A STUDY ON THE RECOGNITION OF LOW BIT-RATE ENCODED SPEECH

An-Tzyh Yu and Hsiao-Chuan Wang
Department of Electrical Engineering, National Tsing Hua University,
Hsinchu, Taiwan 30043
E-mail:hcwang@ee.nthu.edu.tw

c

1. INTRODUCTION

Digital speech communications are the future trend in the Internet and mobile phones. The low bit-rate coding of speech signals is the essential requirement in the concern of channel bandwidth and transmission efficiency. The voice-based services will become more attractive to the service providers. Many voice-driven applications require that users must be authorized and able to be identified. Other applications may allow users to retrieve information or perform some kind of transactions by voice. Traditionally, speech recognition systems need to get the waveform of speech signal and convert it into some specific feature parameters. In order to retain enough speech information, the sampling rate as well as the bit count per sample is high. In other words, these speech recognition systems require high bit-rate speech data. In the contrast, speech coding is to compress the speech information at low bit-rate. To recognize the encoded speech should face the problem of limited and inaccurate feature parameters. One may decode the coded information to recover the speech signal. However, the recovered signal can not retain the same waveform as the original one. Further process of this recovered signal does not gain any extra information.

Many researches have been reported in dealing with the problem of the recognition of encoded speech. Lilly et al.[1] reported the effect of speech coders on speech recognition performance. They examined six speech coders with bit-rate from 40 kbps down to 4.8 kbps and found the recognition performances deteriorate with the reduction of bit-rate. Kuitert et al.[2] investigated the impact of GSM coding on the performance of speaker verification system and found almost no influence on the recognition performance. Sönmez et al.[3] proposed an environment adaptation technique based on adaptive VQ. They preserved the topology transformation for robust recognition and obtained substantial improvement. Sankar et al.[4] utilized a RBF neural network to build robust speech recognizer for mobile applications and achieved satisfactory results. Although the mentioned researches achieved outstanding results, they were all based on decoded speech. In this paper, the encoded parameters of speech features in various speech coding systems are used for a isolated digit speech recognition task and their performances are evaluated. The effect of channel distortion and noise influence is also investigated. The HMM with mixture of continuous Gaussian densities is the framework of the speech recognition system in this study.

Experiments of the speech recognition were conducted based on the encoded speech parameters generated by LPC-10e, FS1016 and GSM coding mechanism. It reveals that the recognition system using features derived from the encoded parameters can achieve satisfactory performance under the noise-free condition. The performance becomes bad under a mismatched noisy environment. In general, the LSP achieves the highest recognition accuracy under clean environment, but degrades substantially with the lowering of SNR. However MFCC is more robust to noise under almost all noisy environments. Experimental results also show that GSM system seems to be more robust to noise than LPC-10e and FS1016 do.

The organization of this paper is as follows: Section 2 describes the speech database. Section 3 presents feature extraction methods. Section 4 interprets the recognition system used herein. Section 5 illustrates experimental results. And conclusion is made finally.

2. SPEECH DATABASE

The speech data provided by 50 male and 50 female speakers was collected in a sound treated environment and sampled at 8 kHz. There are three sessions of data collection, referred to as the clean speech. A speaker utters a set of ten Mandarin digits in each session. Two sessions are used for training and the other session is for testing. End points are roughly detected so that each utterance still contains short periods of pre-silence and post-silence. The specified amount of noise is added to the clean speech with specific SNR values when generating artificially noisy speech. Noises used herein are white noise, F16 noise, factory noise. White noise is artificially generated by computer and the other noises are obtained from NOISEX-92 database.

3. FEATURE EXTRACTION

This study attempts to find the fact that how good is for using the encoded parameters in speech recognition without recovering the speech signal. Three low bit-rate speech coding mechanisms are examined. Those are LPC-10e, FS1016 (CELP), and GSM. For each set of speech feature parameters decoded from a coding

mechanism, we derive other feature parameters and perform the speech recognition for comparison. The initial feature parameters for the three low bit-rate speech coding mechanisms are 10 reflection coefficients (RC) for LPC-10e, 10 line spectrum frequencies (LSP) for FS1016, and 8 log area ratios (LAR) for GSM. From the initial feature parameters, other feature parameters are derived. These include linear prediction coefficients (LPC), cepstral coefficients (CEP), and mel-cepstral coefficients (MFCC).

4. RECOGNITION SYSTEMS

Mandarin digit recognition is a task for evaluating the derived speech features under different environmental conditions. Each digit is modeled by an HMM with mixture of continuous Gaussian densities. The HMM contains four to six states depending on its average duration and variation, and begins with an pre-silence state and ends with a post-silence state. The silence states for all digits are tied together, i.e., share the same statistic parameters. Each state is modeled by a mixture of 4 Gaussian densities. All covariance matrices are diagonal. The model parameters are trained by the segmental K-means algorithm using coded clean speech .
.

5. EXPERIMENTAL RESULTS

Experiments of the speech recognition were conducted based on the encoded speech parameters generated by LPC-10e, FS1016 and GSM coding mechanism.

5.1 Experiments for LPC-10e coding mechanism

For the features derived from the LPC-10e encoded parameters, i.e. 10 reflection coefficients (RC), the dimension is twenty, consisting ten derived coefficients and its first derivate. Figure 1 illustrates the performance under clean environment. It shows that CEP, MFCC and LSP achieve higher recognition accuracy than RC and LAR. Among them, LSP achieves the highest recognition accuracy.

The following experiments investigate the recognition performance under several noisy environments. Tables 1~3 illustrate that there are obvious declines in recognition performance under three noisy environments. These declines are worse under white noise than under f16 noise or factory noise. Tables 1~2 reveal that MFCC is more robust than CEP and LSP under white noise and F16 noise. Table 3 also shows that LSP is more robust than CEP and MFCC under factory noise.

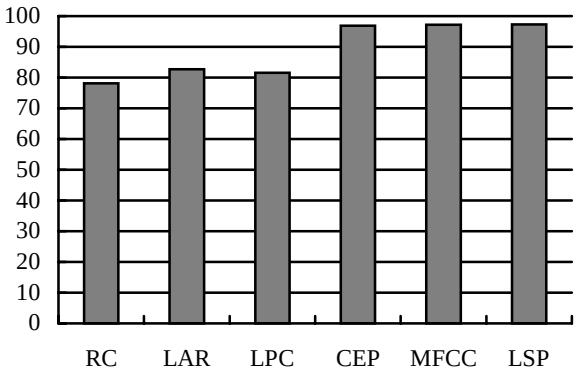

Figure 1: Recognition rates for features derived from LPC-10e parameters under clean environment

	SNR					
	clean	20 dB	15 dB	10 dB	5 dB	0 dB
LSP	97.3	65.8	42.5	29.8	18.0	11.8
CEP	96.9	74.4	56.3	37.5	22.9	16.8
MFCC	97.2	76.00	58.7	35.4	19.6	17.1

Table 1: Recognition rates for features derived from LPC-10e parameters under white noise

	SNR					
	clean	20 dB	15 dB	10 dB	5 dB	0 dB
LSP	97.3	86.8	74.9	59.1	39.4	21.2
CEP	96.9	85.4	72.4	56.7	37.3	22.9
MFCC	97.2	90.9	79.1	61.7	36.7	21.5

Table 2: Recognition rates for features derived from LPC-10e parameters under F16 noise

	SNR					
	clean	20 dB	15 dB	10 dB	5 dB	0 dB
LSP	97.3	85.8	75.5	56.5	36.8	23.2
CEP	96.9	83.6	73.0	54.9	33.9	20.7
MFCC	97.2	85.6	72.9	51.3	27.8	16.4

Table 3: Recognition rates for features derived from LPC-10e parameters under factory noise

5.2 Experiments for FS1016 coding mechanism

Several experiments were conducted to examine the recognition accuracy of various features derived from the FS1016 encoded parameters, i.e. 10 Line spectrum frequencies (LSP). Each feature used here comprises of the original derived feature and its delta feature. The dimension of each feature is twenty. Figure 2 shows the performance in clean speech condition. We can find that CEP, MFCC and LSP achieve higher accuracy than LPC. Among them, LSP is the highest.

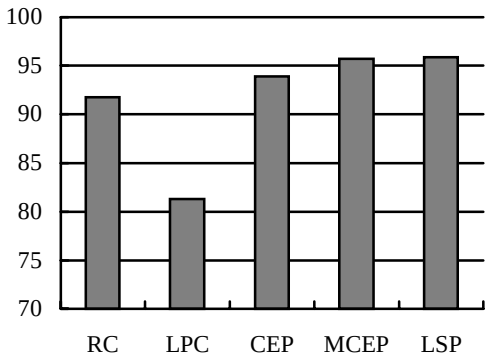

Figure 2: Recognition rates for features derived from FS1016 parameters under clean environment

For noisy environments, the recognition accuracy is shown in Tables 4~6. It is clear that the performance degrades when the environmental noise is high. These tables reveal that MFCC is more robust than other features. ance degradation is critical under white noise than under F16 noise or factory noise.

	SNR					
	clean	20 dB	15 dB	10 dB	5 dB	0 dB
RC	91.8	45.9	36.5	27.3	23.8	15.6
CEP	93.9	57.5	48.3	36.6	28.1	21.7
MFCC	95.7	79.0	71.4	54.7	37.0	27.3
LSP	95.9	56.9	43.3	34.6	26.3	22.2

Table 4: Recognition rates for features derived from FS1016 parameters under white noise

	SNR					
	clean	20 dB	15 dB	10 dB	5 dB	0 dB
RC	91.8	71.4	66.4	55.1	42.8	31.5
CEP	93.9	73.1	66.8	58.0	49.8	40.7
MFCC	95.7	94.2	91.9	83.6	72.3	57.2
LSP	95.9	71.9	67.2	60.0	49.4	38.3

Table 5: Recognition rates for features derived from FS1016 parameters under F16 noise

	SNR					
	clean	20 dB	15 dB	10 dB	5 dB	0 dB
RC	91.8	73.0	70.8	64.6	55.0	45.0
CEP	93.9	76.3	74.3	69.4	60.2	49.5
MFCC	95.7	95.4	94.1	92.0	83.4	72.3
LSP	95.9	76.1	72.3	67.1	60.3	48.2

Table 6: Recognition rates for features derived from FS1016 parameters under factory noise

5.3 Experiments for GSM coding mechanism

For the features derived from GSM encoded parameters, i.e. 8 log area ratios (LAR), each feature consists of the original derived feature and its delta feature, and the dimension of each feature is sixteen. The performance under clean environment is shown in Figure 3. It indicates that all features except LPC can achieve high performance. LSP is the best one.

A series of experiments were continually conducted to examine the recognition accuracy under noisy environments. The performance of the recognition system under noisy environments is shown in Tables 7~9. The performance of the recognition system degrades rapidly under the white noise. In other hand, the degradation is not so serious for F16 noise and factory noise. Tables 8 and 9 also reveal that CEP and MFCC can achieve robust performance under F16 noise and factory noise.

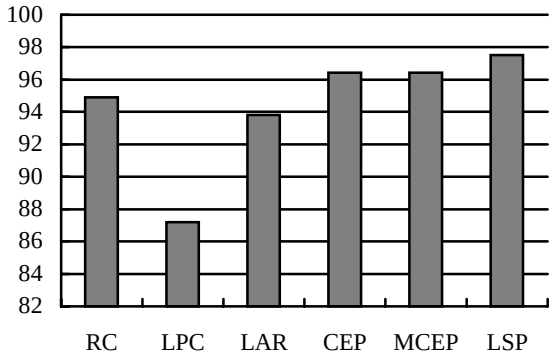

Figure 3: Recognition rates for features derived from GSM parameters under clean environment

	SNR					
	clean	20 dB	15 dB	10 dB	5 dB	0 dB
LAR	94.9	71.2	53.9	36.2	17.2	10
CEP	96.4	78.7	65.2	48.1	31.7	11.8
MFCC	96.1	80.9	61.5	42.4	28.1	12.1
LSP	97.5	74.1	53.2	40.1	26.8	18.2

Table 7: Recognition rates for features derived from GSM parameters under white noise

	SNR					
	clean	20 dB	15 dB	10 dB	5 dB	0 dB
LAR	94.9	94.5	92.5	83.4	63.6	46.7
CEP	96.4	95.5	93.0	85.0	68.0	51.8
MFCC	96.1	95.6	95.1	89.3	70.7	53.7
LSP	97.5	97.4	94.9	85.8	62.5	43.0

Table 8: Recognition rates for features derived from GSM parameters under F16 noise

	SNR					
	clean	20 dB	15 dB	10 dB	5 dB	0 dB
LAR	94.9	94.9	93.6	90.3	80.7	63.6
CEP	96.4	95.6	95.4	93.9	87.5	73.6
MFCC	96.1	96.0	95.9	94.0	87.7	72.2
LSP	97.5	97.5	97.1	93.8	82.7	61.6

Table 9: Recognition rates for features derived from GSM parameters under factory noise

6. CONCLUSION

Experiments of the speech recognition based on encoded speech parameters generated by LPC-10e, FS1016 and GSM coding mechanism were conducted in this study. Experimental results reveal that the recognition system using features derived from the encoded parameters can achieve satisfactory performance under a noise-free condition. The performance degrades under a mismatch noisy environment. The white noise seems to be more harmful to the recognition performance than other types of noises.

The performance of the recognition system degrades rapidly when the SNR is lower. LSP achieves the highest recognition accuracy under clean environment, but degrades substantially with the lower SNR. The feature of MFCC is more robust to noise. As far as the encoding mechanism is concerned, GSM is more robust to noise than LPC-10e and FS1016.

REFERENCE

[1] B.T. Lilly and K.K. Paliwal, “Effect of Speech Coders on Speech Recognition Performance”,ICSLP-96, pp. 2344-2347

[2] M. Kuitert and L.Boves, “Speaker Verification with GSM Coded Telephone Speech”, Eurospeech-97, pp. 975-978

[3] M. K. Sönmez , R. Rajasekaran and J. S. Baras, “Robust Recognition of Cellular Telephone Speech by Adaptive Vector Quantization”, ICASSP-96, pp. 503-506

[4] R. Sankar and N. S. Sethi, “Robust Speech Recognition Techniques Using a Radial Basis Function Neural Network for Mobile Applications”, ICASSP-97, pp. 87-91