2009 ISECS International Colloquium on Computing, Communication, Control, and Management

Noise robust speech recognition by combining speech enhancement in the wavelet domain and Lin-log RASTA

Yang Jie
School of Computer and Information
Shanghai Second Polytechnic University
Shanghai, China
e-mail: yangjie@it.sspu.cn

Wang Zhenli
Department of Electronic Information Engineering
Nanjing Institute of Communications Engineering
Nanjing, China
e-mail: dongwen3619@sina.com

***Abstract*—For improving noise robustness of speech recognition under adverse noise environment, a method of noise robust speech recognition, which combines discrete wavelet transform (DWT), wavelet packet decomposition (WPD) and Lin-log RASTA, is researched in this paper. After one scale of DWT was employed for noisy speech, this method used three scales of DWT and three scales of WPD for the low frequency signal and the high frequency signal, respectively. Multithresholds processing and decision of unvoiced sounds and voiced sounds were also adopted in order to improve the performance of denoising. The Lin-log RASTA coefficients were then computed from the enhanced speech as feature vectors. Cepstral mean subtraction (CMS) was used for compensating the speech distortion and residual noise of the above processing. Experimental results indicate that this method performs better for digital speech recognition than Lin-log RASTA, Spectral Subtraction+ Lin-log RASTA, mel-frequency cepstral coefficients (MFCC) and RASTA.**

Keywords-noise robust speech recognition; speech enhancement; Spectral Subtraction (SS); Cepstral Mean Subtraction (CMS)

I. INTRODUCTION

The robustness of speech recognition system against noise has received much attention for real applications [1-3]. Noise robustness methods of speech recognition can be classified into three major approaches: speech enhancement, robust feature coefficients and recognized model compensation. The goal of speech enhancement is to improve the quality of degraded speech. Conventional spectral subtraction [3] is widely used for its simplicity. Unfortunately, its performance of noise reduction is very limited. Although wavelet transform is a novel method for removing nonstationary noise in recent years, the only discrete wavelet transform and the only wavelet packet decomposition can't obtain good performance against noise. Robust feature coefficients based on auditive perception are well known as MFCC (Mel-Frequency Cepstral Coefficients）[4], RASTA（Relative Spectra）[5], PLP（Perceptual Linear Predictive）[6], RASTA-PLP [7], and Lin-log RASTA [8], etc. The performance of the speech recognition system using these features degrades step by step when noise level increases.

This work is supported by Education High Place Construction Foundation of Shanghai,China

In order to improve noise robustness, this paper adopts the combination algorithm of DWT and WPD, the Lin-log RASTA coefficients for digital speech recognition system. It is organized as follows: In Section Ⅱ, the speech enhancement algorithm in the wavelet domain is investigated. In Section Ⅲ and Section Ⅳ, the concept of Lin-log RASTA and cepstral mean subtraction is introduced, respectively. In Section Ⅴ, the performance of the presented method is compared to the other methods under adverse noise circumstance. In Section Ⅵ, the conclusion is presented.

II. SPEECH EHANCEMENT IN THE WAVELET DOMAIN

It is known that the energies of unvoiced segments are comparable to those of noise. And the noise components are superposed to the unvoiced sounds components in the high frequency band via wavelet transform. The unvoiced sounds is more suppressed in removing noise [9] because DWT can't divide the high frequency band into more partitions. The result is that the performance of the speech recognition system is reduced. Although WPD can overcome the limitation of DWT, it is also applied to the low frequency band signals, which mainly includes the desired signals. Hence, it costs unnecessary computation complexity.

For solving the above problem, the modified speech enhancement algorithm by combining DWT and WPD is presented in this paper. Compared to the algorithm as in [9], the new algorithm can obtain the comparable performance with the low computation complexity by ignoring the step of energy normalization. The flow graph of this algorithm is shown in Fig.1. For improving the quality of the enhanced speech, voiced/unvoiced decision need be applied to the low frequency signal via one scale of DWT. By using three scales of DWT, we define the decomposed high frequency signals as a(1,1), a(2,1), and a(3,1), respectively. Then, the high frequency signal a(1,1), a(2,1) and a(3,1) are normalized with the energy of the original high frequency signal, which is comparable to noise energy. The energies corresponding to the normalized a(1,1), a(2,1) and a(3,1) are defined as E1, E2 and E3, respectively. If one frame speech samples satisfies the condition of E1>E2>E3 [10] and E3/E1<0.99, it is determined as unvoiced segments. Otherwise it is determined as voiced segments.

978-1-4244-4246-1/09/$25.00 ©2009 IEEE

A universal threshold $T = \hat{\sigma}\sqrt{2\ln(N)}$ for removing additive white noise via DWT is presented as in [11], where $\hat{\sigma} = MAD/0.6745$ is the noise level, median absolute deviation (MAD) is estimated in the first scale, N denotes the sample size of noisy signal. In the case of wavelet packet transform, the threshold in literature [11] becomes $T = \hat{\sigma}\sqrt{2\ln(N\log_2(N))}$. The proposed threshold [11] is constant for all decomposed levels, which leads to the bad performance of noise reduction. To preserve more unvoiced sounds and improve the final recognition accuracy, the new method employs various multi-threshold for voiced segments and unvoiced segments. If determined as unvoiced segments, the threshold for the low frequency signal is given as

$$T_j = \hat{\sigma}\sqrt{2\ln(N)/\ln(j+1)} \qquad (1)$$

where j denotes the decomposition level of DWT. And the corresponding threshold for the high frequency signal becomes

$$T_j = \hat{\sigma}\sqrt{2\ln(N\log_2(N))}/\ln(j^2+1) \qquad (2)$$

If determined as voiced segments, the threshold keeps unchanged for the low frequency signal. The threshold for the high frequency signal is adjusted as $T_j = \hat{\sigma}\sqrt{2\ln(N\log_2(N))}/\ln(j+1)$, where j denotes the level of WPD.

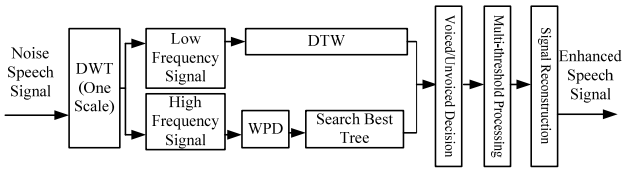


Fig.1. Block diagram of the new algorithm in the wavelet domain.

III. Lin-log RASTA

The relative spectral (RASTA) technique [5] suppresses the spectral components that change more slowly or quickly than the typical range of change of speech. Fig.1 illustrates the process of RASTA. Its transfer function of IIR filter is shown as

$$H(z) = 0.1z^4 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \qquad (3)$$

The low cut-off frequency of this filter determines the fastest spectral change of the log spectrum, which is ignored in the output, whereas the high cut-off frequency determines the fastest spectral change that is preserved in the output parameters. When operating in the logarithmic spectral domain, RASTA effectively diminishes spectral components that are additive in the logarithmic spectral domain, in particular, the fixed or slowly changing spectral characteristics of the environment. These spectral components are convolutive in the time domain and, therefore, additive in the log spectral or cepstral domain.

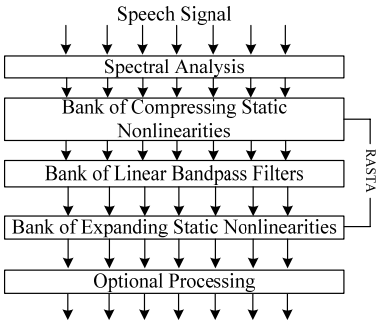


Fig.2 Block diagram of RASTA

However, uncorrelated additive noise components that are additive in the power spectral domain becomes signal dependent after the logarithmic operation on the spectrum and cannot be effectively removed by RASTA band-pass filtering in the logarithmic domain. Thus, the original RASTA processing on the logarithmic spectrum or cepstrum is not particularly appropriate for speech with significant additive noise. The equation (4) is proposed as a substitute for the logarithmic transform of RASTA as in [8]

$$y = \ln(1 + Jx) \qquad (4)$$

where J is a signal-dependent positive constant. The amplitude-warping transform is linear-like for $J << 1$ and logarithmic-like for $J >> 1$. The exact inverse transform of Eq. (4)

$$x = \frac{e^y - 1}{J} \qquad (5)$$

where e is the base of natural logarithm. The approximate inverse $x = e^y / J$ is used to ensure that the value of x is positive for all y [8].

IV. Cepstral Mean Subtraction

As noted earlier, the speech enhancement algorithm in the wavelet domain can suppress the most of noise components, whereas the enhanced speech includes the speech distortion and residual noise. In particular, the processing of Lin-log RASTA also results in a little of signal distortion and residual additive noise. In the presented method, we employ cepstral mean subtraction (CMS) [12] to avoid the above undesired components. Under the slow-variety noise circumstance, it is effective that CMS can reduce the dismatch between training and recognition by convolutive and additive noise. Supposing that the average value of speech signal in the cepstral domain is close to zero, the corresponding mean of noisy speech mainly contains the undesired average values of convolutive and additive noise in the cepstral domain, which result in the dismatch between the recognition parameters and clean training parameters. By subtracting the estimated mean of channel noise in the

cepstral domain, the average-value of noisy speech can be close to zero. Furtherly, the inverse influence of noise is eliminated.

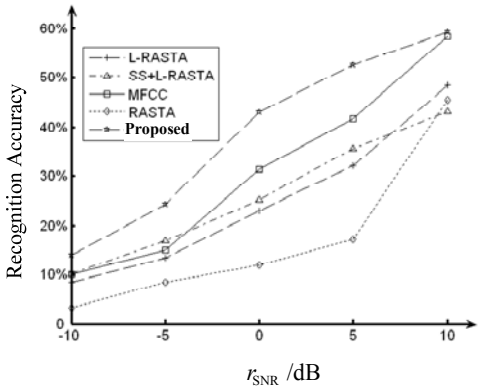
The iterative algorithm [13] is used to compute the average-value m_t, on the assumption that D_t denotes the vector of cepstral coefficients of noisy speech, m_t is the vector which denotes average value in the cepstral domain, and t is the sampling time. The average-value m_t and standard variance $\sigma_t(i)$ are initialized via the first N parameter vectors, where N is the window-width. On the base of moving the window-width N, equation (6) is used to update m_t and the samling mean-square estimation $\overline{s_t^2}$ by the succedent parameter vectors.

$$\begin{cases} m_t(i) = \lambda \cdot m_{t-1}(i) + (1-\lambda) \cdot D_t(i) \\ \overline{s_t^2(i)} = \lambda \cdot \overline{s_{t-1}^2(i)} + (1-\lambda) \cdot D_t^2(i) \end{cases} \quad (6)$$

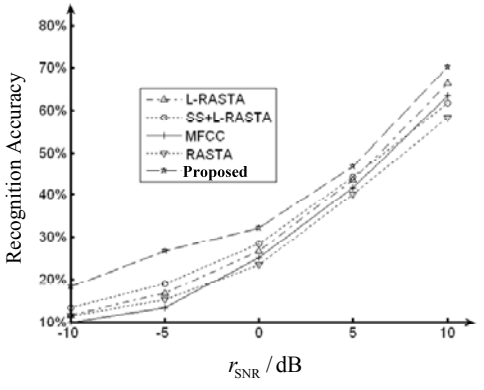
where λ is the updated step. The relationship between λ and N is $1-\lambda^N = 1/\sqrt{2}$.

V. Experimental Results and Analysis

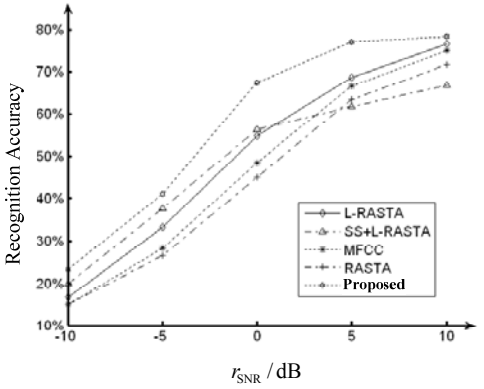
In this section, the comparison experiments are performed by applying L-RASTA（Lin-log RASTA）, SS+L-RASTA, MFCC, RASTA and the proposed method (DWT_WPD+L-RASTA+CMS) to digital speech recogni- tion system, respectively. Each of the mandarin digits 0 to 9 is read 80 times, and the former 20 times is used for training and the latter 60 times is used for recognition. The speech signal is sampled at 8 kHz rate and with each frame size of 22.5 ms (180 taps) by using Hamming window and 10 ms (80 taps) overlapping. The speech samples are quantized to 16 bit precision. The noise data, which includes white Gaussian noise, Babble noise, Leopard noise and Factory noise, is taken from Noisex-92 database. The testinged speech signal are degraded by various additive noise at low SNRs. The MFCC features containing ΔMFCC are 24 dimensional vectors. For L-RASTA with non-linear constant $J = 10^{-6}$, the order of all-pole and the number of critical-band is set to 12, 24, respectively. For each digit, two Gaussian mixtures continuous density HMM (Hidden Markov Model) which is assumed to consist of 5 emitting states are trained using the clean speech data. The topology for all models is left-right with no skips.



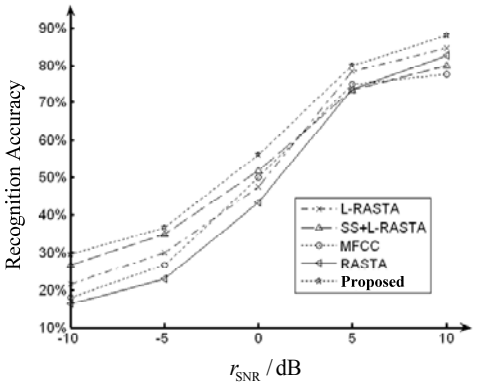
（a） white Gaussian noise



（b） Babble noise



（c） Leopard noise



（d） Factory noise

Fig.3 Comparison of the recognition accuracy at various SNRs

The recognition accuracy curves of five methods are shown in Fig.3 under various additive noise circumstance. From Fig.3 (a)-(d) we can know that the accuracy of the proposed method increases better than L-RASTA, SS+L-RASTA, MFCC and RASTA, when SNR (γ_{SNR}) ranges from -10dB to 10 dB. It is obtained by the combination processing of the proposed method, such as speech enhancement in the wavelet domain, robust parameters of L-RASTA, and the mean compensation in the cepstral domain. The improved performance is related with the noise circumstance. For instance, when γ_{SNR} =0 dB under white Gaussian circumstance, the increased accuracy of the proposed method is 20.00%, 17.83%, 11.67%, and 31.17% respectively, compared with L-RASTA, SS+L-RASTA, MFCC and RASTA. For the Factory noise condition with γ_{SNR} =0 dB, the above increased accuracy becomes 8.5%, 4.33%, 6.00%, 12.70% , respectively. From the above comparison data and the performance curves in Fig.3, the proposed method performs better against noise, compared with the other methods when γ_{SNR} ranges from -10 dB and 10 dB.

VI. CONCLUSION

In this paper, we focused on approaches to noise environmental robustness. A method of robust speech recognition was presented by combining speech enhancement in the wavelet domain and Lin-log RASTA. It reduced the undesired noise components by the speech enhancement algorithm and used Lin-log RASTA parameters for further depressed noise. The final compensation in the cepstral domain eliminated speech distortion and residual noise. The better robustness of the presented method is indicated in simulation experiments under the testing adverse noise circumstance, compared to Lin-log RASTA, SS+Lin-log RASTA, MFCC and RASTA for digital speech recognition.

REFERENCES

[1] Xiaodong Cui, Yifan Gong. “A Study of Variable- Parameter Gaussian Mixture Hidden Markov Modeling for Noisy Speech Recognition”. IEEE Trans.on Audio, Speech, and Language Processing, 2007, vol.15, no.4, pp. 1366-1376.

[2] Chang-Hoon Lee, Soo-Young Lee. “Noise-Robust Speech Recognition Using Top-Down Selective Attention With an HMM Classifier”. IEEE Signal Processing Letters, 2007, vol.14, no.7, pp.489-491.

[3] Steven F B. “Suppression of acoustic noise in speech using spectral subtraction”. IEEE Trans.on Speech and Audio Processing, 1979, vol.27, no.2, pp.113-120.

[4] Davis S B, Mermelstein P. “Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences”. IEEE Trans.on Speech and Audio Processing, 1980, vol.28, no.4, pp.357-366.

[5] Hermansky H, Morgan N. “RASTA Processing of Speech”. IEEE Trans.on Speech and Audio Processing, 1994, vol.2, no.4, pp.578-589.

[6] Hermansky, H. “Perceptual linear predictive (PLP) analysis of speech”. J. Acoust. Soc. Am., 1990, vol.87, no.4, pp. 1738 -1752.

[7] Hermansky H, Morgan N, Bayya A and Kohn P. “RASTA-PLP speech analysis technique”. In Proc. IEEE Int. Conf. Acoust.,Speech,Signal Processing, 1992, vol.1, pp.121-124.

[8] Morgan N, Hermansky H. “RASTA extensions, Robustness to additive and convolutional noise”. In Proc. Workshop Speech Processing Adverse Environments, 1992, pp.115-118.

[9] Wang Zhenli, Zhang Xiongwei, Zheng Xiang, and Yang Jian. “A new wavelet domain speech enhancement method”. Signal Processing, 2006, vol.22, no.3, pp.325 -328. (in Chinese)

[10] Li Chongni; Hu Guangrui. “A modified wavelet domain speech enhancement method”. Journal of China Institute of Communications, 1999, vol.20, no.4, pp.88-91. (in Chinese)

[11] Donoho D L. “De-noising by soft-thresholding”. IEEE Trans.on Inform Theory, 1995, vol.41, no.3, pp.613-627.

[12] F.H. Liu, A. Acero, and R. Stern. “Efficient Joint Compensation of Speech For the Effects of Additive Noise and Linear Filtering”. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992, vol.1, pp.257-260.

[13] Viildu 0, Bye D, Iaurila K. “A recursive feature vector normalization approach for robust speech recognition in noise” [A].Proceedings‘ICASSP'98[C].Seattle, WA, USA: IEEE Acoustics, Speech and Signal Processing Society, 1998, pp.733-736.