TEXT INDEPENDENT SPEAKER RECOGNITION USING MICRO-PROSODY

*YounJeong Kyung, **Hwang-Soo Lee

*, ** Dept. of Information & Communication Engineering, KAIST

** Central R&D Center, SK-telecom

ABSTRACT

The acoustic aspects that differentiate voices are difficult to separate from signal traits that reflect the identity of the sounds. There are two sources of variation among speakers: (1) differences in vocal cords and vocal tract shape, and (2) differences in speaking style. The latter includes variations in both target vocal tract positions for phonemes and dynamic aspects of speech, such as speaking rate. However, most parameters and features are in the former.

In this paper, we propose the use of a prosodic feature that represents micro prosody of utterances. The robustness of the prosodic feature on noise environment becomes clear. Also we propose a combined model. The combined model uses both the spectral feature and the prosodic feature. In our experiments, this model provides robust speaker recognition in noise environments.

1. INTRODUCTION

Although many recent advances and successes in speaker recognition have been achieved, many problems remain. Most of these problems arise from variability, including speaker-generated variability and variability in channel and recording conditions[1]. As for background environment variations, prosodic features and speaking style are not changeable in contrast with spectral features. The performance of spectral features diminishes in noise environments but prosodic features remain robust. This paper focuses on the use of prosodic features. The main points of the paper are as follows: First, we propose the use of a prosodic feature, which represent micro prosody of utterances. Second, we propose a combined model. The combined model uses both the spectral features and prosodic features. In our experiments, this model provides robust speaker recognition on noise environment.

2. SPEAKER RECOGNITION SYSTEM USING SPECTRAL FEATURES

VQ is a source coding technique that has been used successfully in both speech coding and speech recognition. In VQ, each source vector is coded as one of a pre-stored set of codewords, called a codebook, by finding the codeword that minimizes the distortion between itself and the source vector. For speaker recognition, a single-section VQ codebook C is designed to minimize the average distortion[2]. VQ-distortion based speaker recognition method has many advantages[3].

The performance of the VQ model depends on the codebook size. Figure 1 shows the results of speaker recognition experiments by VQ model.

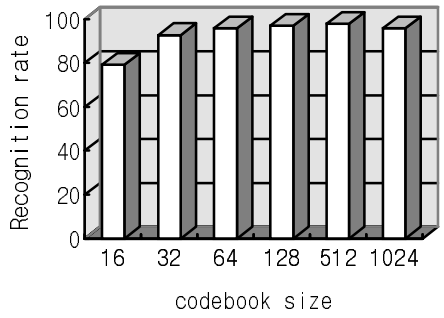


Figure 1 The VQ model performance test results vary according to codebook size

With the VQ-distortion method, a codebook size of 128 is enough for speaker recognition under experimental conditions. We refer to [4] and we select the LPC-based Mel-cepstrum for spectral feature. Figure 2 shows the performance of the VQ model under noise environments.

Figure 2 shows that the VQ model is degraded in noise environments. In the figure, C6020dB is speech mixed with car noise, SNR 20dB. WGN10dB is the speech mixed with white gaussian noise, SNR 10dB.

3. SPEAKER RECOGNITION SYSTEM USING PROSODIC FEATURES

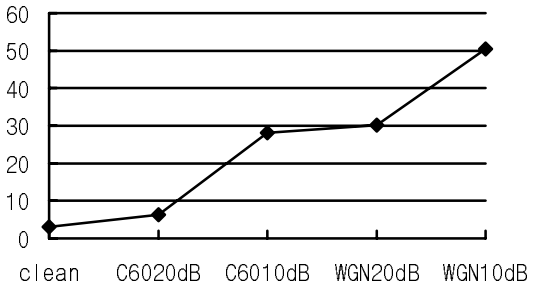
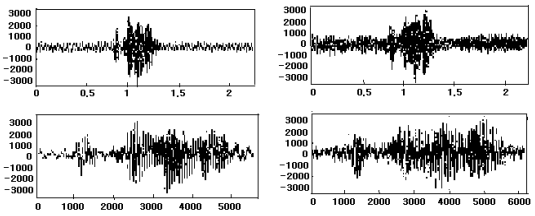


Figure 2 The VQ model performance tests results using spectral feature in noise environments (error rates)

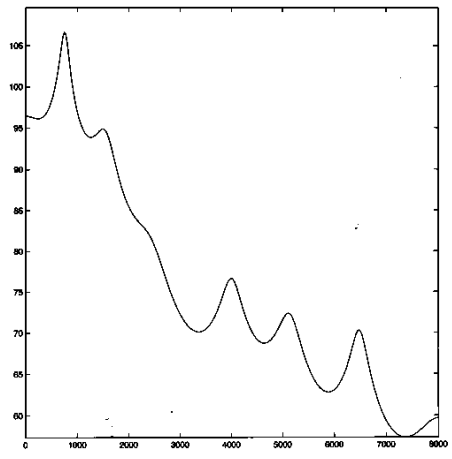
Some researchers have studied prosodic feature[5]-[7]. The acoustic measures of prosodic behavior can be divided into statistical and dynamic measures. Most studies focus on text dependent speaker recognition.

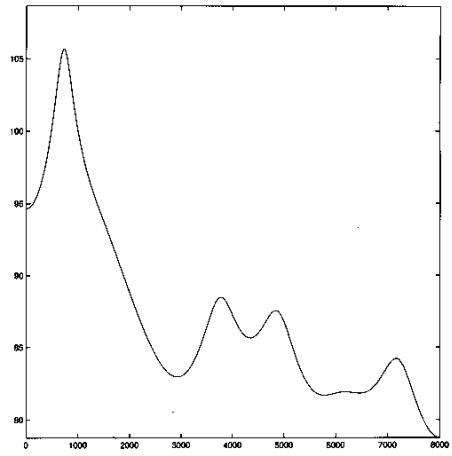
We observe the robustness of the prosody feature in noisy speech. Figure 3 shows this robustness. In Figure 3, (a) is the waveforms. The clean speech data are shown on the left and the noisy speech data are on the right. In (b) LPC envelopes of clean data and noisy data are displayed. In (c) the pitch contours are shown.

(a) waveform

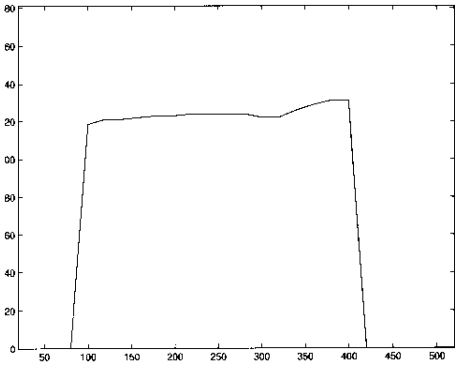


(b) spectrum





(c) pitch contour



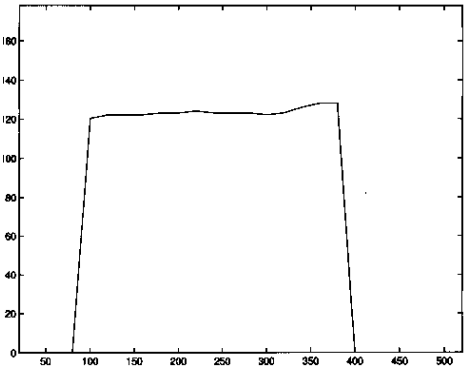


Figure 3 Clean speech data and noisy speech data

In this figure, the robustness of the prosody feature appears. Therefore, we apply the prosodic information to text independent speaker recognition. We observe the micro

prosody on sentences. First, we extract the pitch sequences from the utterances by the MBE pitch detection algorithm[8]. Next we make the VQ codebook using pitch sequences that cut off uniform units. We segment the sequences into P dimension vector units. The next segment is then overlapped as the Q dimension. When the test utterance is entered, we compare the test utterance's pitch sequences with the speaker's prosodic feature in the VQ codebook.

Finding the suitable size for P is critical. If P is too long, we compare the whole sentence. When P is too short, we can't expect the micro prosody. In Table 1, we vary the P value to find the best dimension size.

Table 1 Error rates depend on dimension of segmental pitch contour

CB	9	12	16	18	20
32	32.81	31.64	25.78	31.25	32.03
64	22.27	38.28	21.87	24.22	23.83

We find that the optimal dimension size is sixteen, which is about one syllable length. That is, we use the micro prosody of one syllable. The next experiment concerns the codebook size. In this experiment, we vary the codebook size on the fixed dimension size, sixteen. Figure 4 shows the results.

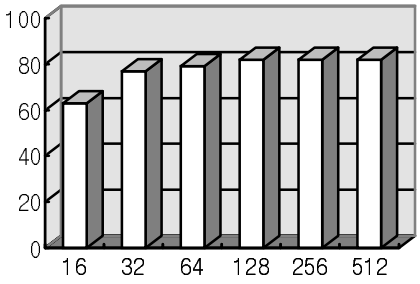


Figure 4 Speaker identification rates vary according to codebook size

We choose 128 for the codebook size. The last experiment relates to noise environments. We test the performance of the prosodic fieature on various noisy speech data. The experimental results are shown in Figure 5.

Although using the prosodic feature is not as good as using the spectral feature in clean speech, the former is more powerful than the latter in noise environments. Therefore, we advocate the use of the prosodic feature. Also, we propose a combined model in next section.

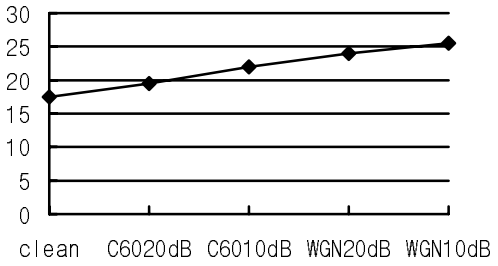


Figure 5 The VQ model performance tests results using prosodic feature in noise environments (error rates)

4. COMBINED MODEL

The speech database consists of 256 short sentences (about 2sec. long) pronounced by 8 male and 8 female speakers in a common laboratory environment. Each speaker uttered 16 sentences. Ten clean speech sentences were used to construct the training vector. Five clean speech sentences and 1024 (256 X 4 noise types) sentences were used for testing.

To construct a combined model, we must consider feature normalization. We use both the spectral feature and the prosodic feature. Their dynamic range is very different. We perform the variance normalization to the features. In Section 3, we already found the effectiveness of the prosodic feature in noise environments. However, we still have not determined how much more effective the prosodic feature is than the spectral feature. In this paper, combined model's distance measure is defined as follows:

Final_distance
= alpha X DSPEC + (1-alpha) X DPROS

Where DSPEC is distance of spectral feature, DPROS is distance of prosodic feature. Alpha is the weighting factor.

Experiments were done to find the optimal alpha weighting value. Table 2 shows that the recognition rate depends on the alpha value. In our experiments, the VQ model is trained with a clean speech database. Four different noisy speech databases were used for testing.

Table 2 Recognition rates for the combined model depend on alpha value

DB	0.1	0.2	0.3	0.4
C6010dB	92.6	92.2	91.0	90.6
C6020dB	97.7	97.7	97.7	97.7
WGN10dB	81.6	73.8	69.1	65.2
WGN20dB	91.8	91.8	90.6	90.2
TOTAL	90.9	88.9	87.1	85.9

0.5	0.6	0.7	0.8	0.9
90.2	90.2	89.8	89.5	89.5
97.7	97.7	97.7	97.7	97.7
60.9	58.6	56.6	52.7	51.2
90.6	90.2	90.2	89.8	89.1
84.9	84.2	83.6	82.4	81.8

5. CONCLUSIONS

VQ-distortion models usually use the spectral feature. Its performance is very good in clean environments, but is seriously degraded in noise environments. In this paper, we propose the use of the micro-prosodic feature. We make the codebook from segmental pitch contours. Also we construct a combined model. The performance of the combined model is better than that for the system using only spectral features. Moreover, the combined model is more robust than other models in noise environments.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

1. R.A.Cole et al., *Survey of the State of the Art in Human Language Technology*, USA, 1995.
2. 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.
3. Kin Yu et.al., "Speaker Recognition Models," Proc. of EUROSPEECH'95, 1995.
4. KAIST, *Man-Machine interface by spoken language*, Technical Report, 1997.
5. B.S.Atal, "Automatic Speaker Recognition based on Pitch Contours," JASA, pp.1687-1697, 1972.
6. J.Kraayeveld et al., "Speaker characterization in dutch using prosodic parameters," Proc. of EUROSPEECH'9, pp.427-430, 1991.
7. B.Yegnanarayana et al., "A speaker verification system using prosodic features," Proc. of ICSLP 94, pp.S31-9.1 - S31-9.4, 1994.
8. D.W.Griffin and J.S. Lim, "Multiband excitation vocoder," IEEE Tran. On Acoustics, Speech and Signal Processing, Vol.36 No.8, pp.1223-1235, August, 1988.
9. L.Philip, *COLEA: A MATLAB software tool for speech analysis*, http://giles,ualr.edu/asd/speech/