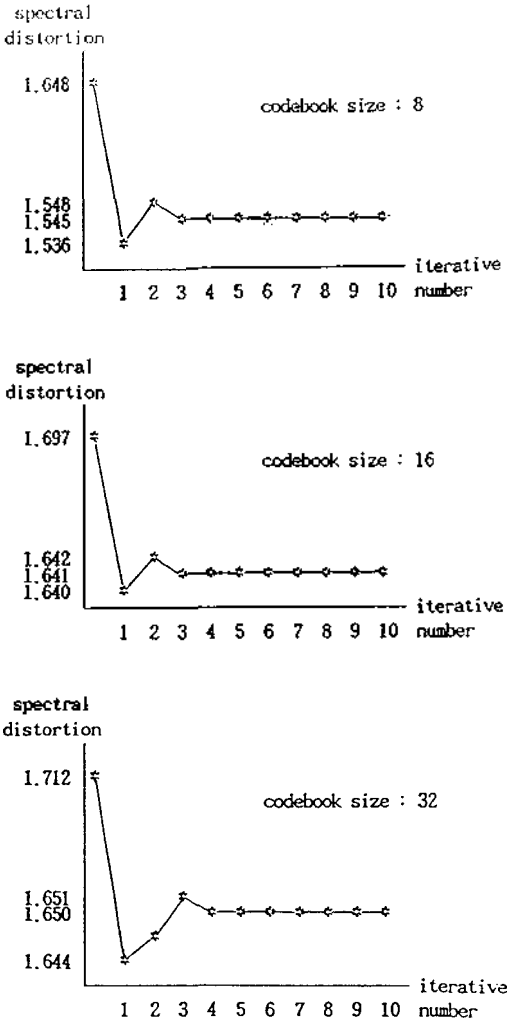

Fig.4 Spectral distortion versus iterative numbers

between several methods, where 'CLEAN' means that input words are not damaged by noise, 'NOISE' means that input words are noisy and not trained by the spectrum mapping, 'L.MAPPED' means that the iterative number for spectral mapping training is 1 and 'N.MAPPED' means that the iterative number is, at least, 4 until to converge.

When codebook size is 32, the rate is not so much improved as when 16, because the codebook has too many codespectra damaged by noise and spectrum features of consonants are similar

to one of noise, so the mapping relationship between clean and noisy codebook is not correctly matched. This means that, for Korean digits, codebook size is more than 13 as described before in the section III-1.

These results show that when using the spectral mapping training, the recognition rates are better than those without this training. The reason of higher performance using 'N.MAPPED' method than that using 'L.MAPPED' method is because the more the iterative number, the less the effect of noise due to the renewal mapped codebook including spectra without noise.

When SNR level is 0 dB, the recognition rate using 'N.MAPPED' method is improved about 10 times more than the one using 'NOISE' method, regardless of many codebook sizes. In fig.7, for example, when using 'NOISE' method, the recognition rate is 5% on codebook size is 16, but, when using 'N.MAPPED' method, 55% is obtained.

20[dB]

CBK SIZE	8		16		32		Avg.	
METHODS	TOP 1	TOP 2	TOP 1	TOP 2	TOP 1	TOP 2	TOP 1	TOP 2
CLEAN	91.6	98.3	96.6	98.3	88.3	93.3	92.3	96.6
NOISE	41.6	63.3	86.6	96.6	38.3	63.3	55.5	74.4
L.MAPPED	43.3	78.3	78.3	93.3	53.3	66.6	58.3	79.4
N.MAPPED	48.3	73.3	90.0	96.6	71.6	85.0	70.0	85.0

10[dB]

CBK SIZE	8		16		32		Avg.	
METHODS	TOP 1	TOP 2	TOP 1	TOP 2	TOP 1	TOP 2	TOP 1	TOP 2
CLEAN	91.6	98.3	96.6	98.3	88.3	93.3	92.3	96.6
NOISE	30.0	43.3	48.3	61.6	25.0	55.0	34.4	53.3
L.MAPPED	48.3	81.6	73.3	93.3	43.3	81.6	54.9	85.5
N.MAPPED	50.0	90.0	90.0	98.3	61.6	80.0	67.2	89.4

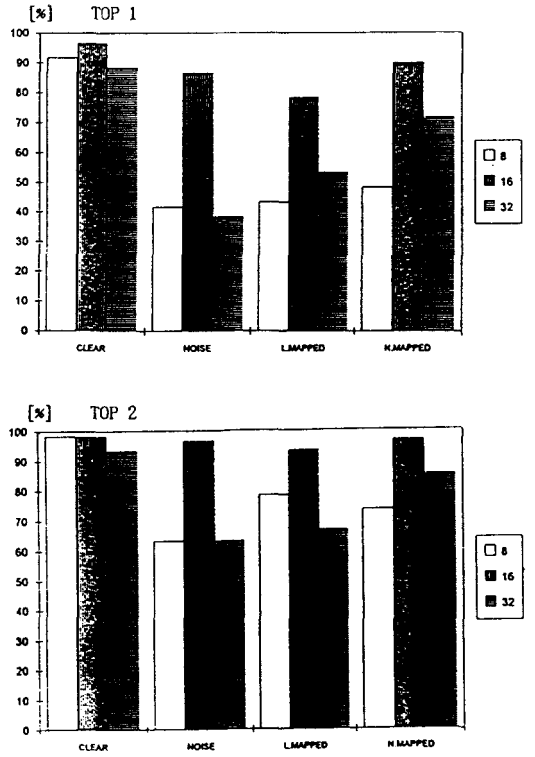

Fig.5 Comparison of recognition rates, SNR = 20[dB]

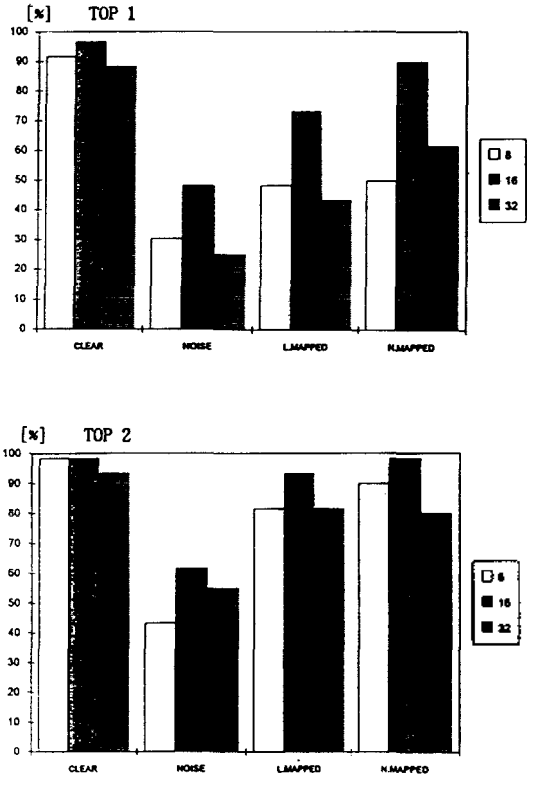

Fig.6 Comparison of recognition rates, SNR = 10[dB]

0[dB]

CBK SIZE	8		16		32		Avg.	
METHODS	TOP 1	TOP 2	TOP 1	TOP 2	TOP 1	TOP 2	TOP 1	TOP 2
CLEAN	91.6	98.3	96.6	98.3	88.3	93.3	92.3	96.6
NOISE	3.3	3.3	5.0	8.3	3.3	26.6	3.6	12.7
L.MAPPED	15.0	18.3	25.0	61.6	28.3	51.6	22.7	43.8
N.MAPPED	36.6	76.6	55.0	86.6	60.0	83.3	50.5	82.1

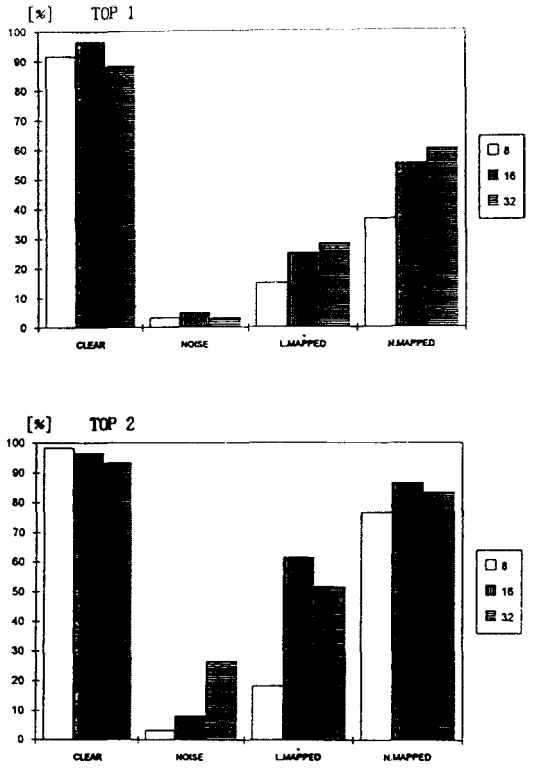

Fig.7 Comparison of recognition rates, SNR = 0[dB]

Ⅳ. CONCLUSION

Spectral mapping training for noisy speech recognition is presented and examined its validity through experiments using Korean digits adding several white noise levels. The experimental results are as follows.

(1) For Korean digits, when codebook size is 16, the recognition performance is much improved.

(2) The lower SNR level, the higher improved recognition rate than the one using the conventional method.

Therefore, it is confirmed that spectral mapping training of speaker adaptation techniques has an ability to improve the recognition performance for noisy speech.

REFERENCES

1. A. Acero, R. M. Stern, "Environmental Robustness in Automatic Speech Recognition," Proc. ICASSP 90, pp. 849-852, 1990
2. J. Porter, S. F. Boll, "Optimal Estimators for Spectra Restoration of Noisy Speech," Proc. ICASSP84, pp. 18A. 2. 1-4., 1984
3. T. G. Stockham, et al., "Blind Deconvolution Through Digital Signal Processing," Proc. IEEE, Vol. 63, pp. 678-692, 1975
4. D. Van Compernolle, "Noise Adaptation in a HMM Speech Recognition System," Computer, Speech and Language, Vol. 3, pp. 151-167, 1989
5. B. Widrow, et al., "Adaptive Noise Cancelling Principles and Applicationa," Proc. IEEE, Vol. 63, pp. 1692-1716, 1975
6. V. R. Viswanathan, C. M. Henry, "Noise-immune Multisensor Speech Input : Formal Subjective Testing in Operational Conditions," ICASSP89, pp. 373-376, 1989
7. K. Shikano, K. F. Lee, R. Reddy, "Speaker Adaptation Through Vector Quantization," Proc. ICASSP 86, pp. 49. 5, 1986
8. R. M. Stern, "Dynamic Speaker Adaptation for Feature-Based Isolated Word Recognition," IEEE Trans. on ASSP, pp. 751-763, 1987
9. M. K. Brown, L. R. Rabiner, "An Adaptive, Ordered, Graph Search Technique for Dynamic Time Warping for Isolated Word Recognition," IEEE Trans. on ASSP, pp. 535-544, 1982
10. J. D. Markel, *Linear prediction of Speech*, Spring-Verlag Berlin Heidelberg New York 1976
11. Y. Linde, A. Buzo, R. M. Gray, "An Algorithm of Vector Quantization Design," IEEE Trans Comm., Vol. COM-28, pp. 84-108, 1980

12. A. H. Gray, Jr., J. D. Markel, "Distance Measure for Speech Processing," IEEE Trans. on ASSP, pp. 380-394, 1976

▲Ki Young Lee

Dept. Electronic Comm. Eng., Kwan Dong Univ.