Relative Mel-Frequency Cepstral Coefficients Compensation for Robust Telephone Speech Recognition

Jiqing Han[*,**], *Munsung Han*[*], *Gyu-Bong Park*[*], *Jeongue Park*[*], *Wen Gao*[**]

*. Language Understanding Lab. , Systems Engineering Research Institute, ETRI, Korea
**. Department of Computer Science and Engineering, Harbin Institute of Technology, P.R. China
email { jqhan, mshan, gbpark, jgpark}@seri.re.kr, wgao@jdl.mcel.mot.com

ABSTRACT

It is a crucial factor to find the robust and simple computation methods for the actual application of telephone speech recognition. In this paper, we propose a new channel compensation method, which uses a RASTA-like band-pass filter on the mel-frequency cepstral coefficients for robust telephone speech recognition. It is shown from the experiments that the proposed method, comparing with the RASTA processing, reduces the computational complexity without losing performance, and it is also better than CMS and two level CMS on the performance. We also verify that it is an effective approach to suppress very low modulation frequencies for robust telephone speech recognition.

1. INTRODUCTION

It is well known that the mismatches between the training and testing conditions can severely degrade the performance of automatic speech recognition systems. The telephone speech is a typical example of such mismatches, and it has been reported[9] that the error rate of a speech recognizer can increase from 1.3% to 44.6% when the testing data are filtered by a pole/zero filter modelling a long-distance telephone line and corrupted by noise at the SNR of 15dB. It is a crucial factor to find the robust and simple computation methods for the actual application of telephone speech recognition. The robustness of telephone speech recognition has been widely discussed, and a variety of approaches have been proposed[1-6]. Cepstral Mean Subtraction (CMS)[1][6] is one of the effective algorithms considering its simplicity. However the effectiveness of CMS is severely limited when the environment can not be adequately modelled by a linear channel. In order to process the nonlinear channel, the two level CMS[5] method was used. However, it needs signal classification, and the system performance depends on the classification accu-racy. The RelAtive SpecTrAl (RASTA) processing [2][3][4] that uses a band-pass filter with a very low cut-off frequency can suppress slowly-varying channel dis-tortions and get good performance. The conventional RASTA processing is applied on the Perceptual Linear Predictive (PLP)[7] log spectrum. However, PLP needs very complex computation.

In this paper, we use mel spectral analysis[8] instead of PLP approach to reduce the computation. Based on the linear relationship between mel-frequency log spectrum and Mel-Frequency Cepstral Coefficients (MFCCs), we extend RASTA processing from mel-frequency log spectrum to MFCCs, and a RASTA-like band-pass filter is proposed for robust telephone speech recognition. Next, we select the pole parameter of the filter by experiments and discuss the selection of the initial value. Comparing with RASTA processing, the proposed Relative MFCCs (RMFCC) processing not only reduces the computational complexity without losing performance, but also shows better performance than CMS and two level CMS techniques. Finally, we discuss the results and verify that it is an effective approach to suppress very low modulation frequencies for robust telephone speech recognition.

2. RELATIVE MEL-FREQUENCY CEPSTRAL COEFFICIENTS COMPENSATION

Perceptual experiments suggest that human speech perception might be able to suppress stationary non-linguistic background and enhance the variable linguistic message[10]. Thus, it is useful to adopt the features based on human hearing for robust speech recognition. In RASTA-PLP technique, only the RASTA processing is used to suppress slowly-varying channel distortions. Mel spectral analysis is also one way simulating the properties of human hearing and simpler than PLP analysis. Specifically, mel spectral analysis does not need the complex equal loudness pre-emphasis, intensity loudness power law and conducting spectral analysis again[7].

If we use $H(Z)$ to represent the RASTA band-pass filter, $Y_{t,i}$ and $\overline{Y}_{t,i}$ each represent the i-th mel-frequency log spectrums at frame t before and after being pro-cessed by RASTA, then we get

$$\overline{Y}_{t,i} = H(Z) \cdot Y_{t,i} \tag{1}$$

MFCCs, which are used as the features in most of the current speech recognizer, are calculated by using Discrete Cosine Transform (DCT) on mel-frequency log spectrum as follows[8]

$$\overline{C}_t(k) = \sum_{i=1}^{B} COS[k(i-0.5) \cdot \frac{\pi}{B}] \cdot \overline{Y}_{t,i}$$

$$= H(Z)\cdot\sum_{i=1}^{B} COS[k(i-0.5)\cdot\frac{\pi}{B}]\cdot Y_{t,i}$$
$$= H(Z)\cdot C_t(k)\,, \qquad (\text{k=1,2,3, ... , K}) \qquad (2)$$

where $\overline{C}_t(k)$ and $C_t(k)$ are the k-th MFCCs at frame t with and without using RASTA processing respectively, B is the number of mel-frequency bands, and K is the dimension of MFCCs.

From equation (2), it is reasonable to extend RASTA processing from log spectrum to MFCCs (i.e. first calculating MFCCs and then processing by a band-pass filter). Generally, B is bigger than K (e.g. we used B=40 and K=12), and thus this kind of RMFCC processing reduces the computation complexity.

The main part of RASTA processing is the IIR filter as

$$H(Z) = G\cdot\frac{Z^4(2+Z^{-1}-Z^{-3}-2Z^{-4})}{1-\rho Z^{-1}} \qquad (3)$$

We also use this kind of filter, and should select the parameters of the filter for our RMFCC processing.

When an input signal $X[t]$ passes through $H(Z)$ in equation (3), the output $Y[t]$ is

$$Y[t] = G\cdot\sum_{n=0}^{4}(n-2)X[t+n] + \rho\cdot Y[t-1] \qquad (4)$$

where t =0,1,2, ... , T-1 is the number of the frames, and the initial value $Y[-1]$ should be selected.

3. THE DATABASE AND THE BASELINE SYSTEM

The database is collected from the local telephone network in Seoul and Taejeon, and many kinds of different hand-sets are used for collection. Since the system is speaker independent, many speakers are selected for experiments. The training database contains utterances from 40 speakers (22 male and 18 female), and the testing utterances from 40 different speakers (22 male and 18 female). The male to female ratio in the database reflects that of the general South Korea population. Every speaker read 93 sentences several times, and then 84 Korean isolated words were manually segmented and labeled. Since some utterances were discarded due to bad recordings, the total number of utterances for training database is 11381, and for testing one 8036 .

We use the Signal-to-Noise Ratio (SNR) as an objective measurement to evaluate our database. In the literature, many SNR measurements have been proposed [e.g. 11]. Since we have no a priori knowledge about the telephone speech, the different SNR measurements were adopted for the training and testing database, and the results are listed in Table.1.

Table.1 *Different SNR measurements for the database*

Measurement	Training database	Testing database
SNR	14.07dB	13.95dB
SEGSNR	13.79dB	13.85dB
MAXSNR	19.78dB	19.00dB

It is shown from the Table.1 that the SNR measurements are very similar between the training database and the testing database, which might be that the database includes relatively sufficient environmental (speakers, channels, noises) features and obeys the statistical theory.

In the experiments, the speech signal is first digitized at a sampling rate of 8KHz, a pre-emphasis filter H(z) = 1 - $0.95\,Z^{-1}$ is applied to the speech samples, and a Hamming window of 240 samples (30ms) is used for every 15ms. Next, the power spectrum of the windowed signal in each frame is computed using a 256-point DFT, and 40 mel-frequency spectral coefficients are derived based on mel-frequency band-pass filters. Then, 12 MFCCs are computed using the DCT. Finally an isolated word, continuous-density HMM recognizer is used as the baseline system.

4. EXPERIMENT AND DISCUSSION

A series of experiments are designed to evaluate the proposed method for robust telephone speech recognition.

4.1. Parameter Selection

In equation (3), the pole parameter ρ should be determined, and a constant ρ=0.98 was used in the previous RASTA processing. For RMFCC, we select the parameter ρ by the comparing experiments. Using the gain G=0.1 consistent with RASTA, we compare the system performances for the different ρ, and the result is shown in Fig.1.

It is shown from the Fig.1 that ρ= 0.92 exhibits an optimum, which is adopted as the pole parameter in the following experiments.

In equation (4), there is also an initial value $Y[-1]$ to be selected to get a better recognition accuracy. In the previous RASTA works, they did not report how to select the initial value $Y[-1]$. All the methods are keeping the silent part before speech, and the results are dependent on the determination of the silence. In noisy environment, it is not easy to determine the silence parts and unfortunately, the silence is often mixed with serious noise. When the integrator is used from the silence, the noise might be introduced again. And moreover, it also needs extra silence processing. We attempt to find the special $Y[-1]$ to get good performance without processing the extra silence. Using the G=0.1 and ρ= 0.92, three kinds of the initial values: zero, cepstral mean and silence speech, are compared, and the results are listed in Table 2. We can see that using zero initial value gets the highest performance. It seems that the zero value normalizes the cepstral coe-fficients for all utterances.

Table.2 *Performance comparison for the different initial values*

Initial Value	Zero	Mean	Silence
Training Database	97.7%	97.6%	97.7%
Testing Database	92.9%	91.9%	92.6%

4.2. Comparing Experiments

CMS is a standard channel compensation techniques, which can remove the time-invariant parts of channel distortion, we implemented a system using CMS as noise compensation method. Although the linear time-invariant channel assumption is almost never satisfied in practice, we still achieved a significant improvement on the performance (97.3% for training database and 92.2% for testing one) comparing with the baseline system (93.5% and 88.2% for training and testing data-base respectively). In order to process the non-linear distortions, a two level CMS technique is implemented as follows,

Step 1: determining the maximum frame energy $E_{\max}$ for every utterance.
Step 2: separating the frames of current utterance into two classes,
if $E_t > \alpha \cdot E_{\max}$, then the frame t belongs to class I, else to class II (α =0.1 is a constant determined by experiment.).
Step 3: calculating the cepstral means for the class I and class II, respectively.
Step 4: subtracting the different cepstral means from the frames of the above two classes.

The experimental results using various types of the noise compensation methods are shown in Table.3. As discussed below, delta-MFCC has a relationship with RMFCC. In this point of view we list the performance when using delta-MFCC processing in Table.3. We also attempt to combine RMFCC with CMS and two level CMS respectively, but the results are not improved comparing with the case of using RMFCC only.
It is shown from the experiments that the performance of RMFCC is significantly superior to that of the base-line system, and a 39.8% reduction in word error rate is got for the testing database. Comparing with delta-MFCC, RMFCC produces a 28.3% reduction in word error rate with slight increase in computational com-plexity. RMFCC is also better than CMS and two level CMS on the performance, and can be implemented straightforwardly. On the other hand, both CMS and two level CMS need calculating the cepstral mean of the utterance and then the mean should be subtracted from every frame. Comparing with RASTA, RMFCC gets nearly same performance but requires a simple compu-tational complexity. With respect to both the perfor-mance and the computational complexity, RMFCC is the best one.

Table.3 *Word error rate using various types of noise compensation methods*

Method	Train Database	Test Database
Baseline	6.5%	11.8%
Delta-MFCC	3.4%	9.9%
CMS	2.7%	7.8%
Two level CMS	2.5%	7.2%
RASTA	2.1%	7.1%
RMFCC	2.3%	7.1%
CMS+RMFCC	2.3%	7.1%
Two level CMS +RMFCC	2.3%	7.1%

4.3. Discussion

RMFCC was shown to yield good performance in section 4.2. The solid line in Fig.2 is the frequency response curve of the RMFCC filter, which can attenuate very low modulation frequencies. Using delta-MFCC, the performance is also better than that of the baseline system which is using MFCC only, and this is consistent with [1]. We noticed that there is a relationship between delta-MFCC and RMFCC. When the denominator of the RMFCC filter in equation (3) is ignored, the RMFCC processing is equivalent to delta-MFCC. Therefore, delta-MFCC is also regarded as a kind of RMFCC processing. The frequency response curve of the filter used in delta-MFCC is the dotted line in Fig.2, and it can also suppress low modulation frequencies. This is the reason why using delta-MFCC is better than using MFCC only. Since delta-MFCC suppresses some parts of useful speech characters that might be included in the low frequencies, the performance is worse than RMFCC. Since CMS processing can be regarded as a kind of high-pass filtering, which can also suppress low modulation frequencies and meanwhile maintain speech characters, the performance is better than using delta-MFCC. The two level CMS, in that the different modulation frequencies are considered and removed by different high-pass filtering, is better than CMS. How-ever, since the low modulation frequencies which are suppressed by either CMS or two level CMS might be just one part which can be suppressed by RMFCC, the performance of RMFCC is better than that of both CMS and two level CMS, and the performances is not improved when CMS and two level CMS are combined with RMFCC, respectively. From the discussion, we know that it is an effective approach to suppress low modulation frequencies for robust telephone speech recognition.

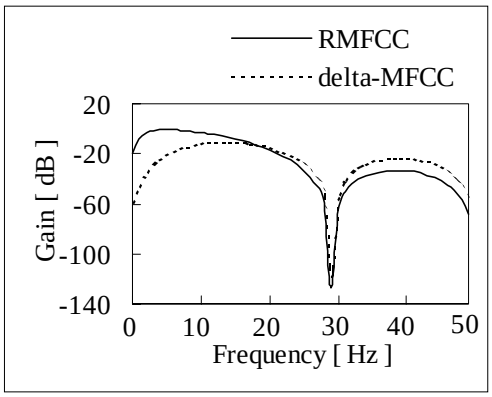

Fig.2 *Frequency responses for RMFCC and delta-MFCC*

5. CONCLUSION

In this paper, we extend RASTA processing from the log spectrum to the MFCCs and propose RMFCC processing method, and also discuss the related issues. It is shown from the experiments that the proposed method reduces the computational complexity without compromising performance comparing with RASTA, and has the advantage which does not have to estimate the long-term spectrum of the communication environment. After discussion, we find that many channel compensation methods are based on the filtering processing, and verify that it is an effective approach to suppress very low modulation frequencies for robust telephone speech recognition.

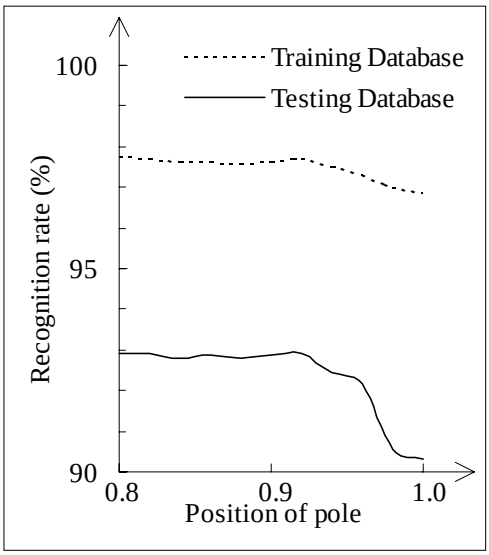

Fig.1 *Performances of different pole positions*

REFERENCES

[1]. S.Furui. Cepstral analysis technique for automatic speaker verification. IEEE Trans on. ASSP 29(2), PP.254-272, April 1981.

[2]. H.Hermansky, N.Morgan, A.Bayya, P.Kohn. RASTA-PLP Speech Analysis Technique. ICASSP, PP.I121-I124, 1992.

[3]. H.Hermansky, N.Morgan, H.Hirsch. Recognition of Speech in Additive and Convolutional Noise Based on RASTA Spectral Processing. ICASSP, PP.II83-II86, 1993.

[4]. J.Koehler, N.Morgan, H.Hermansky, H.Hirsch, G. Tong. Integrating RASTA-PLP into Speech Recognition. ICASSP, PP.I421-I424, 1994.

[5]. A.Sankar, C.Lee. Robust Speech Recognition Based on Stochastic Matching. ICASSP, PP.121-124, 1995.

[6]. C.Mokbel, J.Monne, D.Jouvet. On-Line Adaptation of a Speech Recognizer to Variations in Telephone Line Conditions. Eurospeech'93, PP.1247-1250, 1993.

[7]. H.Hermansky. Perceptual Linear Predictive (PLP) Analysis of Speech. Journal of Acoustical Society of America, 87(4), PP.1738-1752, April 1990.

[8]. J.W. Picone. Signal Modeling Techniques in Speech Recognition. Proc. of the IEEE, 81(9), PP.1215-1247, September 1993.

[9]. S. Lerner, B.Mazor. Telephone Channel Normalization for Automatic Speech Recognition. ICASSP, PP.I261-I264, 1992.

[10]. Q. Summerfield, A. Sidwell, T. Nelson. Auditory Enhancement of Changes in Spectral Amplitude. Journal of Acoustical Society of America, 81(3), PP.700-706, March 1987.

[11].N. Jayant, P. Npll. Digital Coding of waveforms. Prentice Hall, 1984.