Integration of Noise Reduction Algorithms for Aurora2 Task

Takeshi Yamada[1], Jiro Okada[1], Kazuya Takeda[2], Norihide Kitaoka[3], Masakiyo Fujimoto[4], Shingo Kuroiwa[5], Kazumasa Yamamoto[6], Takanobu Nishiura[7], Mitsunori Mizumachi[8], and Satoshi Nakamura[8]

[1]University of Tsukuba, 1–1–1 Tennodai, Tsukuba, Ibaraki, 305–8573 Japan
[2]Nagoya University, [3]Toyohashi University of Technology, [4]Ryukoku University,
[5]University of Tokushima, [6]Shinshu University, [7]Wakayama University,
[8]ATR Spoken Language Translation Research Labs.

Abstract

To achieve high recognition performance for a wide variety of noise and for a wide range of signal-to-noise ratios, this paper presents the integration of four noise reduction algorithms: spectral subtraction with smoothing of time direction, temporal domain SVD-based speech enhancement, GMM-based speech estimation and KLT-based comb-filtering. Recognition results on the Aurora2 task show that the effectiveness of these algorithms and their combinations strongly depends on noise conditions, and excessive noise reduction tends to degrade recognition performance in multicondition training.

1. Introduction

In recent years, the performance of automatic speech recognition has been improved drastically by applying statistical approaches. However, most speech recognizers still have the serious problem that their recognition performance degrades in noisy environments. Noise robustness is an issue to be addressed, since it is inevitable to realize portability and flexibility of speech communication applications.

During the last decade, a number of noise reduction algorithms were proposed as a front-end of speech recognition. However, their effectiveness strongly depends on noise conditions. This means that no algorithm has achieved high recognition performance for a wide variety of noise and for a wide range of signal-to-noise ratios.

One way for solving this problem is to select a suitable algorithm corresponding to each noise condition, while another is to integrate multiple noise reduction algorithms with complementary characteristics. This paper presents the integration of four noise reduction algorithms: spectral subtraction with smoothing of time direction [1], temporal domain SVD-based speech enhancement [2], GMM-based speech estimation [2] and KLT-based comb-filtering [3], and shows recognition results on the Aurora2 task [4].

2. Noise reduction algorithms

2.1. Spectral subtraction with smoothing of time direction

The observation signal x is assumed to be the sum of speech signal s and noise n, namely, $x = s + n$. Spectral subtraction [5] in the power spectral domain is defined as below:

$$|\tilde{S}_i(t)|^2 \quad = \quad |X_i(t)|^2 - \alpha|\tilde{N}_i|^2, \tag{1}$$

where $|\tilde{S}_i(t)|^2$, $|X_i(t)|^2$ are the i-th components of the estimated power spectrum of speech and the power spectrum of observed signals at the time t, respectively, while $|\tilde{N}_i|^2$ is the i-th component of *a priori* estimated power spectrum of noise, and α is the overestimation factor. We can express $|X_i(t)|^2$ as:

$$\begin{aligned} &|X_i(t)|^2 \\ &\quad = \quad |S_i(t)|^2 + |N_i(t)|^2 + 2|S_i(t)||N_i(t)|\cos\theta_i(t), \end{aligned} \tag{2}$$

where $|S_i(t)|$ and $|N_i(t)|$ are the true values for speech and noise, and $\theta_i(t)$ is the phase difference between speech and noise. We suppose that the speech and the noise do not correlate to each other. The definition of (1) stands on the fact that the expectation value of $\cos\theta_i(t)$ in (2) equals zero. However, considering $\cos\theta_i(t)$ as a random variable ranging -1 to 1 and assuming that $\theta_i(t)$ distributes uniformly, the probability density function of $\phi = \cos\theta_i(t)$ becomes $f(\phi) = 1/(\pi\sqrt{1-\phi^2})$, a concave function with sole minimum at $\phi = 0$. Therefore, the term including $\cos\theta_i(t)$ in (2) cannot be removed, even if the noise power can be accurately estimated.

Here, we define the smoothing method as follows [1]:

$$\overline{|X_i(t)|^2} \quad = \quad \sum_\tau \beta_\tau |X_i(t-\tau)|^2, \tag{3}$$

where $\tau = 0, 1, \ldots, T-1$, $\sum_\tau \beta_\tau = 1$. Using (2), (3) becomes

$$\begin{aligned} \overline{|X_i(t)|^2} \quad &= \quad \sum_\tau \beta_\tau |S_i(t-\tau)|^2 + \sum_\tau \beta_\tau |N_i(t-\tau)|^2 \\ &\quad + 2\sum_\tau \beta_\tau |S_i(t-\tau)||N_i(t-\tau)|\cos\theta_i(t-\tau). \end{aligned} \tag{4}$$

Assuming speech and noise are stable for the period T, (4) becomes

$$\begin{aligned} \overline{|X_i(t)|^2} \quad &= \quad \overline{|S_i(t)|^2} + \overline{|N_i(t)|^2} \\ &\quad + 2\sum_\tau \beta_\tau |S_i(t-\tau)||N_i(t-\tau)|\cos\theta_i(t-\tau) \\ &\approx \quad |S_i(t)|^2 + |N_i(t)|^2 + 2|S_i(t)||N_i(t)|\phi, \end{aligned} \tag{5}$$

where $\phi = \sum_\tau \beta_\tau \cos\theta_i(t-\tau)$. Assuming phase differences between speech and noise of successive frames don't correlate to each other, the pdfs of ϕ has the peak at zero and the variance of this term becomes smaller than the original one. Thus, we can assume the third term of (4) is almost zero, and (4) becomes

$$\overline{|X_i(t)|^2} \quad \approx \quad |S_i(t)|^2 + |N_i(t)|^2. \tag{6}$$

Replacing $|X_i(t)|^2$ in (1) with $\overline{|X_i(t)|^2}$, (1) becomes

$$\begin{aligned} |\tilde{S}_i(t)|^2 \quad &= \quad \overline{|X_i(t)|^2} - \alpha|\tilde{N}_i|^2 \\ &\approx \quad |S_i(t)|^2 + |N_i(t)|^2 - \alpha|\tilde{N}_i|^2. \end{aligned} \tag{7}$$

Therefore, we can estimate the speech signal more accurately if we can estimate $|\tilde{N}_i|$ accurately. In this paper, we fixed $T = 3$, $\beta_\tau = 1/3$ and $\alpha = 1.8$ according to preliminary experiments.

2.2. Temporal domain SVD-based speech enhancement

By representing the signal amplitude $a(t)$ with an interval of length N and maximum delay $M-1$, the $N \times M$ dimensional Toeplitz matrix $\mathbf{A}$ is constructed [6].

$$\mathbf{A} = \begin{pmatrix} a(M-1) & \cdots & a(0) \\ \vdots & \ddots & \vdots \\ a(M+N-2) & \cdots & a(N-1) \end{pmatrix}, \quad (8)$$

where $M = 28$ and $N = 173$. At the i-th windowed short time frame, the observed noisy speech signal $x_i(t)$ is assumed to consist of a clean speech signal $s_i(t)$ and an additive noise $n_i(t)$ as follows.

$$x_i(t) = s_i(t) + n_i(t). \quad (9)$$

Therefore, (9) can be represented as (10) in terms of Toeplitz matrices of (8).

$$\mathbf{X}_i = \mathbf{S}_i + \mathbf{N}_i. \quad (10)$$

By applying SVD to $\mathbf{X}_i$, $\mathbf{X}_i$ is decomposed into three matrices and reconstructed as $\mathbf{X}_i = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^T$. As a result, the singular value matrix $\mathbf{\Sigma}_i = \mathrm{diag}(\sigma_m^{\mathbf{X}_i})$ is obtained. Here, the singular value $\sigma_m^{\mathbf{X}_i}$ can be represented as (11) under the assumption that $s_i(t)$ is un-correlate with $n_i(t)$:

$$\sigma_m^{\mathbf{X}_i} = \sigma_m^{\mathbf{S}_i} + \sigma_m^{\mathbf{N}_i}, \quad (11)$$

where $m = 0, \ldots, M-1$. In (11), if $n_i(t)$ is white noise, it can be assumed that the distribution of $\sigma_m^{\mathbf{N}_i}$ is uniform. Therefore, $\sigma_m^{\mathbf{S}_i}$ can be estimated as (12):

$$\hat{\sigma}_m^{\mathbf{S}_i} = \sigma_m^{\mathbf{X}_i} - \bar{\sigma}^{\mathbf{N}_i}, \quad (12)$$

where $\bar{\sigma}^{\mathbf{N}_i}$ is the averaged singular value of $\mathbf{N}_i$. By using estimated $\hat{\sigma}_m^{\mathbf{S}_i}$, the Toeplitz matrix $\hat{\mathbf{S}}_i$ is estimated as (13) [6].

$$\hat{\mathbf{S}}_i = \mathbf{U}_i \mathbf{W}_i \mathbf{\Sigma}_i \mathbf{V}_i^T, \quad (13)$$

$$\mathbf{W}_i = \mathrm{diag}\left(\frac{\sigma_m^{\mathbf{X}_i} - \bar{\sigma}^{\mathbf{N}_i}}{\sigma_m^{\mathbf{X}_i}} \right). \quad (14)$$

In (11), if it can be assumed that the singular values of clean speech $\sigma_m^{\mathbf{S}_i}$ vanish for enough large index of m ($m \geq R$), the remaining singular values can be handled as singular values of the noise [6].

$$\sigma_m^{\mathbf{N}_i} \simeq \sigma_m^{\mathbf{X}_i} \quad (m \geq R). \quad (15)$$

From this fact, the averaged singular value $\bar{\sigma}^{\mathbf{N}_i}$ is estimated as follows:

$$\bar{\sigma}^{\mathbf{N}_i} = \frac{1}{M-R} \sum_{m=R}^{M-1} \sigma_m^{\mathbf{X}_i}. \quad (16)$$

In each frame, the cut-off singular value index number R was set to the index number r, which makes the cumulative contribution rate ($ACR(r,i)$) shown by (16) more than 90%.

$$ACR(r,i) = \left(\sum_{m=0}^{r} \sigma_m^{\mathbf{X}_i} \right) / \left(\sum_{m'=0}^{M-1} \sigma_{m'}^{\mathbf{X}_i} \right) \times 100, \quad (17)$$

$$R = \underset{r}{\mathrm{argmin}} \left\{ ACR(r,i) \geq 90 \right\}. \quad (18)$$

2.3. GMM-based speech estimation

At the i-th frame, the logarithmic output energy of a Mel filter bank of observed noisy speech is represented as follows [7]:

$$\begin{aligned} \mathbf{X}(i) &= \log\left[\exp(\mathbf{S}(i)) + \exp(\mathbf{N}(i))\right] \\ &= \log\left[\exp(\mathbf{S}(i)) \left(1 + \frac{\exp(\mathbf{N}(i))}{\exp(\mathbf{S}(i))}\right)\right] \\ &= \mathbf{S}(i) + \log\left[1 + \exp(\mathbf{N}(i) - \mathbf{S}(i))\right] \\ &= \mathbf{S}(i) + \mathbf{G}(i), \quad (19) \\ \mathbf{G}(i) &= \log\left[1 + \exp(\mathbf{N}(i) - \mathbf{S}(i))\right], \quad (20) \end{aligned}$$

where $\mathbf{X}(i)$, $\mathbf{S}(i)$ and $\mathbf{N}(i)$ denote the vectors that have logarithmic output energy of a Mel filter bank of observed noisy speech, clean speech and noise, respectively. In (19), $\mathbf{G}(i)$ is equivalent to the mismatch factor between $\mathbf{X}(i)$ and $\mathbf{S}(i)$.

First, suppose that $\mathbf{S}(i)$ can be modeled by GMM with K mixture distributions,

$$p(\mathbf{S}(i)) = \sum_{k=1}^{K} P(k) \mathcal{N}(\mathbf{S}(i), \mu_{S,k}, \mathbf{\Sigma}_{S,k}), \quad (21)$$

where $p(\mathbf{S}(i))$ denotes the output probability of $\mathbf{S}(i)$, and $P(k)$, $\mu_{S,k}$ and $\mathbf{\Sigma}_{S,k}$ denotes the mixture weight, mean vector and diagonal covariance matrix of the k-th Gaussian distribution.

Next, suppose that $\mathbf{X}(i)$ can be modeled by GMM with K mixture distributions as well as $\mathbf{S}(i)$. When GMM of $\mathbf{S}(i)$ is given, GMM of $\mathbf{X}(i)$ can be obtained approximately, according to the following description. Let μ_N denote the mean vector of $\mathbf{N}(i)$, which is estimated using the first 10 frames of the observed noisy speech $\mathbf{X}(i)$. Then the mean vector of $\mathbf{X}(i)$ at the k-th Gaussian distribution is estimated as follows, based on (19, 20).

$$\begin{aligned} \mu_{X,k} &\simeq \mu_{S,k} + \log\left[1 + \exp(\mu_N - \mu_{S,k})\right] \\ &= \mu_{S,k} + \mu_{G,k}. \quad (22) \end{aligned}$$

On the other hand, the covariance matrix of $\mathbf{X}(i)$ is not modified as (23), because the estimation accuracy of the covariance matrix of $\mathbf{N}(i)$ estimated from the first 10 frames of the observed noisy speech $\mathbf{X}(i)$ is very poor.

$$\mathbf{\Sigma}_{X,k} \simeq \mathbf{\Sigma}_{S,k}. \quad (23)$$

In (22), $\mu_{G,k}$ corresponds to the mean vector of the mismatch factor at k-th Gaussian distribution. Therefore, the expectation of $\mathbf{G}(i)$ is estimated as weighted average of $\mu_{G,k}$ by using a posterior probability $P(k|\mathbf{X}(i))$ as follows [7]:

$$\hat{\mathbf{G}}(i) = \sum_{k=1}^{K} P(k|\mathbf{X}(i)) \mu_{G,k}, \quad (24)$$

$$P(k|\mathbf{X}(i)) = \frac{P(k) \mathcal{N}(\mathbf{X}(i), \mu_{X,k}, \mathbf{\Sigma}_{X,k})}{\sum_{k'=1}^{K} P(k') \mathcal{N}(\mathbf{X}(i), \mu_{X,k'}, \mathbf{\Sigma}_{X,k'})}. \quad (25)$$

From the procedure described above, the clean speech $\hat{\mathbf{S}}(i)$ is estimated by subtracting $\hat{\mathbf{G}}(i)$ from $\mathbf{X}(i)$ as (26) [7].

$$\hat{\mathbf{S}}(i) = \mathbf{X}(i) - \hat{\mathbf{G}}(i). \quad (26)$$

In the GMM based speech estimation, we have trained the GMM of clean speech with 512 mixture distributions in the logarithmic output energy of the Mel filter bank domain by using the training entire dataset of clean speech contained in the Aurora2 task.

2.4. KLT-based comb-filtering

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

$$S_p(t,i) = (s((t-T-1)K+i), \ldots, s((t+T-1)K+i))^T, \quad (27)$$

and i is from 1 to L which is the frame length. Assuming that noise is additive, we have the noisy input signal:

$$X_p(t,i) = S_p(t,i) + N_p(t,i), \quad (28)$$

where $N_p(t,i)$ is a $(2T+1)$ dimensional noise vector. Now, let H be a $(2T+1) \times (2T+1)$ linear estimator of the clean speech vector as follows:

$$\hat{S_p} = HX_p. \quad (29)$$

The error signal obtained in this estimation is given by

$$r = \hat{S_p} - S_p = (H-I)S_p + HN_p = r_s + r_n, \quad (30)$$

where $r_s = (H-I)S_p$ represents signal distortion and $r_n = HN_p$ represents residual noise [8]. We define the energies of signal distortion $\overline{\varepsilon_s^2}$ and residual noise $\overline{\varepsilon_n^2}$, respectively, as follows:

$$\overline{\varepsilon_s^2} = trE\{r_s r_s^T\} = tr\{(H-I)R_s(H-I)^T\}, \quad (31)$$
$$\overline{\varepsilon_n^2} = trE\{r_n r_n^T\} = tr\{HR_nH^T\}, \quad (32)$$

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

$$\min_H \overline{\varepsilon_s^2}, \quad \text{subject to} : \frac{1}{K}\overline{\varepsilon_n^2} \le \sigma_n^2, \quad (33)$$

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

$$L_H(H,\mu) = \overline{\varepsilon_s^2} + \mu(\overline{\varepsilon_n^2} - K\sigma^2), \quad (34)$$
$$\mu(\overline{\varepsilon_n^2} - K\sigma^2) = 0 \quad \text{for } \mu \ge 0, \quad (35)$$

where μ is the Lagrange multiplier [8]. From $\nabla_H L(H,\mu) = 0$ and (31, 32), we obtain:

$$H = R_s(R_s + \mu R_n)^{-1}. \quad (36)$$

Now, let eigenvalue decomposition of R_s be defined as follows:

$$R_s = U\Lambda_s U^T, \quad (37)$$

where Λ_s is a diagonal $(2T+1)\times(2T+1)$ 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 (37) into (36), we obtain

$$H = U\Lambda_s(\Lambda_s + \mu U^T R_n U)^{-1}U^T. \quad (38)$$

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

$$H = UGU^T, \quad (39)$$

Table 1: Combinations of the noise reduction algorithms.

2nd algorithm \ 1st algorithm	(S)	(T)	(G)	(K)
(S)	(S-S)	(T-S)	-	(K-S)
(T)	(S-T)	(T-T)	-	(K-T)
(G)	(S-G)	(T-G)	-	(K-G)
(K)	(S-K)	(T-K)	-	(K-K)

where

$$G = \mathrm{diag}(g_t(1), g_t(2), \cdots, g_t(2T+1)), \quad (40)$$
$$g_t(i) = \lambda_s^i/(\lambda_s^i + \mu\lambda_n). \quad (41)$$

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

3. Evaluation

3.1. Experimental conditions

The following noise reduction algorithms and their combinations as shown in Table 1 are prepared as front-ends of speech recognition.

- (S) Spectral subtraction with smoothing of the time direction
- (T) Temporal domain SVD-based speech enhancement
- (G) GMM-based speech estimation
- (K) KLT-based comb-filtering

In Table 1, for example, (S-T) means that the output signal of (S) is fed into (T). However, (G) is only the second algorithm, since its output is not waveform. These front-ends are used with cepstral mean normalization and without feature quantization and endpoints detection.

Speech recognition experiments are performed on the Aurora2 connected digit recognition task [4]. Digit HMMs are the standard complex back-end models with 16 states, and each state has a mixture of 20 diagonal Gaussians. All the training data are processed by each front-end before training. The feature vector has 39 components consist of 12 cepstral coefficients together with C0, their first and second derivatives.

3.2. Results

Table 2 shows the recognition results of multicondition training and clean condition training. In Table 2, the best recognition accuracy achieved by one of four algorithms and their combinations is presented in each cell together with the name of the front-end. These results show that the effectiveness of each front-end strongly depends on the noise conditions, such as the type of noise and the SNR. This fact implies that an appropriate selection of front-ends according to each noise condition leads to the improvement of recognition performance. Furthermore, the single algorithms are better than the combined ones in multicondition training, while the combined algorithms tend to achieve the best recognition accuracy in clean condition training. This means that excessive noise reduction results in the degradation of recognition performance in multicondition training.

Table 3 and 4 summarize the relative performance derived from the ETSI advanced front-end. In Table 4, the ETSI advanced front-end are used with the cepstral mean normalization and the feature vector corresponding to our front-ends.

Table 2: Recognition results of the multicondition training and the clean condition training.

Multicondition training, multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	(G) 99.51	(G) 99.43	(S) 99.46	(T) 99.63	99.51	(G) 99.51	(G) 99.43	(S) 99.46	(T) 99.63	99.51	(G) 99.51	(G) 99.49	99.50	99.51
20 dB	(S) 99.14	(S) 99.00	(S) 99.52	(S) 99.38	99.26	(S) 99.02	(S) 99.09	(S) 99.43	(S) 99.32	99.22	(G) 99.17	(G) 98.97	99.07	99.20
15 dB	(S) 98.74	(G) 98.64	(S-G) 99.22	(T) 98.33	98.73	(T) 98.62	(S) 98.61	(S) 98.66	(S-T) 98.55	98.61	(G) 98.80	(S) 98.40	98.60	98.66
10 dB	(K) 96.93	(T) 97.43	(S-G) 98.03	(S) 97.16	97.39	(T) 96.99	(K-G) 96.61	(T) 97.32	(S) 97.13	97.01	(G) 96.59	(K-G) 96.70	96.65	97.09
5 dB	(S) 93.18	(T) 92.44	(S-G) 94.57	(G) 91.89	93.02	(T) 92.85	(S) 91.60	(S) 92.99	(S) 92.29	92.43	(S) 92.94	(S) 90.63	91.79	92.54
0 dB	(S) 80.81	(S) 75.15	(S-G) 82.08	(G) 78.09	79.03	(T) 75.74	(S) 76.00	(S) 79.03	(S) 78.65	77.36	(S) 78.45	(S) 76.69	77.57	78.07
-5dB	(K-S) 49.59	(S) 41.32	(S-S) 51.77	(T-G) 47.18	47.47	(S) 40.80	(K-S) 45.28	(S) 46.82	(S-S) 47.73	45.16	(S) 48.14	(S) 40.05	44.10	45.87
Average	93.76	92.53	94.68	92.97	93.49	92.64	92.38	93.49	93.19	92.93	93.19	92.28	92.73	93.11

Clean training, multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	(G) 99.66	(G) 99.70	(G) 99.64	(G) 99.85	99.71	(G) 99.66	(G) 99.70	(G) 99.64	(G) 99.85	99.71	(G) 99.66	(G) 99.70	99.68	99.71
20 dB	(K-G) 98.77	(G) 99.15	(G) 99.34	(T-G) 98.98	99.06	(G) 99.20	(K-S) 98.94	(G) 99.19	(G) 99.17	99.13	(K-G) 98.93	(K-G) 98.88	98.91	99.06
15 dB	(K-G) 97.27	(G) 98.16	(T-G) 98.78	(K-G) 97.01	97.81	(G) 97.54	(K-G) 97.97	(G) 98.78	(K-G) 98.12	98.10	(K) 97.21	(K-G) 97.43	97.32	97.83
10 dB	(K-G) 93.34	(K-G) 95.44	(T-G) 96.24	(T-G) 93.12	94.54	(G) 94.41	(K-G) 93.92	(G) 96.21	(K-G) 94.82	94.84	(K-S) 92.29	(K-G) 92.84	92.57	94.26
5 dB	(K-G) 84.25	(K-G) 85.79	(T-G) 87.68	(T-G) 83.06	85.20	(K-G) 83.73	(K-G) 85.10	(T-G) 86.91	(K-G) 85.62	85.34	(T-K) 81.09	(K-G) 81.74	81.42	84.50
0 dB	(K-G) 62.27	(K-G) 61.76	(K-G) 67.01	(K-G) 62.82	63.47	(K-G) 61.68	(K-G) 62.12	(K-G) 66.18	(S-G) 63.19	63.29	(K-S) 56.92	(K-G) 58.68	57.80	62.26
-5dB	(K-G) 32.30	(K-G) 26.63	(S-K) 33.40	(T-G) 32.95	31.32	(K-G) 27.11	(K-G) 35.04	(K-G) 33.07	(S-G) 32.64	31.97	(K-S) 30.09	(K-S) 30.32	30.21	31.36
Average	87.18	88.06	89.81	87.00	88.01	87.31	87.61	89.45	88.18	88.14	85.29	85.91	85.60	87.58

Table 3: Relative performance derived from the ETSI advanced front-end.

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	-3.75%	-11.54%	-4.21%	-6.96%
Clean	-2.27%	-1.59%	-2.94%	-2.13%
Average	-3.01%	-6.57%	-3.57%	-4.55%

Table 4: Relative performance derived from the ETSI advanced front-end with the cepstral mean normalization and the feature vector corresponding to our front-ends.

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	-6.10%	-12.45%	-13.62%	-10.14%
Clean	19.89%	14.00%	4.34%	14.42%
Average	6.90%	0.77%	-4.64%	2.14%

4. Conclusions

This paper presented the integration of four noise reduction algorithms. The recognition results for the Aurora2 task showed that the effectiveness of each front-end strongly depends on the noise conditions, and that excessive noise reduction tends to degrade recognition performance in the multicondition training. As future work, we plan to develop an optimum selection of front-ends according to each noise condition.

5. Acknowledgements

This work was supported in part by the Telecommunications Advancement Organization of Japan and Strategic Information and Communications R&D Promotion Scheme of the Ministry of Public Management, Home Affairs, Posts and Telecommunications of Japan.

6. References

[1] N. Kitaoka, S. Nakagawa, "Evaluation of spectral subtraction with smoothing of time direction on the AURORA 2 task," Proc. ICSLP2002, pp. 465–468, 2002.

[2] M. Fujimoto, Y. Ariki, "Combination of temporal domain SVD based speech enhancement and GMM based speech estimation for ASR in noise – evaluation on the AURORA2 task –," Proc. Eurospeech2003, 2003.

[3] M. Ikeda, K. Takeda, F. Itakura, "Speech enhancement by quadratic comb-filtering," Technical Report of IEICE, SP96-45, pp. 23–30, 1996.

[4] H. G. Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," ISCA ITRW ASR2000, 2000.

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoustics, Speech and Signal Proc., Vol. 27, No. 2, pp. 113–120, 1979.

[6] C. Uhl, M. Lieb, "Experiments with an extend adaptive SVD enhancement scheme for speech recognition in noise," Proc. ICASSP2001, 2001.

[7] J. C. Segura, A. de la Torre, M. C. Benitez, A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. experiments using AURORA II database and tasks," Proc. Eurospeech2001, Vol. I, pp. 221–224, 2001.

[8] Y. Ephraim, H. L. Van-Trees, "A signal subspace approach for speech enhancement," IEEE Trans. Speech and Audio Proc., Vol. 3, No. 4, pp. 251–266, 1995.