Robust Speech Recognition Features Based on Temporal Trajectory Filtering and Non-Uniform Spectral Compression

Sang-Ho Lee, Jeong-Hyun Ha, Woo-Young Lee and Jae-Keun Hong
Department of Electronics Graduate School, Kyungpook National University
1370 Sankyuk-dong, Buk-gu, Daegu 702-701, South Korea
E-mail: bluepine74@ee.knu.ac.kr

Abstract: This paper proposes a new feature extraction method based on temporal trajectory filtering and non-uniform spectral compression and examines its performance with two tasks in noisy environments. Temporal trajectory filtering is effective for robust speech recognition in noisy environments, due to human hearing is more sensitive to relative values rather than absolute values and the effect of additive noise which varies slowly may be removed. However, even if noise is stationary, it is not removed exactly due to the random fluctuation. Thus we use non-uniform spectral compression after temporal trajectory filtering and this method shows better performances than the respective methods.

1. Introduction

The importance of the noise robust recognition system is increasing, and a lot of research has been done. In the presence of noise, the performance of automatic speech recognition systems may drastically degrade due to a mismatch between training and test environments. To alleviate this problem, many robust speech recognition techniques have been developed by many researchers. These techniques for the noise robust recognition are generally classified into three categories. The three categories are noise robust speech feature, speech enhancement, and model compensation.

The method of robust speech feature extraction is to reduce the variation of the speech representation caused by noise. The methods in this category include such techniques as relative spectra (RASTA) filtering [1], cepstral mean normalization (CMN) [2], short-time modified coherence (SMC) [3], one-sided autocorrelation LPC (OSALPC) [4], differential power spectrum (DPS) [5], relative autocorrelation sequence (RAS) [6][7] and perceptually non-uniform spectral compression (PNSC) [8-10].

Speech enhancement needs initial information about noise and tries to remove this noise from noisy speech. The methods in this category are spectral subtraction (SS) [11], the minimum mean square error short-time spectral amplitude estimator (MMSE-STSA) [12] and Wiener filtering [13].

The compensation techniques try to remove the mismatch between the trained models and the noisy speech to improve the recognition rate. Methods such as parallel model combination (PMC) [14] and maximum likelihood linear regression (MLLR) adaptation [15] are in this category.

This paper focuses on the robust speech feature extraction approach specially. MFCC based on the characteristics of the human's acoustic sense is widely used for speech recognition. Despite its widespread popularity for speech recognition, MFCC is found to be sensitive in noisy environments. Therefore, many studies have been carried out into the representation of speech signals to improve the performance of speech recognizers in noisy environments. When speech is corrupted by uncorrelated additive noise, the noise component is additive with the speech in the power spectral domain and in the autocorrelation domain. Also the sensitivity of human hearing is greater to modulation frequencies around 4 Hz than to lower or higher modulation frequencies. Therefore, the noise is removed and the relative value of speech is also emphasized by filtering the temporal trajectories of short-time power spectrum or one-sided autocorrelation sequences of speech. However, noise is not removed exactly due to the random fluctuation of noise. So the result of temporal trajectory filtering is similar to the spectrum of the original noisy speech rather than the filtering result of the clean speech in the stationary regions especially. We use non-uniform spectral compression, which applies larger energy compression to broadband-like high frequency bands of the power spectrum of each frame instead of a fixed compression for all frequency bands, after temporal trajectory filtering.

2. Extraction of Robust Features

2. 1 Temporal trajectory filtering

In this paper, we focus on the effects of additive noise on speech recognition. If $x(m,n)$ is the clean speech signal and $v(m,n)$ is the additive noise signal, then the noisy speech signal $y(m,n)$ can be modeled as

$$y(m,n) = x(m,n) + v(m,n),$$
$$0 \le m \le M-1,\ 0 \le n \le N-1, \quad (1)$$

where m is the frame index, n the discrete time index in a frame, M the number of frames and N the frame length.

If the noise is uncorrelated with the speech,

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{vv}(m,k),$$
$$0 \le m \le M-1,\ 0 \le k \le N-1, \quad (2)$$

where k is the autocorrelation index and $r_{yy}(m,n)$, $r_{xx}(m,n)$ and $r_{vv}(m,n)$ are the one-sided autocorrelation sequences of the noisy speech, clean speech and noise respectively.

Moreover, if the noise is stationary,

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{vv}(k), \quad (3)$$

$$\frac{\partial r_{yy}(m,k)}{\partial m} = \frac{\partial r_{xx}(m,k)}{\partial m}. \quad (4)$$

The differential of noisy speech autocorrelation is equal to the differential of clean speech autocorrelation. Therefore temporal trajectory filtering can emphasize the relative value of the speech and also reduce the effects of noise.

We approximate Eq. (4) to Eq. (5) in a manner similar to the RAS. Moreover, this may be applied to the power spectrum or band filter outputs as well as the autocorrelation.

$$\frac{\partial r_{yy}(m,k)}{\partial m} \cong \frac{\sum_{t=-L}^{L} t r_{yy}(m+t,k)}{\sum_{t=-L}^{L} t^2}. \quad (5)$$

where L is the length of the temporal trajectory filter.

For the calculation of one-sided autocorrelation sequence, we use the unbiased estimator as below.

$$r_{yy}(m,k) = \frac{1}{N-k}\sum_{j=0}^{N-1-k} y(m,j)y(m,j+k), \quad 0 \le k \le N-1. \quad (6)$$

2. 2 Perceptually non-uniform spectral compression

From the knowledge of psychoacoustics, spectral compression is a process that converts sound intensity into loudness. The uniform spectral compression is expressed as

$$\hat{P}(k) = P(k)^{\gamma}, \quad 0 \le \gamma \le 1, \quad (7)$$

where $\hat{P}(k)$ is the compressed power spectrum of the speech signal, $P(k)$ is the power spectrum of the original speech, γ is the compression factor and k is the DFT point or the filter band index. As γ is smaller, the mismatch or variation caused by noise is much reduced and at the same time considerable amount of information is lost. Thus the spectral compression technique is a trade-off between information and pattern mismatch. Therefore, non-uniform spectral compression may be available.

$$\hat{P}(k) = P(k)^{\gamma(k)}, \quad 0 \le \gamma(k) \le 1. \quad (8)$$

The degree of spectral compression is determined by $\gamma(k)$. As $\gamma(k)$ decreases, the effect of noise decreases, but the loss of speech information increases. Therefore the decision of $\gamma(k)$ is very important process in the spectral compression. As the result, the part that has important information is needed to be emphasized without the loss of information for noisy speech recognition, The power of speech gradually decrease as its frequency become high, but the power of white noise is flat, therefore the low frequency band has higher SNR and is robuster than high frequency band. In addition, the voiced sound segment has more energy and is robuster than the unvoiced sound segment. This means that high frequency band and the unvoiced sound segment should be compressed more strongly than low frequency band and the voiced sound segment. Therefore the compression factor is defined as

$$\gamma(k) = Ae^{-\lambda k} + A_0, \quad (9)$$

where λ is the decay parameter, and A and A_0 are used for restricting the dynamic range of the compression factor between $A + A_0$ and A_0. We set A_0 to constants. The parameters A and λ are varied according to the frame energy ρ as

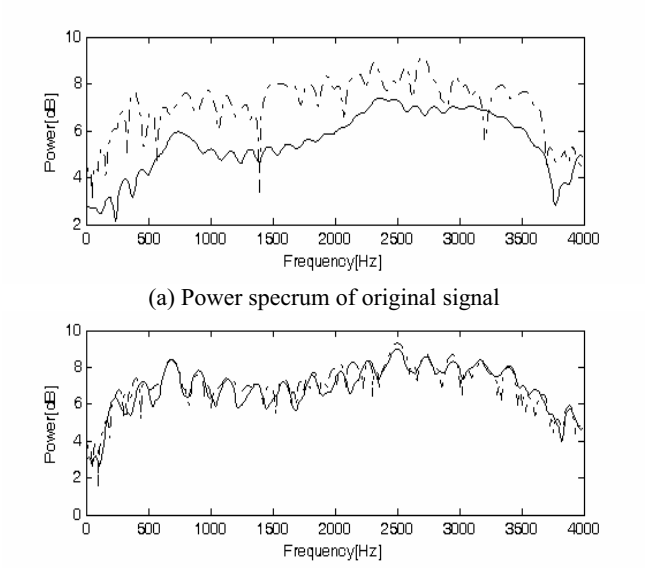
$$A = (1 - A_0)\left(\frac{1}{1 + e^{-(\rho-\mu)/\sigma}}\right) \quad (10)$$

$$\lambda = (\lambda_u - \lambda_l)\left(1 - \frac{1}{1 + e^{-(\rho-\mu)/\sigma}}\right) + \lambda_l \quad (11)$$

where μ and σ are the mean and standard deviation of frame energy against all of the frame of an utterance, and λ_u and λ_l are respectively the upper and lower bound of the decay parameter.

2. 3 Description of proposed method

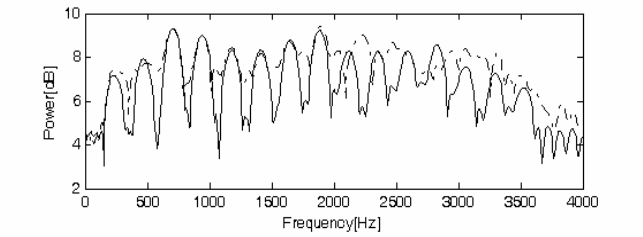
Temporal trajectory filtering can emphasize the relative value of the speech and reduce the effects of noise. However, due to the random fluctuation of noise, noise is not removed exactly. The power spectrum of the filtered signal is similar to the power spectrum of the original signal especially in the voiced segments as illustrated in Fig. 2. So in order to emphasize the temporal transition of the speech signal and also reduce the effects of unremoved noise, we propose to use non-uniform spectral compression after temporal trajectory filtering



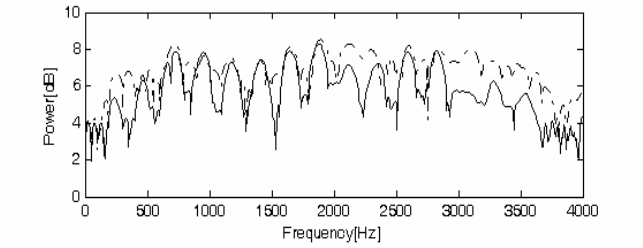
(a) Power specrum of original signal

(b) Filtered power spectrum

Fig. 1 Comparison of power spectra of clean signal (solid line) and noisy signal (dashed line) in the transition region before and after temporal trajectory filtering



(a) Power specrum of original signal



(b) Filtered power spectrum

Fig. 2 Comparison of power spectra of clean signal (solid line) and noisy signal (dashed line) in the voiced region before and after temporal trajectory filtering

Fig. 1 shows that the temporal transition of the speech signal is emphasized effectively by temporal trajectory filtering due to the transition of the speech signal is bigger than the noise. However, in Fig. 2, the results of temporal trajectory filtering depend on the random fluctuation in proportion to the magnitude of spectrum due to the speech signal as well as the noise is stationary, and are similar to the spectrums of the original signals. Therefore, High frequency band of filtered spectrum is also weaker to noise than low frequency band. These problems may be removed effectively by using PNSC after temporal trajectory filtering.

3. Experiments and Results

To evaluate the performance of the proposed method as feature extraction method for a speech recognition system, the experiments were performed using the Korean isolated-word 445DB and the Aurora2 DB.

For experiments using 445DB, The speech corpus was collected from 40 male and female speakers uttering the 445 Korean words twice at a sampling rate of 16 kHz, and following speech recognition system was used. The baseline recognition system was implemented on the HTK (Hidden Markov model Toolkit) with tri-phone model used as a basic acoustic unit. Each model is made up of 7 states and the number of mixtures per state is 6. Speech signals were analyzed with the hamming window of length-20 ms every 10 ms. The 13th order static cepstral coefficient vector was obtained from a set of 22 Mel-spaced rectangular filters. The resulted 13th order feature parameters were augmented by their delta and acceleration parameters, consequently, 39th feature parameters were obtained for speech recognition. Also, recognition experiments using Aurora DB were performed to evaluate the performance of the proposed method in real noise environments. In our experiments, clean training mode and test set A were used.

Table 1
Comparison of recognition rates for the various feature types with 445DB in white noise corruption

Feature type	Clean	30dB	25dB	20dB	15dB	10dB
MFCC	96.0	92.6	85.5	65.5	32.8	11.5
MFCC+PNSC	96.0	93.8	90.8	84.1	67.5	41.6
Log_Subband TF	94.0	91.0	88.0	81.0	70.4	51.0
Subband TF	95.1	93.0	89.9	80.2	56.7	28.2
DFT TF	96.1	94.8	91.8	80.4	53.5	27.0
Autocorrelation TF	93.3	91.2	88.7	84.0	74.2	51.9
Log_Subband TF+PNSC	93.8	91.4	89.3	83.6	71.9	51.0
Subband TF+PNSC	94.6	93.3	92.1	86.8	75.0	48.6
DFT TF+PNSC	96.0	94.6	92.6	87.3	73.6	48.2
Autocorrelation TF+PNSC	93.1	91.5	89.4	85.9	79.1	64.1

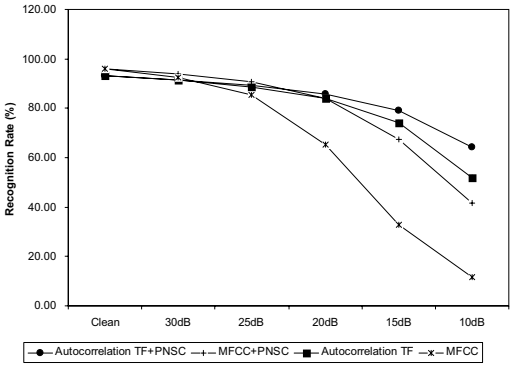


Fig. 3. Comparison of recognition rates for the methods of MFCC extraction with 445DB in autocorrelation domain.

In clean-condition training, 8440 utterances collected from 55 male and 55 female adults are used. Test set A is composed of 4004 utterances divided into four subsets with 1001 utterances in each subset. Subway, babble, car and exhibition hall noises are added to the above mentioned four subsets at SNRs of 20, 15, 10, 5, 0 and -5dB. Speech signals were analyzed with the hamming window of length-25 ms every 10 ms and the feature vectors were obtained in the same manner as 445DB. All the procedures for training and recognition were exactly identical to the reference experiments stated in Aurora2 documentation. We used a filter with $L = 2$ for temporal trajectory filtering and λ_u = 0.03, $\lambda_l = 0.01$ and $A_0 = 0.3$ for PNSC.

Table 1 shows the recognition rates using the different features for speech recognition in the presence of white noise corruption. Temporal filtering (TF) was performed in the various domains, i.e. logarithmic subband domain, subband domain, DTF domain and one-sided autocorrelation domain and the results obtained by processing temporal trajectory filtering were used for extraction of MFCC instead of the original respectively. all operations of PNSC were performed in the filter band for a simple computation, not in the DFT point. PNSC and temporal filtering in the various domains outperform conventional MFCC in severe noise conditions. Also, in all the domains except logarithmic domain, PNSC after temporal trajectory filtering shows a significant improvement. In the logarithmic domain, PNSC does not improve the performance of a speech recognizer due to the

property of logarithmic subtraction, while temporal trajectory filtering shows remarkable improvement. The result of the autocorrelation domain shows the best performance and Fig. 3 shows the results for the various methods of MFCC extraction in the autocorrelation domain.

Table 1 and Fig. 4 show the recognition rates for Aurora DB in the autocorrelation domain. All methods have achieved the better performance than standard MFCC, especially the proposed method shows the best performances in low SNR conditions, even if that results are similar to the results of PNSC

4. Conclusion

A new feature extraction technique for robust speech recognition has been developed in the part of MFCC extraction. In this paper, we proposed the new MFCC using non-uniform spectral compression after temporal trajectory filtering in order to emphasize the temporal transition of the speech signal and also reduce the effects of noise. Experimental results demonstrate improvements using the proposed method, especially in white noise corruption. Even if the proposed method didn't show remarkable improvement in real noise environments, if the parameters are optimized, much better recognition rate would be shown.

Table 2
Comparison of recognition rates for the various feature types with Aurora DB in real noise environments

Feature type	Clean	20dB	15dB	10dB	5dB	0dB	-5dB
MFCC	98.9	97.6	93.8	80.3	46.4	22.7	12.4
MFCC+PNSC	99.0	97.4	94.9	88.2	71.4	42.9	15.6
Autocorrelation TF	97.9	95.8	93.2	84.6	62.0	30.4	14.9
Autocorrelation TF+PNSC	98.0	96.0	93.5	87.2	72.5	45.5	19.6

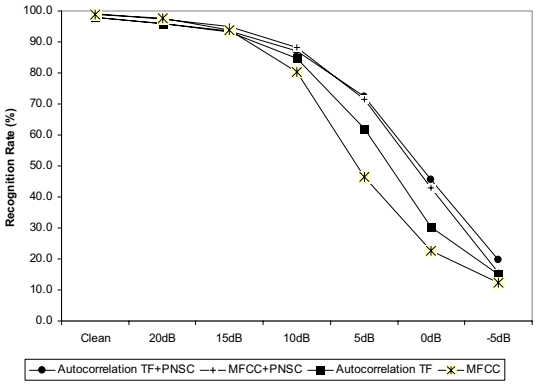


Fig. 4. Comparison of recognition rates for the methods of MFCC extraction with Aurora DB in autocorrelation domain.

References

[1] H. Hermansky, N. Morgan, "RASTA processing of speech", *IEEE Trans. Speech Audio Processing 2*, 578-589, 1994.

[2] A. Acero, R.M. Stern, "Robust speech recognition by normalization of the acoustic space", *Proceedings of IEEE International Conference of Acoust., Speech, Signal Process '91*, pp. 893-896, 1991.

[3] D. Mansour, B.H. Juang, "The short-time modified coherence representation and noisy speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing 37 (6)*, pp. 795-804, 1989a.

[4] J. Hernando, C. Nadeu, "Linear prediction of the onesided autocorrelation sequence for noisy speech recognition", *IEEE Trans. Speech Audio Processing 5 (1)*, pp. 80-84, 1997.

[5] J. Chen, K.K. Paliwal, S. Nakamura, "Cepstrum derived from differentiated power spectrum for robust speech recognition", *Speech Communication 41 (2–3)*, pp. 469–484, 2003.

[6] K.-H. Yuo, H.-C. Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences", *Speech Communication 28*, pp. 13–24. 1999.

[7] G. Farahani, S.M. Ahadi, M.M. Homayounpour, "Features based on filtering and spectral peaks in autocorrelation domain for robust speech recognition", *Computer Speech and Language 21*, pp. 187-205, 2007

[8] K. K. Chu, S. H. Leung and C. S. Yip, "Perceptually non-uniform spectral compression for noisy speech recognition", *Proc. ICASSP 2003*, pp. 404-407, 2003.

[9] K. K. Chu, S. H. Leung, "Feature extraction based on perceptually non-uniform spectral compression for speech recognition", *Proc. ISCAP 2003*, pp. 726-729, 2003.

[10] K. K. Chu and S. H. Leung, "SNR-dependent non-uniform spectral compression for noisy speech recognition", *Proc. ICASSP 2004*, pp. 973-976, 2004.

[11] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Transactions on Acoustic, Speech and SignalProcessing 27 (2)*, pp. 113–120, 1979.

[12] Y. Ephraim and D. Malah, " Speech enhancement using a minimum mean square error short-time spectral magnitude estimator, " *IEEE Trans. on ASSP*, vol. ASSP-32, pp.1109-1121, 1984.

[13] C.-H. Lee, F.K. Soong, K.K. Paliwal, "Automatic Speech and Speaker Recognition", *Kluwer Academic Publishers, Norwell*, 1996.

[14] M.J.F. Gales, S.J. Young, "Robust continuous speech recognition using parallel model combination", *IEEE Transactions Speech, Audio Processing 4 (5)*, pp. 352–359, 1996.

[15] C. Leggetter and P. Woodland, "Maximum likelihood linear regress ion for speaker adaptation of HMMs" , *Computer Speech and Language,* vol. 9, pp. 171–186, 1995.