音声言語情報処理　44－3
(2002. 12. 19)

スペクトルサブトラクションとピッチ同期KLTの複数の雑音抑圧手法を組合せたAurora2データベースに対する音声認識性能の改善

朴 性俊†*　池田 幹男‡　武田 一哉†　板倉 文忠†
†名古屋大学工学部 〒464-8603 名古屋市千種区不老町1
‡四日市大学経済学部 〒512-8512 三重県四日市市萱生町1200
*KT 〒137-792 韓国 ソウル 瑞草区 牛眠洞17
E-mail: †sjpark@itakura.nuee.nagoya-u.ac.jp, †{takeda, itakura}@nuee.nagoya-u.ac.jp,
‡ikeda@yokkaichi-u.ac.jp, *sjpak@kt.co.kr

あらまし　本研究では雑音抑圧手法の組合せによる音声認識性能の改善について検討した。具体的にスペクトルサブトラクション, ピッチ同期 KLT, およびその両者の組合せを Aurora2 のデータベースに適用し、音声認識結果を調査した。実験の結果、スペクトルサブトラクションの後でピッチ同期 KLT を適用した場合に最も音声認識性能が向上した。

キーワード　Aurora2, スペクトルサブトラクション, KLT, 櫛形フィルタ

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

Sung-Joon PARK†*　Mikio IKEDA‡　Kazuya TAKEDA†　and　Fumitada ITAKURA†
†Nagoya University, Furo-cho1, Chikusa-ku, Nagoya, Aichi 464-8603, Japan
‡Yokkaichi University, Kayou-cho1200, Yokkaichi-shi, Mie-ken 512-8512, Japan
* Service Development Institute, KT, 17 Umyeon-dong, Seocho-gu, Seoul, 137-792, Korea
E-mail: †sjpark@itakura.nuee.nagoya-u.ac.jp, †{takeda, itakura}@nuee.nagoya-u.ac.jp,
‡ikeda@yokkaichi-u.ac.jp, *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, and their results are compared. The results show that the performance is improved in the recognition accuracy when KLT based comb-filtering was applied after spectral subtraction.

Keyword　Aurora2, Spectral Subtraction, KLT, Adaptive Comb-filtering

1. Introduction

In most speech recognition systems, degradation of speech quality caused by undesirable background noise is common in adverse conditions. During the decades, the problem of removing uncorrelated noise components from noisy speech has been widely studied, and various approaches to the problem have been adopted.

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 first describe spectral subtraction and Karhunen-Loève transform (KLT) based adaptive comb-filtering that all belong to speech enhancement approaches. Additionally, cepstral mean subtraction is incorporated.

Spectral subtraction is a traditional method for removing stationary background noise in a single channel system. It attempts to estimate the short-time spectral magnitude of speech by subtracting a noise estimation from the noisy speech, and combine it with the phase of the noisy speech. It has gained popularity because it is relatively easy to understand and implement. The major drawback of this method is the characteristics of the residual noise called musical noise. It represents nonstationary residual noise due to the time varying filter applied to the noisy signal.

In KLT based comb-filtering, it is assumed that the input signal can be represented as the linear sum of basis vectors, which are assumed linearly independent. The basis vectors are obtained from a number of pitch periods used forward and backward in time [5]. The linear estimation is performed by modifying the KLT components by a gain function determined by the estimation criterion. The enhanced signal is obtained from the inverse KLT of the altered components. In this paper, the combination of spectral subtraction and KLT based comb-filtering is experimented assuming that one method cannot remove the noise completely and the remaining noise after the application of one method may be removed by the other method. In addition to the combination of the spectral subtraction and KLT based comb-filtering, cepstral mean subtraction is implemented to reduce the influence of a slowly changing acoustic environment of a telephony transmission channel.

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, we give our conclusions.

2. 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 our 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 the KLT based comb-filtering becomes the input of the Aurora2 front-end. Accordingly, the same Aurora2 front-end can be used except when the cepstral mean subtraction is incorporated into it. In the following subsections, each method is described.

2.1 Spectral subtraction

The 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

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

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

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

If the current frame is speech, the previous noise is used. The 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 [6]. This method decomposes noisy speech into its components along the axes of a KLT-based vector space of the clean speech [7]. 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 used in our experiments, 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 [5]. 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.

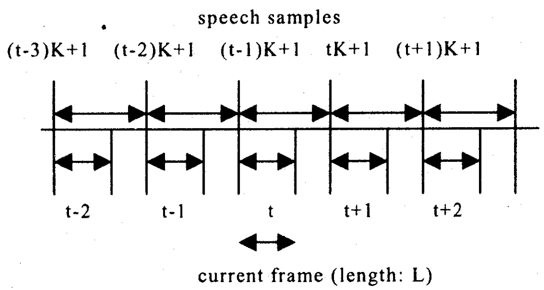

Fig. 1. Speech samples and frmaes

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:

$$\hat{X}_p = HY_p \quad (5)$$

The error signal obtained in this estimation is given by

$$\begin{aligned} r &= \hat{X}_p - X_p = (H-I)X_p + HN_p \\ &= r_x + r_n \end{aligned} \quad (6)$$

where $r_x = (H-I)X_p$ represents signal distortion and $r_n = HN_p$ represents residual noise [6]. Define the energies of signal distortion $\overline{\varepsilon_x^2}$ and residual noise $\overline{\varepsilon_n^2}$, respectively as follows:

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

and

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

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

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

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 [6].
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 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 eigenvectors. 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

$$H = UGU^T \quad (15)$$

where

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

Hence, the signal $\hat{X}_p = HY_p$ is obtained by applying the KLT to the noisy signal, appropriately modifying the components of the 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:

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

where

$$\begin{aligned} m_1 &= \arg \min_{1 \le m \le (2T+1)} \left(\sum_{j=1}^{L}[v(m,j)]^2\right) \\ m_2 &= \arg \min_{\substack{1 \le m \le (2T+1) \\ (m \ne m_1)}} \left(\sum_{j=1}^{L}[v(m,j)]^2\right) \end{aligned} \tag{18}$$

and

$$v(m,j): m^{th} \ \textit{element of} \ U^T Y_p(t,j). \tag{19}$$

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

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

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

$$\begin{aligned} G &= diag(g_t(1), g_t(2), \ldots, g_t(2T+1)), \\ g_t(m) &= \max\left(0, \left(1-\mu \frac{\sigma_t^2}{\sum_{j=1}^{L}[v(m,j)]^2 / L}\right)^{\gamma}\right) \end{aligned} \tag{21}$$

2.3 Cepstral mean subtraction

Cepstral mean subtraction is one of the earliest and the simplest methods used to remove channel distortion from signal. The long time average $\overline{C}(t,k)$ of the k-th cepstral coefficients $C(t,k)$ of the transmitted speech is computed by (22) and $\overline{C}(t,k)$ is subtracted from the observed cepstral vectors in order to remove the channel effect. By cepstral mean subtraction, the performance could be improved. However, N must be determined not to introduce too long additional delay.

$$\overline{C}(t,k) = \begin{cases} \dfrac{(t-1)\overline{C}(t-1,k) + C(t,k)}{t} & , if\ t > N \\ \dfrac{1}{N}\displaystyle\sum_{i}^{N} C(i,k) & , otherwise \end{cases} \tag{22}$$

3 Experiments

All speech data of Aurora2 database are derivatives of the TIdigits database. Three different sets of speech data are taken for the recognition. Set "a" consists of TIdigits test data downsampled at 8 kHz, filtered with a G712 characteristic and noise artificially added at several SNRs (20dB, 15dB, 10 dB, 0dB, -5dB, *clean*no noise added). The noises are the same as for the multi-condition training. Set "b" consists of TIdigits test data downsampled at 8 kHz, filtered with a G712 characteristic and noise artificially added. The noise types are different from those seen in the training data. Set "c" consists of TIdigits test data downsampled at 8 kHz, filtered with a MIRS characteristic and noise added. The noises are the same as used in test set "a" and "b". The intention of test set "c" is the consideration of a different frequency characteristic (MIRS instead of G712).

The experiments used HMMs trained in the manner described by HTK 20mix configuration of the Aurora2 tasks. Each of the 11 digits were modeled as strictly left-right whole-word HMMs, each with 16 states and each state consisted of 20 Gaussian mixtures. The silence had only 3 states, each with 36 Gaussians per state. The 1-state short pause model was tied to the second state of silence model. The HTK toolkit [8] was used for both training and testing.

First experiments applied spectral subtraction and KLT based comb-filtering separately. In spectral subtraction, the values of α and λ were experimentally chosen to be 1.2 and 0.95. The speech frame length and the frame shift are 25ms and 10 ms, which are the same values as in the Aurora 2 front-end. There are more parameters in KLT based comb-filtering. For T, λ, μ, γ, and L, we used 3, 0.995, 2.0, 1.2, and 64. 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. There are two kinds of combinations. One is the application of KLT based comb-filtering after spectral subtraction (Table 3). The other is spectral subtraction after KLT based comb-filtering (Table 4). KLT based comb-filtering after spectral subtraction shows better performance than others. In the final experiments, cepstral mean subtraction was incorporated. It was applied after spectral subtraction and KLT based comb-filtering. Table 5 and 6 show the results. In the experiments of Table 5, cepstral mean was calculated using the whole frames of the input speech. In the experiments of Table 6, only 30 frames were used in calculating the cepstral mean to reduce the additional delay.

4 Conclusions

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

References

[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] 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.

[6] 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.

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

[8] S. Young, D. Kershaw, J. Odell, "The HTK book," Entropic, 1999.

Aurora 2 Reference Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	11.93%	12.78%	15.44%	12.97%
Clean	41.26%	46.60%	34.00%	41.94%
Average	26.59%	29.69%	24.72%	27.46%

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	8.15%	9.70%	14.15%	9.97%
Clean	27.98%	29.98%	27.54%	28.69%
Average	18.07%	19.84%	20.85%	19.33%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	38.17%	36.22%	18.00%	33.35%
Clean	39.89%	43.72%	28.94%	39.23%
Average	39.03%	39.97%	23.47%	36.29%

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

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	8.52%	11.59%	13.70%	10.79%
Clean	35.21%	40.92%	24.89%	35.43%
Average	21.87%	26.25%	19.29%	23.11%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	31.34%	16.02%	18.70%	22.69%
Clean	23.17%	13.32%	39.00%	22.39%
Average	27.25%	14.67%	28.85%	22.54%

Table 2. Results of KLT based comb-filtering.

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	7.41%	8.46%	13.42%	9.03%
Clean	22.07%	24.74%	18.28%	22.38%
Average	14.74%	16.60%	15.85%	15.70%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	42.21%	42.68%	19.29%	37.81%
Clean	56.47%	55.44%	57.22%	56.21%
Average	49.34%	49.06%	38.25%	47.01%

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

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	8.69%	12.02%	14.00%	11.09%
Clean	37.06%	41.82%	26.86%	36.93%
Average	22.88%	26.92%	20.43%	24.01%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	29.11%	9.37%	15.63%	18.52%
Clean	12.74%	9.66%	29.63%	14.89%
Average	20.93%	9.51%	22.63%	16.70%

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

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	6.80%	7.79%	8.35%	7.50%
Clean	22.21%	22.70%	17.18%	21.40%
Average	14.50%	15.24%	12.77%	14.45%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	47.45%	48.24%	45.19%	47.32%
Clean	54.01%	60.36%	60.73%	57.89%
Average	50.73%	54.30%	52.96%	52.61%

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

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	7.48%	8.40%	9.40%	8.23%
Clean	23.72%	24.55%	18.49%	23.01%
Average	15.60%	16.48%	13.94%	15.62%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	42.55%	43.59%	38.11%	42.08%
Clean	50.67%	57.00%	57.44%	54.56%
Average	46.61%	50.30%	47.77%	48.32%

Table 6. Results of Spectral subtraction, KLT based comb-filtering, and local cepstral mean subtraction with N=30.