ISCA Archive
http://www.isca-speech.org/archive

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

ROBUST FEATURE EXTRACTION IN A VARIETY OF INPUT DEVICES ON THE BASIS OF ETSI STANDARD DSR FRONT-END

Satoru TSUGE *Shingo KUROIWA* *Masami SHISHIBORI* *Fuji REN* *Kenji KITA*

Department of Information Science & Intelligent Systems
Faculty of Engineering, Tokushima University
E-mail: {tsuge, kuroiwa, bori, ren, kita}@is.tokushima-u.ac.jp

ABSTRACT

This paper reports an evaluation of European Telecommunications Standards Institute (ETSI) standard Distributed Speech Recognition (DSR) front-end through continuous word recognition on a Japanese speech corpus and proposes a method, the Bias Removal Method (BRM), that reduces the distortion between feature vector and VQ codebook. Experimental results show that using non-quantized features in acoustic model training procedure can improve the recognition performance of DSR fornt-end features and that the proposed method can improve recognition performances of DSR front-end feature.

1. INTRODUCTION

As potable terminals, such as cellular phones and PDAs (Personal Digital Assistants), lately become very popular, the necessity to use wireless and mobile gears is increasing. These portable terminals are typically small in size and are difficult to manipulate. In this respect, speech is a very convenient and reasonable interface. Hence, the speech recognition with portable terminals is demanded. However, all the speech recognition processes on a large or a middle vocabulary task have not performed in the portable terminal. To solve this problem, it is consider that the speech recognition system is located on server side. This is accomplished by transmitting the coded parameter to the server, which reconstructs the speech signal, parameterizes it and performs recognition. However, it is well known that lower recognition performances than uncoded speech[1]. To avoid these problems, several researchers have proposed the feature extraction method that the recognition features are computed directory from the transmitted information that is codec parameters[2][3][4][5].

On the other hand, a Distributed Speech Recognition (DSR) system is proposed to overcome these problems[6]. It separates the structural and computational components of recognition into two parts – the front-end processing on the terminal and the speech recognition engine on server. This separation of tasks can be a cause of a flexible architecture with great potential. That brings some advantages for using the DSR method as follows:

- It is possible to avoid the influence of the channel distortion, because the front-end part sends the back-end not to the speech signal but to the feature parameters. Therefore, we can get the improvement of the recognition performance.
- The bit rate is low because the bitstream, which the front-end part sends to the back-end part, only includes the information available for the speech recognition.
- Because there is no restriction of frequency band, it is possible to use the information of low and high frequency.

The DSR systems need a common bitstream format between front-end and back-end. To address this need, the European Telecommunications Standards Institute (ETSI) is producing the published standard DSR front-end algorithm based on Mel-Cepstrum technology[7]. It is reported that there are several researches that use this front-end[6][8][9].

In this paper, we describe following evaluation and a proposed method.

1. We evaluate ETSI standard DSR front-end through continuous word recognition on a Japanese speech corpus.
2. The recognition performance may be degraded because of the differences of the channel characteristics such as an A/D in DSR front-end and a microphone. To solve this problem, we perform Cepstral Mean Subtraction (CMS) on the back-end in our experiments. Because a vector quantization (VQ) codebook is fixed in ETSI standard DSR front-end, the increased distortion of the codebook and the feature vector causes the degradation of recognition performance if the difference of this channel characteristic grows. Therefore, we propose a method that reduces VQ distortion.

2. DISTRIBUTED SPEECH RECOGNITION

2.1. Speech Recognition style

Figure 1 shows a block diagram of the DSR system. The DSR system is constructed in two parts, the front-end and the back-end.

At the front-end part, the speech signal is sampled and parameterized. These are then compressed to obtain a lower bit rate (4.8kbps) for transmission. The compressed parameters are formatted into a defined bitstream for transmission. The defined bitstream is sent to the back-end in the remote server where parameters received with transmission errors are detected and the bitstream is decompressed to reconstitute the features. DSR system needs a common front-end between front-end and back-end. To address this need, the ETSI is producing the published standard DSR front-end algorithm based on Mel-Cepstrum technology[7].

2.2. Feature Extraction Algorithm

In this section, we brief the feature extraction algorithm of the ETSI standard DSR front-end. For full details, see the ETSI standard DSR standard document[7]. Figure 2 shows a block diagram of a DSR front-end.

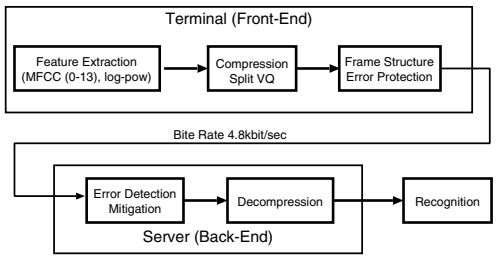

Fig. 1. DSR System

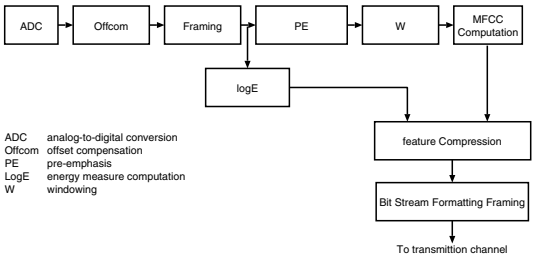

Fig. 2. ETSI standard DSR front-end

The sampled speech signal is parameterized by using a Mel-Cepstrum algorithm. This process generates 12 Mel-Frequency Cepstral Coefficients (MFCCs), C0 and a log energy parameter. A feature extraction conditions are described as follows:

- sampling rate: 8, 11, 16 kHz
- window: Hamming
- frame length: 25msec (8 and 16kHz sampling), 23.27msec (11kHz sampling)
- frame shift: 10msec
- pre-emphasis coefficient: 0.97
- filter bank order: 23
- offset compensation filter

$$s_{of}(n) = s_{in}(n) - s_{in}(n-1) + 0.999 \times s_{of}(n-1), \quad (1)$$

where, $s_{in}(n)$ and $s_{of}(n)$ indicate input speech signal and output speech signal, respectively.

2.3. Feature Compression Algorithm

The feature parameters, which are described in the previous section, are compressed to obtain a lower bit rate for transmission. A split vector quantization (VQ) algorithm is used for this compression algorithm. A codebook of size 64 is used for each pair of cepstral coefficients from MFCC[1] to MFCC[12]. Also, a size of code book which is used for C0 and a log energy is 256. The CRC, header information and error protection code have been applied to the compressed data.

3. VQ DISTORTION REDUCTION METHOD

When the MFCCs change under the influence of channel characteristic, the distortion between MFCCs that are extracted from ETSI standard DSR front-end and VQ codebook increases. As a result, there is degradation in the recognition performance. Hence, we propose VQ distortion reduction methods, which are described in following two sections.

3.1. Bias Removal Method 1

In this section, we propose Bias Removal Method 1 (BRM1) for reducing the distortion between VQ codebook and feature vector. The following steps perform this proposed method.

1. Calculate an average feature vector of each test sentence.

$$\boldsymbol{a}_{test} = \frac{\sum_{n=1}^{N} \boldsymbol{x}_n}{N}, \quad (2)$$

where, $\boldsymbol{a}_{test}$ and $\boldsymbol{x}_n$ indicate the average feature vector of each test sentence and the feature vector of each frame, respectively. N is the number of frames in a test sentence.

2. Subtract a difference between average feature vector of training sentences and average feature vector of a test sentence.

$$\tilde{\boldsymbol{x}}_n = \boldsymbol{x}_n - (\boldsymbol{a}_{test} - \boldsymbol{a}_{train}), \quad (3)$$

where, $\boldsymbol{a}_{train}$ indicates the average feature vector of VQ codebook creating data. $\tilde{\boldsymbol{x}}_n$ is the modified feature. Because it is actually difficult to require these training data, we use the average values of VQ codebook for $\boldsymbol{a}_{train}$.

3.2. Bias Removal Method 2

In this section, we propose the Bias Removal Method 2 (BRM2). The feature vector approximates the VQ codebook through the repetition of this method. This method is based on Generalized Lloyd Algorithm (GLA).

Given the test data x_n^0, ($n = 1, \ldots, N$, N is the number of frames in a test sentence), centroid decision function $q(v)$ and the proposed method iteratively performs the following steps.

1. The distortion between a test datum and VQ codebook is defined as

$$d_n = (x_n^i - q(x_n^i))^2, \quad (4)$$

where, i indicates the iteration number.

2. The distortion of a test sentence is calculated as

$$D = \sum_n d_n \quad (5)$$

$$= \sum_n (x_n^i - q(x_n^i))^2. \quad (6)$$

3. We estimate the bias, h, which minimizes a distortion, $\tilde{D}$.

$$\tilde{D} = \sum_n ((x_n^i - h) - q(x_n^i)) \quad (7)$$

$$\frac{\partial \tilde{D}}{\partial h} = \frac{\partial (\sum_n ((x_n^i - h) - q(x_n^i))^2)}{\partial h} \quad (8)$$

$$h = \frac{\sum_n x_n^i - q(x_n^i)}{N}. \quad (9)$$

4. The modified test data which are used in the next iteration, x_n^{i+1}, are calculated as

$$x_n^{i+1} = x_n^i - h. \tag{10}$$

5. Repeat 1, if the distortion less than threshold.

In this way, the feature vectors are changed to fit VQ codebook.

4. EXPERIMENT

The DSR front-end and the proposed method are evaluated through a Japanese standard dictation task (JNAS newspaper corpus of ASJ[10]).

4.1. Conditions

A total of 5,168 sentences by 103 male speakers are used for the training. For the open test set, 100 sentences by 23 male speakers are used.

The feature vector for the experiment is 25 MFCC's (12 static MFCCs extracted from the ETSI standard DSR front-end + 12 delta + logpower-delta). This feature vector is quantized and we call it *MFCCs with VQ* in this paper. In order to examine the effects of vector quantization, we use the feature vector without vector quantization, namely *MFCC w/o VQ*. For comparison with traditional codec method, we use the feature vector that is extracted from coded and decoded speech. We use G723.1 (5.3kbps)[11] for the codec method and call this feature vector *G723.1 MFCCs w/o VQ*. All the feature vectors are performed CMS on the back-end.

For the acoustic model, shared state triphone HMMs with sixteen Gaussian mixture components per state are trained. We set the number of states at about 1,000.

We use following filters for evaluating the proposed method under the condition of distortion speech.

- high pass filter (H/P)

$$s_{of}(n) = s_{in}(n) - 0.9 \times s_{in}(n-1) \tag{11}$$

- moving average filter (M/A)

$$\begin{aligned} s_{of}(n) &= 0.25 \times (s_{in}(n) + s_{in}(n+1) \\ &\quad + s_{in}(n+2) + s_{in}(n+3)) \end{aligned} \tag{12}$$

Where, $s_{in}(n)$ and $s_{of}(n)$ indicate input speech signal and output speech signal, respectively.

The index of the recognition performance is Word Error Rate (WER) given by

$$\text{WER} = \frac{I + S + D}{N} \cdot 100(\%), \tag{13}$$

where, I, D and S are insertion errors, deletion errors and substitution errors, respectively. We compute the WER under the optimum condition of 1st and 2nd pass width. We use Julius for the speech recognizer[12].

Table 1. Recognition performnaces of ETSI standard DSR front-end (WER (in %))

feature vector (MFCCs)		sampling rate	
training	testing	8kHz	16kHz
(1) *w/o VQ*	*w/o VQ*	10.4	9.6
(2) *with VQ*	*with VQ*	15.3	10.7
(3) *G723.1 w/o VQ*	*G723.1 w/o VQ*	18.6	–
(4) *w/o VQ*	*with VQ*	10.8	9.6

4.2. Experimental Results and Discussion

4.2.1. Experimental result of DSR front-end

We perform speech recognition experiments under the condition that the *MFCCs with VQ* are used for acoustic model training and testing. The experimental results are shown in table 1. For comparison, this table also includes the result of using *MFCCs w/o VQ* both in acoustic model training and in recognition and of using *G723.1 MFCCs w/o VQ* in them.

Comparing (2) with (3) in this table, we can observe that the recognition performances of the ETSI standard DSR front-end, *MFCCs with VQ*, is better than traditional codec, *G723.1 MFCCs w/o VQ*. In the G723.1 codec, the greatest part of transmitted information is assigned to personal information of speech, e.g., pitch. Hence, the transmitted information may not included the important information for speech recognition. On the other hand, in DSR, compressed feature parameters are transported to back-end. Therefore, the lack of information available for speech recognition is less frequent than it is when G723.1 codec method is used. We conclude that the difference of transmitted information between DSR and traditional codec influences the recognition performance.

Comparing (1) and (2) in this table, we can see that the VQ causes the degradation of recognition performances. In fact, we conjecture that the dispersion of the feature parameter by the vector quantization negatively affects the acoustic model training. In order to examine the effects of vector quantization in acoustic model training, we carry an experiment in which we perform acoustic model training with *MFCCs w/o VQ* and recognition of *MFCCs with VQ*. One can tell from table 1 that using non-quantized feature vectors in acoustic model training can improve the recognition performance of *MFCCs with VQ*. This result is almost the same as when *MFCCs w/o VQ* is used both in acoustic model training and recognition. Therefore, it can safely be said that the dispersion of feature parameters in acoustic model training has bad effects on recognition performance. To settle this problem, as the experiment shows, the acoustic model training with non-quantized features or the dispersion of HMM would be solutions.

Moreover, according to those experiments, when the recognition of quantized MFCCs with non-quantized acoustic models is performed, the recognition performance of a sampling frequency of 16kHz is a bit higher than in the case of 8kHz. It is interesting because this shows a possibility that when the bit-rate is the same, broadening the analysis band can improve the accuracy of recognition.

Table 2. The result of the recognition with proposed methods in 16kHz sampling (WER in %)

adaptation method	filter		
	no filter	H/P	M/A
(1) *no adapt. w/o VQ*	9.63	9.63	9.32
(2) *no adapt. with VQ*	9.58	9.38	11.22
(3) *BRM1*	9.46	9.19	10.65
(4) *BRM2*	9.07	9.51	9.07

Table 3. The result of recognition with proposed methods in 8kHz sampling (WER in %)

adaptation method	filter		
	no filter	H/P	M/A
(1) *no adapt. w/o VQ*	10.40	10.59	11.67
(2) *no adapt. with VQ*	10.78	10.59	14.14
(3) *BRM1*	10.21	10.46	11.67
(4) *BRM2*	10.59	10.46	11.60

4.2.2. Efficiency of the proposed method

In this section, we describe the efficiency of the proposed method. In previous section, we described that the acoustic model trained with non-quantized feature vector could improve recognition performance of quantized feature vector in comparison with one trained with quantized features. Therefore, in this section, we use the acoustic model trained with non-quantized feature vector, *MFCCs w/o VQ*. Table 2 and 3 compare speech recognition performances obtained by using various feature parameters under the condition of sampling frequencies 16kHz and 8kHz, respectively.

From these tables, we can see that the proposed methods, *BRM1* and *BRM2*, do not degrade the recognition performances under the clean condition. The recognition performances of *no adapt. w/o VQ* are almost the same as those in a clean condition, because we perform CMS on the back-end in the experiments. However, because of the influence of the increased distortion caused by M/A filter, the recognition performance of *no adapt. with VQ* is significantly degraded, especially when the sampling frequency is 8kHz. On the other hand, when the H/P filter is used, the recognition performances of *no adapt. with VQ* are almost the same as they are in clean conditions. For elucidating that reason, we investigated the distortion between feature vector and VQ codebook. After the investigation, we discovered that the distortion of H/P filter did not increase so much as it did in clean conditions. Therefore, it can be said that using a H/P filter does not degrade recognition performances.

The proposed method, *BRM1*, helps improve the recognition performances of *no adapt. with VQ* in all the conditions. When we investigate the distortion between feature vector and VQ codebook, it is understood that the proposal method, *BRM1*, decreases the distortion comparison with *no adapt. with VQ*. The *BRM2* improves the recognition performances as well as the *BRM1* dose in almost all the cases so far. Especially, when a M/A filter that causes large distortion is used, *BRM2* shows the best performances. In those experiments, we conclude that the these proposed methods are effective to improve the degradation of the recognition performances cased by the distortion.

5. CONCLUSIONS

In this paper, we evaluated ETSI standard DSR front-end through continuous word recognition on a Japanese speech corpus and proposed the methods, the Bias Removal Methods (BRMs), that reduced the distortion between feature vector and VQ codebook.

We could confirm that using non-quantized feature vectors in acoustic model training could improve the recognition performance of quantized feature vector that was extracted from ETSI standard DSR front-end.

The proposed methods could decrease the distortion between feature vector and VQ codebook. Experimental results showed that the proposed methods could improve recognition performances of quantized feature vector.

In our experiments, we applied the proposed method only to the test sentences. Because it is consider that there is distortion of the training data, we plan to apply the proposed methods to the training data.

6. REFERENCES

[1] B. Lilly and K. Paliwal, "Effect of speech coders on speech recognition performance," *Proc. ICSLP*, pp. 2344–2347, 1996.

[2] B. Raj, J. Migdal, and R. Singh, "Distributed speech recognition with codec parameters," *ASRU*, 2001.

[3] J. Huerta and R. Stern, "Speech recognition from GSM codec parameters," *Proc. ICSLP*, pp. 1463–1466, 1998.

[4] H. Kim and S. Member, "A bitstream-based front-end for wireless speech recognition on IS-136 communications system," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 558–568, 2001.

[5] T. Uchibe, S. Kuroiwa, and N. Higuchi, "The method to translate codes of Cs-Acelp into acoustic parameters for speech recognition," *Proceedings of the 2000 IEICE General Conference*, vol. 6, pp. 195, 2000, (in Japanese).

[6] D. Pearce, "Enabling new speech driven services for mobile devices: An overview of the etsi standards activities for distributed speech recognition front-ends," *AVIOS*, 2000.

[7] "ETSI ES 201 108 v1.1.2 distributed speech recognition; front-end feature extraction algorithm; compression algorithm," 2000.

[8] C. Broun, W. Campbell, D. Pearce, and H. Kelleher, "Distributed speaker recognition using the ETSI distributed speech recognition standard," *The 2001 International Conference on Artificial Intelligence (IC-AI'2001)*, 2001.

[9] S. Tsuge, S. Kuroiwa, F. Ren, and K. Kita, "A speech recognition using etsi standard dsr front-end," *ASJ*, pp. 171–172, 2002, (in Japanese).

[10] K. Ito, M. Yamamoto, and et al., "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition," *J. Acoust. Soc. Jap.*, vol. 20, no. 3, pp. 199–206, 1995.

[11] "ITU-T recommendation G.723.1 dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s," 1996.

[12] A. Lee, T. Kawahara, and K.Shikano, "Julius — an open source real-time large vocabulary recognition engine," *Proc. EuroSpeech*, pp. 1691 – 1694, 2001.