스펙트럴 서브트렉션과 비동기 KLT 잡음 감소 기법의 조합에 의한 음성 인식 성능 개선

박 성 준
KT 서비스개발연구소

Improvement of the ASR Robustness using Combinations of Spectral Subtraction and KLT-based Adaptive Comb-filtering

Sung-Joon Park
Service Development Institute, KT, 17 Umyeon-dong, Seocho-gu, Seoul
E-mail : sjpak@kt.co.kr

Abstract

In this paper, the combinations of speech enhancement techniques are experimented. Specifically, the spectral subtraction, KLT based comb-filtering, and their combinations are applied to the Aurora2 database. The results show that recognition accuracy is improved when KLT based comb-filtering is applied after spectral subtraction.

I. Introduction

The ETSI STQ-AURORA DSR Working Group Aurora has initiated the standardization of front-end for Distributed Speech Recognition (DSR) where the speech analysis is done in the telecommunication terminal and the recognition at a central location in the telecom network [1]. The framework for the performance evaluation of speech recognition systems under noisy conditions was prepared [2] and various methods were proposed [3, 4]. Robustness can be achieved by an appropriate extraction of robust features in the front-end and/or by the adaptation of the references to the noise situation. In this paper, we describe spectral subtraction and Karhunen-Loève transform (KLT) based adaptive comb-filtering that both belong to speech enhancement approaches. Additionally, cepstral mean subtraction is incorporated.

This paper is organized as follows. In section 2, the noise reduction methods used in the experiments are described. In section 3, the experimental results are shown. Finally, the conclusions are given.

II. The enhancement methods

The Aurora2 front-end is a cepstral analysis scheme where 13 Mel frequency cepstral coefficients (MFCCs), including the coefficients of order 0, are determined for a speech frame of 25ms length [2]. The frame shift is 10 ms. In the experiments, the spectral subtraction and KLT-based comb-filtering use several parameters which do not necessarily coincide with those of the front-end.

For the convenience of the experiments, the spectral subtraction and KLT-based comb-filtering are implemented being separated from the Aurora2 front-end, which means that the outputs of each method are raw speech signals, and they are again the input of the Aurora2 front-end. In case of using spectral subtraction and KLT-based comb-filtering in sequence, for example, the output of the spectral subtraction is the input of KLT-based comb-filtering and the output of KLT-based

comb-filtering becomes the input of the Aurora2 front-end.

2.1 Spectral subtraction

Processing of the spectral subtraction is done on a frame-by-frame basis in frequency domain. It is mainly composed of two phases. The first phase is the calculation of the noise and the second is noise subtraction. The frame length and the frame shift are the same as in the Aurora2. Hanning window is applied. Let $S_y(w,t)$ denote be the short term fast Fourier transform of input signal y(n) at the t-th frame. The estimator of the clean speech is given by

$$| \hat{S}_x(w,t) | = \max(0, | S_x(w,t) | - \alpha | \hat{S}_n(w,t) |) \quad (1)$$

where $\hat{S}_n(w,t)$ is the estimated noise. Noise is estimated from the non-speech frames of input signal. If the current frame is determined as noise, noise is adapted by

$$| \hat{S}_n(w,t) | = \lambda | \hat{S}_n(w,t-1) | + (1-\lambda) | S_y(w,t) |. \quad (2)$$

If the current frame is speech, the previous noise is used. Detection of speech pauses is done simply by comparing the power of the current frame with a threshold that is the power of noise multiplied by α. If the power of the current frame is larger than the threshold, the current frame is considered as speech. The initial power of noise is calculated from the first segment of the input signal. The estimated clean speech is generated by the inverse FFT.

2.2 KLT-based comb-filtering

A signal subspace approach for speech enhancement was suggested by Ephraim and Van-Trees [5]. This method decomposes noisy speech into its components along the axes of a KLT-based vector space of the clean speech [6]. In this method, a block of data is used to estimate noisy speech covariance matrix. Then, an eigenvalue decomposition is applied to perform KLT. This approach requires repeated eigenvalue decomposition that consumes much time. In KLT-based comb-filtering, a vector of the input signal is composed of the samples separated with the pitch period that is determined at the current frame. Speech enhancement is performed by scaling each channel output of the quadrature comb-filter and reconstructing the speech signal from the scaled outputs [7]. This processing reduces the dimension of the covariance matrix of the input vector and the load of matrix computation.

In KLT-based comb-filtering, each sample of the clean speech signal $X(t)$ of the t-th frame is reconstructed from the estimation of (2T+1)-dimensional vectors $X_p(t,i)$ at the t-th frame, where

$$X_p(t,i) = (x((t-T-1)K+i), \ldots, x((t+T-1)K+i))^T \quad (3)$$

and i is from 1 to L which is the frame length. Speech samples and frames are shown in Fig.1.

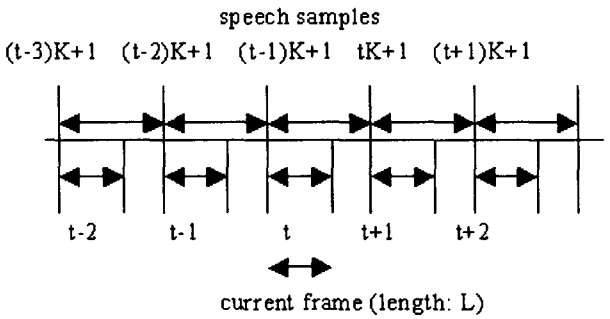

Fig. 1. Speech samples and frames

Assuming that noise is additive, we have the noisy input signal:

$$Y_p(t,i) = X_p(t,i) + N_p(t,i) \quad (4)$$

where $N_p(t,i)$ is (2T+1)-dimensional noise vector.

Now, let H be a (2T+1) x (2T+1) linear estimator of clean speech vector as follows:

$$\hat{X}_p = H Y_p . \quad (5)$$

The error signal obtained in this estimation is given by

$$r = \hat{X}_p - X_p = (H-I) X_p + H N_p = r_x + r_n \quad (6)$$

where $r_x = (H-I) X_p$ represents signal distortion and $r_n = H N_p$ represents residual noise [5]. Define the energies of signal distortion $\overline{\varepsilon_x^2}$ and residual noise $\overline{\varepsilon_n^2}$, respectively as follows:

$$\overline{\varepsilon_x^2} = trE\{ r_x r_x^T \} = trE\{(H-I) R_x (H-I)^T\} \quad (7)$$

and

$$\overline{\varepsilon_n^2} = trE\{ r_n r_n^T \} = trE\{H R_n H^T\} \quad (8)$$

where R_x and R_n are covariance matrices of clean signal and noise vector, respectively. Now, assuming R_x and R_n are provided, the linear estimator is obtained from

$$\min_H \overline{\varepsilon_x^2} \quad subject\ to: \frac{1}{K}\overline{\varepsilon_n^2} \le \sigma_n^2 \quad (9)$$

where σ^2 is a positive constant. H is a stationary feasible point if it satisfies the gradient equation of the Lagrangian

$$L_H(H,\mu) = \overline{\varepsilon_x^2} + \mu(\overline{\varepsilon_n^2} - K\sigma^2) \quad (10)$$

and

$$\mu(\overline{\varepsilon_n^2} - K\sigma^2) = 0 \quad for\ \mu \ge 0 \quad (11)$$

where μ is the Lagrange multiplier [5].

From $\nabla_H L(H,\mu) = 0$ and (7, 8), we obtain:

$$H = R_x (R_x + \mu R_n)^{-1} \quad (12)$$

Now, let the eigenvalue decomposition of R_x be defined as follows:

$$R_x = U \Lambda_x U^T \quad (13)$$

where Λ_x is a diagonal (2T+1) x (2T+1) matrix that contains clean signal covariance matrix eigenvalues and U contains its eigenvalues. U is called the inverse KLT and the unitary U^T is called KLT.

Substituting (13) in (12), we obtain:

$$H = U \Lambda_x (\Lambda_x + \mu U^T R_n U)^{-1} U^T . \quad (14)$$

Assuming that noise is white, $R_n \cong \lambda_n I$, where λ_n is the variance of white noise. From this assumption, we can rewrite the estimator as

$$H = UG U^T \quad (15)$$

where

$$G = diag(g_t(1), g_t(2), \ldots, g_t(2T+1)), \quad g_t(i) = \frac{\lambda_x^i}{\lambda_x^i + \mu \lambda_n} . \quad (16)$$

Hence, the signal $\hat{X}_p = H Y_p$ is obtained by applying KLT to the noisy signal, appropriately modifying the components of KLT $U^T Y_p$ by a gain function, and by inverse KLT of the modified components.

White noise was assumed in the derivation of the estimator of clean signal. In real environments, however, noise is not white and is difficult to estimate. Hence, we assume a more realistic approximation for noise model as follows:

$$\sigma_{n,t}^2 = \sqrt{\frac{\sum_{j=1}^{L} [\nu(m_1, j)]^2}{L} \frac{\sum_{j=1}^{L} [\nu(m_2, j)]^2}{L}} \quad (17)$$

where

$$m_1 = \arg \min_{1 \le m \le (2T+1)} (\sum_{j=1}^{L} [\nu(m, j)]^2) \quad (18)$$

$$m_2 = \arg \min_{1 \le m \le (2T+1), (m \ne m_1)} (\sum_{j=1}^{L} [\nu(m, j)]^2)$$

and

$$\nu(m, j) = m^{th} element\ of\ U^T Y_p(t, j). \quad (19)$$

Namely, the noise is calculated from the two low square averages of the coefficients that are obtained from $U^T Y_p(t, j)$. Before it is used for the gain function, σ_t^2 is adapted by

$$\sigma_t^2 = (1.0 - \lambda)\sigma_{n,t}^2 + \lambda \sigma_{t-1}^2 . \quad (20)$$

Using σ_t^2 ,the gain is calculated as follows:

$$G = diag(g_t(1), g_t(2), \ldots, g_t(2T+1)), \quad g_t(m) = \max(0, (1 - \mu \frac{\sigma_x^2}{\sum_{j=1}^{L} [\nu(m, j)]^2 / L})^r) \quad (21)$$

III. Experiments

The experiments used the multi-condition training HMMs trained in the manner described by HTK 20 mix configuration of the Aurora2 tasks.

First, spectral subtraction and KLT-based comb-filtering were each experimented. Table 1 shows the word error rates and improvements obtained by spectral subtraction. Table 2 is the results of KLT-based comb-filtering. Next, the combinations of two methods were experimented. Two kinds of combinations were experimented. One is the application of KLT-based comb-filtering after spectral subtraction. The other is spectral subtraction after KLT-based comb-filtering. Table 3 and 4 show the results. KLT-based comb-filtering after spectral subtraction shows better performance than others. In the final experiment, cepstral mean subtraction was incorporated (Table 5). It was applied after spectral subtraction and KLT-based comb-filtering.

Table 1. Aurora 2 reference word error rates, spectral subtraction word error rates and the related relative improvements

	Set A	Set B	Set C	Overall
Reference Word Err. Rate	11.93%	12.78%	15.44%	12.97%
Word Err. Rate	8.15%	9.70%	14.15%	9.97%
Rel. Imp.	38.17%	36.22%	18.00%	33.35%

Table 2. Results of KLT based comb-filtering

	Set A	Set B	Set C	Overall
Word Err. Rate	8.52%	11.59%	13.70%	10.79%
Rel. Imp.	31.34%	16.02%	18.70%	22.69%

Table 3. Results of KLT based comb-filtering after spectral subtraction

	Set A	Set B	Set C	Overall
Word Err. Rate	7.41%	8.46%	13.42%	9.03%
Rel. Imp.	42.21%	42.68%	19.29%	37.81%

Table 4. Results of spectral subtraction after KLT based comb-filtering.

	Set A	Set B	Set C	Overall
Word Err. Rate	8.69%	12.02%	14.00%	11.09%
Rel. Imp.	29.11%	9.37%	15.63%	18.52%

Table 5. Results of Spectral subtraction, KLT based comb-filtering, and cepstral mean subtraction.

	Set A	Set B	Set C	Overall
Word Err. Rate	6.63%	7.23%	8.05%	7.16%
Rel. Imp.	49.87%	52.89%	47.81%	50.67%

Ⅳ. Conclusions

In this paper, we applied the spectral subtraction and KLT-based comb-filtering together to the Aurora 2 database to improve the recognition performance. In the experiments, KLT-based comb-filtering after the spectral subtraction shows better performance than the spectral subtraction only, KLT-based comb-filtering only, and the spectral subtraction after KLT-based comb-filtering. When the cepstral mean subtraction is incorporated, performance is improved a litter more. In the spectral subtraction, the parameter values were experimentally chosen. In KLT-based comb-filtering, the parameter values were not optimized fully. Hence, performance improvement may be expected with the optimization of the parameter values.

참고문헌

[1] D. Pearce, "Enabling New Speech Driven Services for Mobile Device: An overview of the ETSI standards activities for Distributed Speech Recognition Front-ends," Applied Voice Input/Output Society Conference (AVOIS2000), San Jose, CA, May 2000

[2] H.G. Hirsch and D. Pearce, "The Aurora2 experimental framework for the performance evaluation of speech recognition systems under noisy conditions," ISCA ITRW ASR 2000, Sep. 2000

[3] Proceedings of Eurospeech 2001

[4] Proceedings of ICSLP 2002

[5] Y. Ephraim and H.L. Van-Trees, "A signal subspace approach for speech enhancement," IEEE Trans. on Speech and Audio Processing, vol. 3, pp. 251-266, July 1995

[6] L.L. Sharf, Statistical Signal Processing: Detection, Estimation, and Time Series Analysis, New York: Addison-Wesley, 1990

[7] M. Ikeda, K. Takeda, and F. Itakura, "Speech enhancement by quadratic comb-filtering," Technical report of IEICE, DSP96-70, SP96-45, pp. 23-30, Sep. 1996