BLIND EQUALIZATION TECHNIQUES FOR ETSI STANDARD DSR FRONT-END

Shingo KUROIWA *Satoru TSUGE*

Department of Information Science & Intelligent Systems
Faculty of Engineering, University of Tokushima

ABSTRACT

In this study we present blind equalization techniques for ETSI standard Distributed Speech Recognition (DSR) front-end which compensate for acoustic mismatch caused by input devices. The DSR front-end employs vector quantization (VQ) for feature parameter compression so that the mismatch does not only cause a shift of parameters but also increases VQ distortion. Although CMS is one of the most effective methods to compensate for the shift, it can not decrease VQ distortion in DSR. To compensate for the shift and decrease VQ distortion simultaneously, the proposed methods estimate the shift in the input data necessary to match the VQ codebook distribution. The methods do not need the acoustic likelihood which is calculated in a decoder on the server side. Therefore, they are applicable to the DSR front-end. Japanese Newspaper Article Sentences database (JNAS) was used for the equalization experiments. While the word error rate (WER) for ETSI standard DSR front-end was 18.6 % under acoustic mismatched condition, our propsed method yielded a rate of 12.3 %.

1. INTRODUCTION

Portable terminals, such as cellular phones and PDAs (Personal Digital Assistants), lately have become very popular. These portable terminals are typically small in size and it is inconvenient and inefficient to use their conventional input devices provided to feed in complex command sequences. Speech is a more convenient and reasonable interface. Hence, portable terminals speech recognition demand. However, due to hardware limitations, all speech recognition processes on a large or middle scale vocabulary task can not be performed in the portable terminal.

One solution to this problem is to move the speech recognition system to the server side. In this method, the system has to recognize coded speech, such as VSELP, PSI-CELP, and ACELP. It is well known, however, that these systems achieve lower recognition performance than uncoded speech because of influences of the codec and channel distortion[1]. To avoid the influence of the speech decoding process, several researchers have proposed feature extraction methods where recognition features are computed directly from the transmitted information, i.e. codec parameters[2, 3, 4, 5].

Distributed Speech Recognition (DSR) has been proposed to overcome these problems of codec speech[6]. DSR separates the structural and computational components of recognition into two parts – front-end processing on the terminal and speech recognition engine on the server. This separation of tasks permits a flexible architecture with great potential. DSR has the following advantages:

- It is possible to avoid the influence of channel distortion because the front-end part sends the back-end not to the speech signal but to the feature parameters. Therefore, one can get improvement in 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 necessary for speech recognition.
- Because there is no restriction on the frequency band, it is possible to use information of low and high frequencies.

To enable widespread applications of DSR in the market place, a front-end standard is needed to ensure compatibility between the terminal and the remote recognizer. The European Telecommunications Standards Institute (ETSI) is producing a published standard DSR front-end algorithm based on Mel-Cepstrum technology[7].

In this paper, we consider the influence on recognition performance of DSR with acoustic mismatches caused by input devices. Cepstral Mean Subtraction (CMS) is one of the most effective methods to compensate for these mismatches. However, DSR employs a vector quantization (VQ) algorithm for feature compression so that the VQ distortion is increased by such mismatches. Large VQ distortion increases the speech recognition error rate. The recommendation of DSR front-end standardizes a VQ codebook so that CMS can not be applied on the terminal side, and is applied on the server side. Therefore, CMS can not decrease VQ distortion. To overcome these problems, this paper proposes the blind equalization techniques (BEQ), which decrease both mismatch and distortion simultaneously. Although BEQ is similar to the Signal Bias Removal (SBR)

0-7803-7663-3/03/$17.00 ©2003 IEEE

Table 1. Influence of the frequency characteristic (WER)

	Filter	
Sampling rate	no	M/A
8 kHz	13.5 %	58.2 %
16 kHz	12.2 %	32.2 %

technique [10], rather than maximize acoustic likelihood it minimizes VQ distortion. Hence, BEQ does not need information from a decoder and can be adopted in the DSR front-end.

2. BLIND EQUALIZATION METHOD

It is known that log-spectral based feature vectors, e.g. MFCC, shift if the frequency characteristic of the input device is altered. With DSR, these mismatches also increase VQ distortions. As a result, the speech recognition error rate increases. In this section, we propose blind equalization techniques which decrease these distortions.

2.1. Preliminary Experiment

We investigated the influence on the recognition performance of the acoustic mismatch caused by the input device. The moving average (M/A) filtering, which is shown in equation (9), was performed to the JNAS speech corpus to simulate the mismatch of the frequency characteristic of input devices. Other experimental conditions were the same as in section 3.1.

The results are presented in Table 1. This table also contains the results of non-filtering. From this table, the M/A filtering increased the word error rate (WER) compared to non-filtering. To investigate this result in more detail, we viewed a scatter chart of the VQ codebook and feature parameters with M/A filtering (Fig.1). We noticed a difference between the distribution of the feature paramters and the VQ codebook. This difference increased VQ distortion. CMS on the server side could not decrease this distortion.

2.2. Blind Equalization Method 1

In this section, we propose Blind Equalization Method 1 (BEQ1) which equalizes the mean of the input data with the mean of VQ centroids. The followings show the steps of BEQ1.

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

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

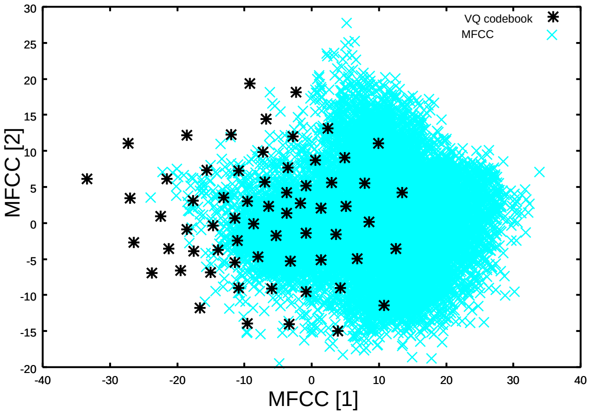

Fig. 1. The distortion between feature parameter and VQ codebook. (1st and 2nd order MFCC)

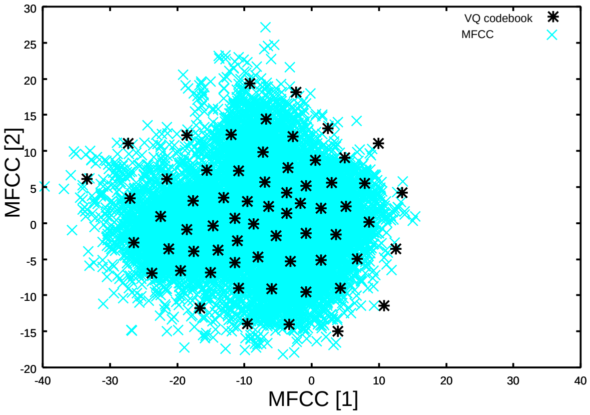

Fig. 2. Analysis example of BEQ1 (1st and 2nd order MFCC)

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

$$\tilde{\boldsymbol{x}}_n = \boldsymbol{x}_n - (\boldsymbol{a}_{test} - \boldsymbol{a}_{train}), \qquad (2)$$

where, $\boldsymbol{a}_{train}$ indicates the average feature vector of VQ codebook training data. $\tilde{\boldsymbol{x}}_n$ is the feature parameter which is applied BEQ1. Because it is actually difficult to require these training data, we use the average values of VQ codebook centroids for $\boldsymbol{a}_{train}$.

Fig.2 illustrates the effectiveness of the proposed method, BEQ1, under the acoustic mismatched condition. By applying the BEQ1, the distribution of the feature parameters approaches the VQ codebook. Consequently, the BEQ1 can decrease VQ distortion and may improve recognition performance.

2.3. Blind Equalization Method 2

In this section, we propose Blind Equalization Method 2 (BEQ2). The feature vectors approximate the VQ codebook through the repetition of this method. This method is based on the Generalized Lloyd Algorithm (GLA)[9].

The Signal Bias Removal (SBR) method has been proposed to compensate the convolution noise caused by differences in the input device's frequency characteristic[10]. The SBR calculates the bias which maximizes the acoustic likelihood. The proposed method, BEQ2, is similar to SBR. However, in BEQ2, the bias does not maximize the acoustic likelihood, but minimizes VQ distortion. Therefore, the BEQ2 can be adopted in the DSR front-end.

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

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

$$d_n = |\boldsymbol{x}_n^i - q(\boldsymbol{x}_n^i)|^2, \qquad (3)$$

where, i indicates the iteration number.

2. The distortion of a test sentence is calculated as

$$D = \sum_n d_n. \qquad (4)$$

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

$$\begin{aligned} \tilde{D} &= \sum_n |(\boldsymbol{x}_n^i - \boldsymbol{h}) - q(\boldsymbol{x}_n^i)|^2 & (5) \\ \frac{\partial \tilde{D}}{\partial \boldsymbol{h}} &= \frac{\partial (\sum_n |(\boldsymbol{x}_n^i - \boldsymbol{h}) - q(\boldsymbol{x}_n^i)|^2)}{\partial \boldsymbol{h}} = 0 & (6) \\ \boldsymbol{h} &= \frac{\sum_n \boldsymbol{x}_n^i - q(\boldsymbol{x}_n^i)}{N}. & (7) \end{aligned}$$

4. The modified test data which are used in the next iteration, $\boldsymbol{x}_n^{i+1}$, are calculated as

$$\boldsymbol{x}_n^{i+1} = \boldsymbol{x}_n^i - \boldsymbol{h}. \qquad (8)$$

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

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

When the BEQ2 is applied to the feature parameters in Fig.1, we can obtain almost the same result as in Fig.2 (BEQ1 result).

2.4. Technique to process in real-time

The proposed methods need the whole sentence to equalize input data. Thus, they can be applied only after the whole utterance ends and it is a disadvantage in real-time operation. This means that the methods are not applicable to DSR systems. To overcome this problem, we applied an idea, which has been used in some real application systems[11], and which had previously been developed by one of the authors. The idea is that the system makes the best use of the previous utterance. It calculates the shift of the previous utterance for use in the subsequent equalization. We evaluate this idea in the following experiments.

3. EVALUATION OF THE PROPOSED METHODS

3.1. Experimental Conditions

We evaluated the proposed methods through continuous speech recognition experiments. A total of 5,168 sentences by 103 male speakers were used for the training. For the open test set, 100 sentences by 23 male speakers were used.

The feature vector for the experiment was 25 MFCC's (12 static MFCCs extracted from the ETSI standard DSR front-end + 12 of their deltas + one delta-logpower).

For the acoustic model, shared-state triphone HMMs with sixteen Gaussian mixture components per state were trained. We set the number of states at about 1,000. In previous work[8], we described that the acoustic model trained with non-quantized feature vectors could improve recognition performance using DSR. Therefore, we used the acoustic model trained with non-quantized feature vector by the following experiments.

The following moving average filter (M/A) was used to simulate the mismatch of the frequency characteristic of input devices.

$$\begin{aligned} s_{of}(n) &= 0.25 \times (s_{in}(n) + s_{in}(n+1) \\ &\quad + s_{in}(n+2) + s_{in}(n+3)) \qquad (9) \end{aligned}$$

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

3.2. Experimental results

Tables 2 and 3 show the speech recognition performance obtained by using various equalization methods at sampling frequencies of 16 kHz and 8 kHz, respectively. The "Baseline" indicates ETSI ES 201 108 v.1.1.2 DSR front-end. The results of ETSI blind equalization ("ETSI"), which were described in the final draft of ETSI new robust DSR front-end in 2002[12], are also presented in Table 3 (8 kHz). In "non real-time" conditions, the equalizations were performed by using the whole utterance, while "real-time" methods used the previous utterance for equalization.

Table 2. Word error rates in 16 kHz sampling.

Equalization	Filter	
method	no	M/A
Baseline	12.2 %	32.2 %
non real-time		
CMS	9.6 %	11.2 %
BEQ1	10.4 %	10.2 %
BEQ2	10.0 %	11.2 %
real-time		
BEQ1	9.6 %	10.3 %
BEQ2	10.3 %	10.2 %

Table 3. Word error rates in 8 kHz sampling.

Equalization	Filter	
method	no	M/A
Baseline	13.5 %	58.2 %
non real-time		
CMS	10.8 %	14.1 %
BEQ1	10.8 %	12.3 %
BEQ2	10.9 %	12.4 %
real-time		
ETSI	13.3 %	18.6 %
BEQ1	12.8 %	13.7 %
BEQ2	11.6 %	14.0 %

These tables show that the proposed methods, BEQ1 and BEQ2, helped improve the recognition performances of the baseline under all conditions. Compared to the baseline under the M/A filter condition, BEQ1 (non real-time) achieved 78.9 % (at 8 kHz) and 68.3 % (at 16 kHz) improvement in the error rate. The BEQ2 (non real-time) yielded 78.7 % (at 8 kHz) and 65.2 % (at 16 kHz) improvement in the error rate. These results were better than the results of the baseline under the condition witout M/A filter. The proposal methods can compensate for acoustic mismatches caused by input devices.

Although CMS was applied, the M/A filtering increased the WER compared to the condition without the filter. CMS was applied on the server side, so that the CMS could not decrease VQ distortion. These results indicate the advantage of the proposed methods which decrease VQ distortion.

The proposed methods for "real-time" versions achieved similar performance with methods for "non real-time" versions in 16 kHz sampling. In 8 kHz sampling, "real-time" BEQ1 and BEQ2 slightly degrade recognition performance compared with "non real-time" values. However, in acoustic mismatched condition, the error rate significantly decreased from 18.6 % for ETSI blind equalization to 13.7 % for our blind equalization technique (BEQ1).

4. SUMMARY

In this paper, we proposed blind equalization techniques for ETSI DSR standard front-end, which decrease acoustic mismatches and VQ distortion simultaneously. Experimental results showed that the proposed techniques could improve the recognition performance under acoustic mismatched conditions. We also confirmed that the proposed methods could decrease VQ distortion. Although the proposed methods could not be performed in real-time as is, we also proposed an adoption to use them in real-time applications. Experimental results under real-time conditions revealed the advantages of the methods.

5. ACKNOWLEDGMENT

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 14350204, 2002 and International Communications Foundation (ICF).

6. REFERENCES

[1] B. Lilly and K. Paliwal, "Effect of speech coders on speech recognition performance," *Proc. ICSLP96*, pp. 2344–2347, 1996.

[2] B. Raj, J. Migdal, and R. Singh, "Distributed speech recognition with codec parameters," *Proc. ASRU2001*, 2001.

[3] J. Huerta and R. Stern, "Speech recognition from GSM codec parameters," *Proc. ICSLP98*, 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," *Proc. 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," *Proc. AVIOS2000*, 2000.

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

[8] S. Tsuge, S. Kuroiwa, M. Shishibori, F. Ren, and K. Kita, "Robust feature extraction in a variety of input devices on the basis of etsi standard dsr front-end," *Proc. ICSLP2002*, pp. 2221–2224, Sep 2002.

[9] *Learning from Data: Concepts, Theory, and Methods*, JOHN WILEY & SONS, INC., 1998.

[10] M. Rahim and B. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 190–202, 1996.

[11] T. Kato, S. Kuroiwa, and N. Higuchi, "Area code, country code, and time difference information system and its field trial", *Proc. IVTTA'98*, pp. 5 – 10, 1998.

[12] "ETSI ES 202 050 v1.1.1 stq; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," 2002.