2014

ISSN : 0974 - 7435 Volume 10 Issue 20

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(20), 2014 [12349-12354]

A noise reduction method with modified running spectrum filtering

Zhang Yuxin*, Ding Yan
School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, 130022, (CHINA)
E-mail : zyx@cust.edu.cn

ABSTRACT

In this paper we proposed a modified running spectrum filtering(RSF) method for noise reduction. RSF is an efficient noise reduction method. Because conventional RSF used FIR filter with 240 orders for reducing noise twice, the calculation cost is very high. Thus, removing low-frequency components with a bandpass filter can reduce the additive noise in power spectra. On the other hand, the Cepstrum Mean subtraction Algorithm (CMS) method is used to reduce the remanent noise in whole frequency domain in logarithm spectra. CMS is simpler than RSF. The calculation cost is far lower than that of RSF. The experiment shows proposed method not only improves the recognition accuracy of automatic speech recognition, but also reduces the calculation cost of conventional RSF.

KEYWORDS

Running spectrum filtering; Cepstrum mean subtraction; Automatic speech recognition.

INTRODUCTION

In the early research of speech recognition, the standard speech databases are recorded on the quiet circumstances. Thus, the better recognition accuracy can be got with the recognition system that speeches are trained or created to reference models on the quiet circumstances. As the application of speech recognition system, the recognition environment is more complexity. In real noise environment, the recognition performance is drastically reduced because the feature vectors are discrepant between the noisy speeches and the reference models, which are created under the quiet circumstances[1].

The noise reduction technique can reduce the noise and extract the real speech from the noisy speech. It tries to increase the acoustic feature of real speech signal possibly, in order to improve the recognition accuracy of ASR system. The running spectrum filtering (RSF)[2,3] and Cepstrum mean subtraction Algorithm (CMS)[4,5] are popular noise reduction methods. RSF removes low-frequency components with a high-pass filter can reduce the noise. On the other hand, it use a band-pass filter to separate speech from noise. The additive noise is reduced in the power spectra and the multiplicative noise is reduced in the logarithm spectra. Because conventional RSF used FIR filter with 240 orders for reducing noise twice, the calculation cost is very high.

In this paper, we propose a modified RSF method. We remove low-frequency components with a bandpass filter can reduce the additive noise in power spectra. On the other hand, the CMS method is used to reduce the remanent noise in whole frequency domain in logarithm spectra. CMS is simpler than RSF. The calculation cost is far lower than that of RSF. The experiment shows proposed method not only improves the recognition accuracy of automatic speech recognition, but also reduces the calculation cost of conventional RSF.

NOISE REDUCTION METHOD

Influence of additivity and multiplicative noises

Usually, there are two kinds of noise by interrelation between single and noise. One is additivity noise, the other is multiplicative noise. The additivity noise and speech signal are independent with each other. It exists in all the time whether there are speech signal or not. We can only reduce the influence of additivity noise, but cannot eliminate the additive noise completely. Thus, the additivity noise can affect the speech signal inevitably. The multiplicative noise is usually caused with the unfavorable channel. It exists with the presence of speech signal. If the speech signal disappear, then the multiplicative noise is also disappear. In the time domain, we assume the interfered speech signal by additivity noise is $x(t)$:

$$x(t) = s(t) + n(t) \tag{1}$$

The $x(t)$ is made fourier transform, then the corresponding relation is follow in the frequency domain and power spectrum.

$$|X(t;\ i)|^2 = |S(t,i) + N(t,i)|^2$$

$$= |S(t,i)|^2 + |N(t,i)|^2 + 2|S(t,i)||N(t,i)|cos\big(\theta(t,i)\big) \tag{2}$$

Where $X(\cdot)$ is spectrum of the mixed superimposed signal, $S(\cdot)$ is spectrum of speech signal, $N(\cdot)$ is spectrum of additivity noise. t is frame index, i is the frequency components index of the t frame. $\theta(t,i)$ is the phase separation between speech signal and additivity noise on the i^{th} point. If the speech signal and additivity noise are assumed as independent distribution of zero-mean, then

$$|X(t,i)|^2 \approx |S(t,i)|^2 + |N(t,i)|^2 \tag{3}$$

If we can extrapolate the $|N(t,i)|^2$, then the additivity noise can be removed in the frequency component $|S(t,i)|^2 = |X(t,i)|^2 - |N(t,i)|^2$, e.g., spectral subtraction(SS) method. These methods are based on that additivity noise is considered to approximately invariable. In fact, it is very difficult to extrapolate the power of additivity noise accurately. After subtracting the $|N(t,i)|^2$, a few additivity noise is still left. Furthermore, the distribution of additivity noise is variable, but the method is same.

Moreover, we can analyze the frequency spectral by the $|N(t,i)|^2$ in the all spectrum components. The frequency components which is most of $|N(t,i)|^2$ can be filtered with filter. The method can remove most of noise, but it is also very difficult to confirm the frequency of additivity noise. Some additivity noise is still left.

It is impossible that the multiplicative noises are removed with aforementioned two methods. Because of the multiplicative noise is appeared alongside of speech noise. In order to remove the multiplicative noise, the interfered speech noise must be processed. We assume the interfered speech noise by multiplicative noise is

$$x(t) = s(t) \otimes h(t) \tag{4}$$

The $x(t)$ is made fast Fourier transform (FFT), then the $x(t)$ is transformed as

$$X(t,i) = S(t,i) \cdot H(t,i) \quad \textbf{(5)}$$

Where $X(\cdot)$ is spectrum of the mixed superimposed signal, $S(\cdot)$ is spectrum of speech signal, $H(\cdot)$ is spectrum of multiplicative noise, t is frame index, i is the frequency components index of the t^{th} frame. Eq. (5) is made logarithms transformation on both sides.

$$log|X(t,i)| = log|S(t,i)| + log|H(h,i)| \quad \textbf{(6)}$$

then, made cepstrum transformation on both sides.

$$X_{cep}(t,n) = S_{cep}(t,n) + H_{cep}(t,n) \quad \textbf{(7)}$$

Where $X_{cep}(\cdot)$ cepstrum of the mixed superimposed signal is, $S_{cep}(\cdot)$ is cepstrum of speech signal, $H_{cep}(\cdot)$ is cepstrum of additivity noise. n is the number of channel. Then, it is same as the additivity noise, we can extrapolate the $H_{cep}(t,n)$, and then the multiplicative noise can be removed in the frequency component $H_{cep}(t,n) = X_{cep}(t,n) - S_{cep}(t,n)$.

RSF algorithm

RSF is similar to relative spectral (RASTA), which is proposed by Hermansky et al.[6,7] RASTA is that speech signal is filtered by a band-pass filter in each frequency channel, according to time tract of speech parameter. RASTA uses a bandpass filter with a sharp spectral zero at the zero frequency to cut-off slowly changing or steady-state factors in speech spectrum. RASTA uses an infinite impulse response (IIR) filter. We know that the output value is calculated with current input and last output values. Hence, the effect of steady background noise is still residue after much iteration. In order to cut-off the effect of input signal, the RSF uses FIR filter instead of IIR filter.

In the modulation spectrum, we have found that the noise spectrum is concentrated in the direct component (DC). Most of the noise energy is distributed in the low-frequency band of the modulation spectrum. Through analyzing the power spectrum of speech and noise, we found additivity noise on frequency band [0,1]Hz exerts such tremendous effect on speech signal. To speech recognition, the important information of speech is about in the frequency band[1,16]Hz. The multiplication noise exerts such tremendous influence on frequency band of close to 0Hz.

Thus, removing low-frequency components with a high-pass filter can reduce the noise. On the other hand, we can use a bandpass filter to separate speech from noise. The additive noise is reduced in the power spectra and the multiplicative noise is reduced in the logarithm spectra by RSF.

CMS algorithm

CMS is a simple method of reducing noise. White noise is uniformly distributed in a spectrum. After feature extraction, the MFCC feature vectors are obtained in the cepstral domain. In a long-time range, almost all speech features are changed with the progress of time. On the other hand, the time-invariant noise features in such a range are considered as almost constant. The subtraction of the time-invariant features from noisy speech features result in the reduction of noise components. We assume that a speech waveform is divided into h short frames. $f_i(t)$ is the t^{th} component of the i^{th} frame. Noise reduction is then executed as Eq. (8).

$$f_i'(t) = f_i(t) - \frac{1}{h}\sum_{j=1}^{h} f_j\,(t) \quad \textbf{(8)}$$

Dynamic range adjustment algorithm (DRA)

Usually, when white noise is added to a speech waveform, observing the speech waveform is more difficult than observing the clean speech. In addition, when RSF or CMS is applied for noise reduction, the signal amplitude is typically reduced.

DRA[8] can be used to compensate for this difference using the following normalization.

$$f_i'(t) = \frac{f_i(t)}{arg\,max_{j=1,\cdots,h}|f_i(t)|} \quad \textbf{(9)}$$

DRA makes it possible to obtain similar cepstrum data for clean speech and noisy speech after CMS or RSF. However, the shape of waveform is kept same to original one.

Proposed RSF algorithm

In real environment, the additivity and multiplicative noises are simultaneous. Hence, the mixed superimposed speech waveform is as follow in time domain.

$$x(t) \; = \; s(t) \otimes h(t) + n(t) \qquad \textbf{(10)}$$

Where $x(t)$ is noisy speech signal, $s(t)$ is speech signal, $h(t)$ is multiplicative noise, and $n(t)$ is additivity noise. The Eq. (10) is Fourier transformed on both sides. In frequency and power spectrums, the equation is followed, which is effected by the additivity and multiplicative noises.

$$X(t,i) = S(t,i)H(t,i) + N(t,i) \qquad \textbf{(11)}$$

$$\begin{aligned} &|X(t,i)|^2 = |S(t,i)H(t,i) + N(t,i)|^2 \\ &= |S(t,i)H(t,i)|^2 + |N(t,i)|^2 + 2Re[S(t,i)H(t,i)N(t,i)] \\ &= |S(t,i)|^2|H(t,i)|^2 + |N(t,i)|^2 \\ &+2|S(t,i)||H(t,i)||N(t,i)|cos\big(\theta(t,i)\big) \end{aligned} \qquad \textbf{(12)}$$

Where $\theta(t,i)$ is the phase separation between speech signal and additivity noise on the i^{th} point. Because that the speech and noise can be supposed as mutually independent zero-mean distribution, the desired value of last item is zero in Eq. (12). Although instantaneous value of each frame is not zero in this item, the output value of each filter unit is equal to weighted sum of energies of all points when computing Mel-filter. Hence, Mel-energy of noisy speech signal is approximately equal to

$$P_x(t,i) \;\approx\; P_s(t,i)P_h(t,i) + P_n(t,i) \qquad \textbf{(13)}$$

where $P_x(\cdot)$,$P_s(\cdot)$,$P_h(\cdot)$, and $P_n(\cdot)$ are Mel-energy of noisy speech, clean speech, additivity noise, and multiplicative noise.

In logarithm spectrum, we defined X^{log}, S^{log}, N^{log}, and H^{log} are as values of vector for noisy speech, clean speech, additivity noise, and multiplicative noise. So

$$X^{log} = S^{log} + H^{log} + log(I + e^{(N^{log} - S^{log} - H^{log})}) \qquad \textbf{(14)}$$

Similarly, we defined X^{cep}, S^{cep}, N^{cep}, and H^{cep} are as values of cepstrum feature vector for noisy speech, clean speech, additivity noise, and multiplicative noise in cestrum spectrum. So

$$X^{cep} = S^{cep} + H^{cep} + Dlog\big(I + e^{D^{-1}(N^{cep} - S^{cep} - H^{cep})}\big) \qquad \textbf{(15)}$$

where D is discrete cosine transformation (DCT) matrix.

In Eq. (15), the H^{cep} can be almost removed by RSF, but the effect of $Dlog\big(I + e^{D^{-1}(N^{cep} - S^{cep} - H^{cep})}\big)$ is in the whole modulation frequency domain.

We know the calculation cost of RSF algorithm is high, since the high order (240) is used. Moreover the conventional RSF algorithm is used for reducing noise twice. One is in power spectrum; the other is in logarithm spectrum. Hence, the calculation time of ASR system with RSF is relatively high. In order to improve the performance of ASR system, we remove the RSF for noise reduction in power spectra. After cepstrum computing, we use RSF with band-pass filter to reduce the noise. And then, CMS method is used to reduce the remanent noise in whole frequency domain. CMS is simpler than RSF. The calculation cost is far lower than that of RSF. The flowchart of this method is shown in Figure 1.

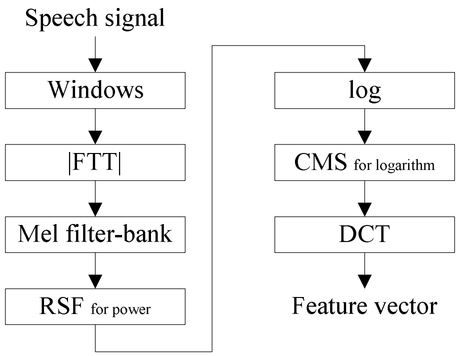

Figure 1 : Overview of modified RSF method

EXPERIMENT AND RESULT

Our noise reduction is implemented in Matlab. The pattern comparison method is dynamic time warping (DTW) with nonlinear media filter method (NMF), which has been proposed in paper[9]. The reference database comprises reference speech waveforms (utterances) for 100 isolated Japanese words. For each word, the database contains 50 or waveforms for words uttered by 60 distinct speakers. Further details of the experimental settings and parameters appear in TABLE 1.

TABLE 1 : The experimental settings and parameters

parameter	value
Recognition task	100 isolated Japanese words
Sampling	11.025 kHz, 16 bits
Window length	23.2 ms (256 samples)
Shift length	11.6 ms (128 samples)
Band of band-pass filter	1-16Hz
Feature vector	MFCC, 38 dimensions
Noise type	White noise, babble noise

TABLE 2 : Recognition accuracy for five noise reduction methods (%)

Noise		Noise reduction method				
Name	SNR	CMS+DRA	RSF+DRA	CMS+RSF+DRA	DRA	Nothing
White	10dB	84.74	86.54	86.58	59.62	24.64
	20dB	97.5	97.7	97.74	92.04	80.52
Babble	10dB	81.2	81.74	81.82	71.34	33.74
	20dB	96.56	95.6	96.14	93.76	84.92
Clean		98.96				

TABLE 2 shows the recognition accuracy of five noise reduction methods. The recognition accuracy improves greatly with CMS, RSF and DRA. RSF is better than CMS method. They show that CMS&DRA and RSF&DRA outperform other noise reduction methods. Furthermore, it shows that modified RSF better than conventional RSF approach. Moreover, the recognition accuracy of proposed method is close to that in clean environment.

CNCLUSION

In this paper, we propose a modified RSF method. The additivity noise on frequency band [0,1]Hz exerts such tremendous effect on speech signal. To speech recognition, the important information of speech is about in the frequency band[1,16]Hz. RSF is similar to RASTA that speech signal is filtered by a band-pass filter in each frequency channel. However, in order to cut-off the effect of input signal, the RSF uses FIR filter instead of IIR filter. Thus, the calculation cost of RSF algorithm is high, since the high order (240) is used for FIR. Moreover the RSF algorithm is used for reducing noise twice. Because the RSF only uses bandpass filter, the noise cannot be reduced, which is distributed in the whole frequency spectrum. Thus, we use the CMS instead of RSF in logarithm spectrum. The experiment shows proposed method not only improves the recognition accuracy of ASR, but also reduces the calculation cost of conventional RSF.

REFERENCES

[1] J.C.Junqua, J.Haton; Robustness in Automatic Speech Recognition: Fundamentals and Applications, 1st Ed., Kluwer Academic, **(1995)**.

[2] N.Hayasaka, K.Khankhavivone, Y.Miyanaga, K.Songwatana; New robust speech recognition by using nonlinear running spectrum filter, Paper presented at International Symposium on Communications and Information Technologies, 133–136 **(Oct. 2006)**.

[3] N.Hayasaka, Y.Miyanaga; Spectrum filtering with FRM for robust speech recognition, Paper presented at IEEE International Symposium on Circuits and Systems, 3285–3288 **(Nov. 2006)**.

[4] D.Naik; Pole-filtered cepstral mean subtraction, Paper presented at International Conference on Acoustics, Speech, and Signal Processing, 157–160 **(May 1995)**.

[5] M.Rahim, B.H.Juang, W.Chou, E.Buhrke; Signal conditioning techniques for robust speech recognition, Paper presented at IEEE Signal Processing Letters, 107–109 **(Apr. 1996)**.

[6] M.Holmberg, D.Gelbart, W.Hemmert; "Automatic speech recognition with an adaptation model motivated by auditory processing, Paper presented at IEEE Transactions on Audio, Speech, and Language Processing, 43–49 **(Jan. 2006)**.

[7] M.Grimaldi, F.Cummins; Speaker identification using instantaneous frequencies, Paper presented at IEEE Transactions on Audio, Speech, and Language Processing, 1097–1111 **(Aug. 2008)**.

[8] N.Wada, S.Yoshizawa, N.Hayasaka, Yoshikazu Miyanaga; Robust speech feature extraction using RSF/DRA and burst noise skipping, Paper presented at Transactionson Electrical Engineering, Electronics, and Communications (ECTI-EEC), 100–107 **(Aug. 2005)**.

[9] Y.Zhang, Y.Miyanaga, C.Siriteanu; Robust Speech Recognition with Dynamic Time Warping and Nonlinear Median Filter, Journal of signal processing, **16(2)**, 147-157 **(2012)**.