논문 2008-03-01

A Noise Reduction Method Combined with HMM Composition for Speech Recognition in Noisy Environments

Guanghu Shen, Ho-Youl Jung, Hyun-Yeol Chung*

Abstract : In this paper, a MSS-NOVO method that combines the HMM composition method with a noise reduction method is proposed for speech recognition in noisy environments. This combined method starts with noise reduction with modified spectral subtraction (MSS) to enhance the input noisy speech, then the noise and voice composition (NOVO) method is applied for making noise adapted models by using the noise in the non-utterance regions of the enhanced noisy speech. In order to evaluate the effectiveness of our proposed method, we compare MSS-NOVO method with other methods, i.e., SS-NOVO, MWF-NOVO. To set up the noisy speech for test, we add White noise to KLE 452 database with different SNRs range from 0dB to 15dB, at 5dB intervals. From the tests, MSS-NOVO method shows average improvement of 66.5% and 13.6% compared with the existing SS-NOVO method and MWF-NOVO method, respectively. Especially our proposed MSS-NOVO method shows a big improvement at low SNRs.

Keywords : Modified spectral subtraction, Wiener filtering, HMM composition, NOVO

I. Introduction

Automatic speech processing systems are employed more and more often in real environments. However, they are confronted with high ambient noise levels and their performance degrades drastically. Thus, there has been a strong need to improve the performance of these systems in noisy environments.

The research work in noisy speech recognition may be classified into three broad categories:

a) Noise reduction method prior to classification. Techniques in this category include spectral subtraction (SS)[1], Wiener filtering (WF)[2].

b) Hidden markov model (HMM) composition method is to adapt the speech model, which has the effects of noise. Methods using this approach including bias compensation algorithm[3] and parallel model combination (PMC)[3-5].

c) Use features that are robust to noise.

Usually, SS and WF are used to reduce noise from noisy speech. These methods, however, are not able to remove noise completely, and they create new problems in that insufficient or excessive reduction processing leads to remaining noise or speech distortion, respectively. The HMM composition method, i.e. noise and voice composition (NOVO) and PMC, are well-known as effective noise adaptation methods that can improve speech recognition performance in noisy environments. However, performance of HMM composition method like NOVO or PMC, is not sufficient when SNR is low, because the feature of speech is buried in the noise. Approaches that combine the HMM composition method with noise reduction

*Corresponding Author
Manuscript received Feb. 25, 2008 ;
accepted Apr. 10, 2008.
Guanghu Shen, Ho-Youl Jung, Hyun-Yeol Chung : Yeungnam University

method such as SS, modified Wiener filtering (MWF)[6] have been proposed, for example SS-NOVO[6], MWF-NOVO[6]. These approaches include adaptation processing to handle the inevitable speech distortion caused by noise reduction method, but they have difficulties to get high improvement in low SNR environments. Therefore, we propose MSS-NOVO method, which is a combination of NOVO and modified spectral subtraction (MSS) instead of Wiener filter, to improve the recognition performance in noisy environments by decreasing the distortion of speech in the process of noise reduction.

The rest of this paper is organized as follows. In section 2, we describe the proposed noise processing method, and experimental results are shown in Section 3. Finally, we conclude this paper in section 4.

II. Noise Processing Method

2.1 Noise Reduction Method

2.1.1 Modified Wiener Filtering (MWF)

The noise reduction process is based on a short-time spectral amplitude (STSA) estimation[2]. Furthermore, the background noise is considered additive and uncorrelated to the speech signal. The noisy speech can be modeled as

$$x(\lambda) = s(\lambda) + n(\lambda), \qquad (1)$$

where λ, $s(\lambda)$, $n(\lambda)$, and $x(\lambda)$ denote the time index, the original speech, the noise and the input noisy speech, respectively.

A Wiener filter is the optimal Bayesian linear filter that minimizes the expected mean squared error (MSE) $E[|\hat{s}(\lambda) - s(\lambda)|^2]$ for the noise corruption model in equation (1).

In achieving noise reduction based on STSA estimation, the speech spectrum $\tilde{S}(\lambda,k)$ is estimated as being obtained from $X(\lambda,k)$ by a multiplicative nonlinear gain function $G(\lambda,k)$ in the frequency domain.

$$\tilde{S}(\lambda,k) = G(\lambda,k)X(\lambda,k) \qquad (2)$$

The gain function based on Wiener filter, $G(\lambda,k)$, is given by

$$G(\lambda,k) = \frac{|S(\lambda,k)|^2}{|S(\lambda,k)|^2 + |N(\lambda,k)|^2}, \qquad (3)$$

where $S(\lambda,k)$, $N(\lambda,k)$ are respectively the frequency spectrum of the original speech and noise. Figure 1 illustrates the block diagram of Wiener filtering for noise reduction.

In [7], the noise reduction method minimizes speech distortion by adding a part of the untreated input noisy speech $X(\lambda,k)$ to the output of the enhanced speech $\hat{S}(\lambda,k)$. Finally, the speech after noise reduction $\tilde{S}(\lambda,k)$ is given by

$$\tilde{S}(\lambda,k) = (1-\alpha)X(\lambda,k) + \alpha\hat{S}(\lambda,k). \qquad (4)$$

The added untreated input noisy speech masks the distorted components of the speech so that speech quality will be improved in a poor SNR environment.

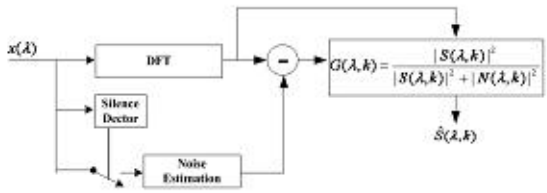

Fig. 1. Block diagram of Wiener filtering

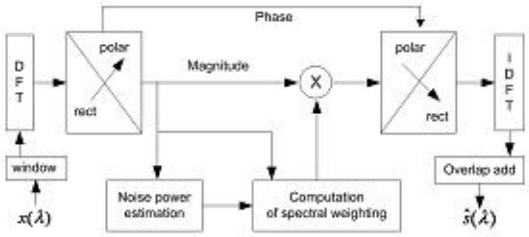

Fig. 2. Block diagram of spectral subtraction

2.1.2 Modified Spectral Subtraction (MSS)

A block diagram of the basic spectral subtraction method is shown in Fig. 2. Spectral subtraction relies upon the assumption that the background noise signal has an almost constant magnitude spectrum and the speech signal is short-time stationary. Let $s(\lambda)$, $n(\lambda)$, and $x(\lambda)$

represent the original speech, noise and input noisy speech, respectively. Then noisy speech can be modeled as equation (1).

Spectral processing is based on a discrete fourier transform (DFT) filter bank with W_{DFT} sub-bands. The phase of the disturbed signal is not modified. We denote the data window by $h(\lambda)$ and the DFT of the windowed disturbed signal $x(\lambda)$ by

$$X(\lambda,k)=\sum_{\mu=0}^{W_{DFT}-1} x(\lambda)\cdot h(\mu)\cdot \exp(-j\frac{2\pi\mu k}{W_{DFT}}), \quad (5)$$

where k denotes frequency bin. Typically we use a DFT length of $W_{DFT}=256$. The enhanced speech are converted back to the time domain using an inverse DFT. The synthesized enhanced speech is denoted by $\hat{s}(\lambda)$, the corresponding magnitude spectra by $|\hat{S}(\lambda,k)|$.

Let $P_n(\lambda,k)$ and $\overline{|X(\lambda,k)|^2}$ denote the estimated noise power specturm and input noisy speech power spectrum, respectively. To obtain the stable power spectrum of input noisy speech, we do smoothing process with a first order recursive network as follows. ($\gamma \le 0.9$)

$$\overline{|X(\lambda,k)|^2}=\gamma\cdot\overline{|X(\lambda-1,k)|^2}+(1-\gamma)\cdot|X(\lambda,k)|^2 \quad (6)$$

Following the proposal of Berouti et. al. [8], we subtract the magnitude spectra with an over-subtraction factor $osub(\lambda,k)$ and a limitation of the maximum subtraction by a spectral floor constant subf ($0.001 \le subf \le 0.05$).

$$|\hat{S}(\lambda,k)|=\begin{cases}\sqrt{subf\cdot P_n(\lambda,k)} & if\ |X(\lambda,k)|\cdot Q(\lambda,k) \le \sqrt{subf\cdot P_n(\lambda,k)}\\ |X(\lambda,k)|\cdot Q(\lambda,k) & otherwise\end{cases}$$

$$where\ Q(\lambda,k)=\left(1-\sqrt{osub(\lambda,k)\frac{P_n(\lambda,k)}{|X(\lambda,k)|^2}}\right) \quad (7)$$

While a large over-subtraction factor $osub(\lambda,k)$ essentially eliminates residual spectral peaks ('musical noise'), it also affects speech quality such that some of the low energy phonemes are suppressed. To limit this undesirable effect the over-subtraction factor is computed as a function of the subband $SNR_k(\lambda,k)$ and the frequency bin k, i.e., $osub(\lambda,k)=f(\lambda,k,SNR_k(\lambda,k))$. In general we use less over-subtraction for high SNR conditions and for high frequencies than for low SNR conditions and for low frequencies. When noise power spectrum is estimated to be larger than its real value, distortion will be happened and speech quality will be degraded.

According to the idea in [7], we also reduce the speech distortion by adding a small amount of the untreated input speech to the output enhanced speech obtained by the conventional spectral subtraction method.

And, to find the optimum value of α, we also did some preliminary experiments. From the results, we found that 0.8 is the best value of the adding ratio α and this value is also used in subsequent experiments with NOVO method. The block diagram of modified spectral subtraction method is shown in Fig. 3.

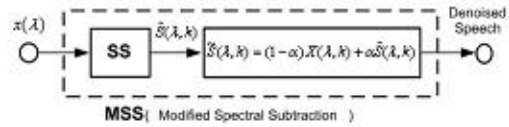

Fig. 3. Block diagram of modified spectral subtraction

2.2 HMM Composition Method

2.2.1 Noise and Voice Composition (NOVO)

HMM composition method assumes that the adapted HMM obtained by combining two or more "source HMMs" will adequately model the input noisy speech. The source HMMs may model clean speech recorded in noise-free conditions or the various noise sources, such as stationary or non-stationary noises, background voices, etc. The NOVO HMM is the product of two or more source HMMs. An example of this process is

represented in Fig. 4. All parameters of the NOVO HMM except its output probabilities can be directly deduced from the source HMMs as product of the corresponding parameters of the source HMMs. Usually, Mel Frequency Cepstral Coefficients (MFCCs) is efficient for modeling speech in noisy environments. Therefore, we investigated the application of HMM composition method to MFCCs.

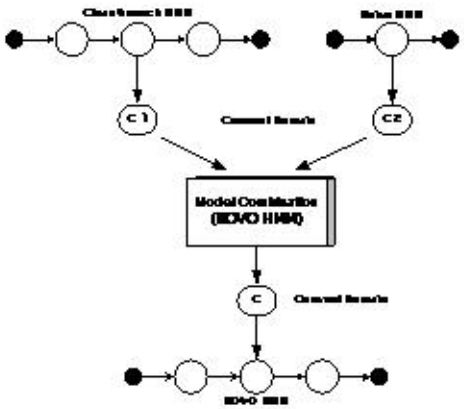

Fig. 4. Basic process of NOVO method

To compute the output probabilites of the NOVO HMM, a domain in which the relationship between the sources can be seated explicitly and as simply as possible is need to easily deal with the distributions of the corresponding random variables. The linear spectrum was chosen because clean speech and noise are additive within it.

Therefore the random variables are related by

$$R_{ln} = S_{ln} + N_{ln}[k(SNR)] \qquad (8)$$

$k(SNR)$ is a weighting factor that depends on the SNR between the clean speech and the noise. Figure 5 shows in detail how the distributions of the NOVO HMM can be inferred for one Gaussian output probability distribution. This process has to be repeated for all states and for all mixture components of the speech and noise HMMs. We assume that all the distributions handled in the cepstrum domain are normal, so it will be enough to obtain their first and second moments to determine them.

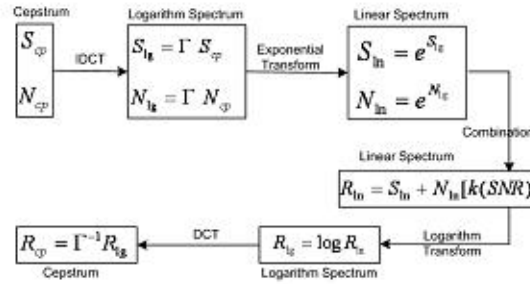

Fig. 5. Transform process of NOVO method

Let S_{cp} and N_{cp} are supposed to have a multivariate Gaussian distribution. Therefore, their linear combination is still Gaussian distribution and the parameters of N_{lg} (same for S_{lg}) are given by

$$\mu^{N_{lg}} = \Gamma \cdot \mu^{N_{cp}} \quad ; \quad \mu^{N_{lg}} = \Gamma \cdot \Sigma^{N_{cp}} \cdot \Gamma^{T}, \qquad (9)$$

where Γ is the transform operator.

2.3 MSS-NOVO

An approach that combines the HMM composition method with noise reduction method is the solution which solves speech distortion happened in noise reduction method and low recognition performance of HMM composition method at low SNR. For example, the NOVO method is effective in noisy environments with the normal level SNRs, but it has a significant problem that its recognition performance is severely degraded at low SNR. To address this problem, we propose the combined MSS-NOVO method, which improves the SNR after noise reduction by using the MSS method and compensates the remaining noise by using the NOVO method. In this case, the NOVO method should be showed higher performance, because it handles the speech data which has higher SNR than before. And

we can also further expect that the combined MSS-NOVO method shows a good performance in low SNR environments.

Figure 6 shows the framework of MSS-NOVO method. At first, it reduces noise with MSS method. Then the noise that extracted from the non-utterance regions of the enhanced speech is used to adapt clean speech HMM to yield the noise adapted NOVO HMM.

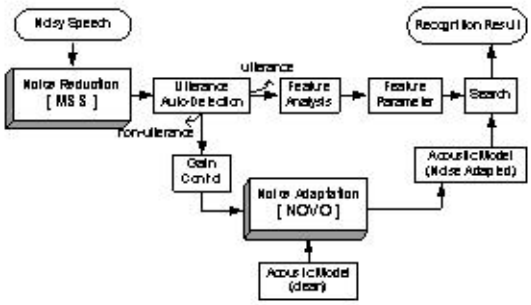

Fig. 6. Block diagram of MSS-NOVO method

Ⅲ. Experimental Results

3.1 Parameters

To evaluate the performance of the proposed method, the speech data used are KLE (center for Korean Language Engineering) 452 database, which is a phonetically balanced isolate word database. The corpus includes 38 male speakers and each speaker pronounces 452 words for one times. Those were divided into two subsets. One has 15,280 words of 35 speakers' utterance, which used for training clean speech HMMs. Another has 1,356 words of 3 speakers' utterance, which used for testing. To set up the noisy speech for test, we add White noise to clean speech data with different SNRs range from 0dB to 15dB, at 5dB intervals.

All speech data were sampled at 16 kHz with accuracy 16 bits, pre-emphasized with $1-0.97z^{-1}$. Each frame is multiplied by a Hamming windows with 25 msec and is computed every 10 msec.

The recognition system in our experiments is based on the HM-Net[10], which have 2000 states (4 mixtures/state) for a network. The acoustic models employ the 26 dimensions of features, which contains 12 MFCCs plus the logarithmic frame energy, as well as their the 1st order derivatives.

3.2 Experimental Results

Recognition experiments are carried out for noisy speech data in different SNR environments. Figure 7 shows that recognition performance of MSS is higher than of other noise reduction methods. When we compared MSS with SS and MWF, we can get better results of 24.4%, 23.8% at 0dB, 53.5%, 30.2% at 5dB, 40.2%, 9.7% at 10dB, 21.2%, 0.7% at 15dB.

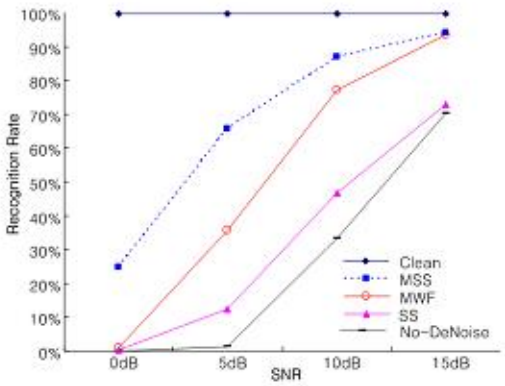

Fig. 7. Comparison of recognition rates for noise reduction methods

Figure 8 shows the recognition results for three combined methods i.e. MSS-NOVO, SS-NOVO, and MWF-NOVO at different SNRs. We could also find out that MSS-NOVO method showed the best performance by improving recognition rates of 52.8%, 34.6% in 0dB, 78.9%, 18.2% in 5dB, 75.0%, 2.7% in 10dB, 59.3%, and -1.2% in 15dB, comparing with SS-NOVO and MWF-NOVO, respectively.

From the above experiments, we find that SS-NOVO method shows the relative worse performance of recognition. This is because its noise reduction is imperfect and leads to

severe speech distortion. However, MSS-NOVO method shows much improvement of recognition performance in noisy environments, especially in low SNR environments such as at 0dB~10dB. The reason is the lower speech distortion happened from noise reduction with MSS method comparing with the SS method case. And this is also the benefit obtained from MSS method. Then the NOVO method is possible to shows its better effect for the test noisy speech which has lower distortion. Therefore, the combined MSS-NOVO method shows further higher recognition performance than other combined methods.

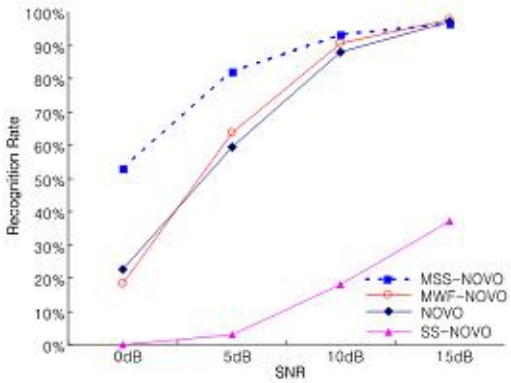

Fig. 8. Comparison of recognition rates for the combined methods

Ⅳ. Conclusions

In this paper, a MSS-NOVO method that starts with noise reduction with modified spectral subtraction (MSS) to enhance the input noisy speech, then the noise and voice composition (NOVO) method is applied for making noise adapted models by using the noise in the non-utterance regions of the enhanced noisy speech. From the tests, MSS-NOVO method showed average improvement of 66.5% and 13.6% compared with the existing SS-NOVO method and MWF-NOVO method, respectively. Especially, the proposed MSS-NOVO shows high effectiveness at low SNRs.

References

[1] Y. Ephraim, D. Malah, B. H. Juang, "On the Application of Hidden Markov Models for Enhancing Noisy Speech," Proc. IEEE Trans. Acoust., Speech Signal Processing, vol. 37, pp. 1846-1856, 1989.

[2] J. S. Lim, A. V. Oppenheim, "Enhanced and Bandwidth Compression of Noisy Speech," Proc. IEEE, vol. 67, no. 12, pp. 1586-1604, 1979.

[3] M. J. Gales, S. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," Proc. IEEE trans. on Speech and Audio Processing, vol. 4, pp.352-359, 1996.

[4] M. J. Gales, S. J. Young, "Robust Speech Recognition in Additive and Convolutional Noise using Parallel Model Combination," Proc. Computer Speech and Language, pp. 289-307, 1995.

[5] M. J. Gales, S. Young, "An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise," Proc. ICASSP, I-233-236, 1992.

[6] K. Satoshi, S. Sumitaka, "Robust Speech Recognition Based on HMM Composition and Modified Wiener Filter," Proc. ICSLP pp. 2045-2048, 2004.

[7] S. Sakauchi, A. Nakagawa, Y. Haneda, A. Kataoka, "Implementing and Evaluating of an Audio Teleconferencing Terminal with Noise and Echo Reduction," Proc. IWAENC, pp. 191-194, 2003.

[8] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," Proc. ICASSP, pp. 208-211, 1979.

[9] A. P. Varga, R. K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," Proc. ICASSP, pp. 845-848, 1990.

[10] Se-Jin Oh, Chul-Jun Hwang, Bum-Koog Kim, Hyun-Yeol Chung, Akinori Ito, "New State Clustering of Hidden Markov Network with Korean Phonological Rules for Speech Recognition," Proc. IEEE 4th workshop on Multimedia Signal Processing, pp. 39-44, 2001.

Biograph

Guanghu Shen

Aug. 2002: BS in Mathematics Education from Yanbian Univ., China
Aug. 2005: MS in Information Communication Engineering from Yeungnam Univ.
Sep. 2005~Present: PhD course in Information Communication Engineering from Yeungnam Univ.

Research interest: Speech Signal Processing, Robust Speech Recognition
Email : guanghosin@yu.ac.kr

Ho-Youl Jung

Aug. 1988: BS in Electronics Engineering from Ajou Univ.
Aug. 1990: MS in Electronics Engineering from Ajou Univ.
Apr. 1998: PhD in Electronics Engineering from the INSA, France
Mar. 1999~Present : Associate Professor, Electrical Engineering and Computer Science, Yeungnam Univ.

Research interest: Digital Watermarking, Speech Signal Processing, JPEG/JPEG-2000, Digital Image Processing
Email : hoyoul@yu.ac.kr

Hyun-Yeol Chung

Mar. 1975: BS in Electrical Engineering from Yeungnam Univ.
Aug. 1981: MS in Electrical Engineering from Yeungnam Univ.
Mar. 1989: PhD in Information Engineering from Tohoku Univ., Japan
Mar. 1990~present : Professor, School of Electrical Engineering and Computer Science, Yeungnam Univ.

Research interest: Speech Recognition, Speaker Recognition, Digital Signal Processing, Pattern Recognition,
Email : hychung@yu.ac.kr