2003－SLP－47　(18)
2003／7／19

雑音下音声認識のための複数の前処理手法の統合とそのAURORA-2Jによる評価

山田武志[1]　岡田治郎[1]　武田一哉[2]　北岡教英[3]　藤本雅清[4]
黒岩眞吾[5]　山本一公[6]　西浦敬信[7]　水町光徳[8]　中村　哲[8]

[1]筑波大学　[2]名古屋大学　[3]豊橋技術科学大学　[4]龍谷大学
[5]徳島大学　[6]信州大学　[7]和歌山大学　[8]ATR音声言語コミュニケーション研究所

概要　様々な雑音条件下でロバストな音声認識を実現するためには，複数の雑音抑圧手法の統合が有効であると考えられる．本稿では，4つの雑音抑圧手法（時間方向スムージングを用いたスペクトルサブトラクション法，時間領域SVDに基づく音声強調，GMMに基づく音声信号推定，ピッチ同期KLT）とそれらの組合せの有効性を，AURORA-2Jを用いて調べた．その結果，雑音条件によって最適な手法・組合せは異なっており，雑音条件に適したものを適宜選択することにより，認識性能を大幅に改善できることが明らかとなった．また，Multicondition trainingの場合は，雑音抑圧量を増やしても，必ずしも認識性能の改善につながらないことが分かった．

Integration of Noise Reduction Algorithms for AURORA-2J 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, [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 AURORA-2J 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　はじめに

近年，統計的手法の導入により音声認識の性能は飛躍的に向上し，音声ワープロなどのアプリケーションが市販されるまでに至っている．しかし，音声認識の利用が想定される環境には周囲雑音が存在することが多く，特にマイクロホンから離れて発話するような状況下では，認識性能が著しく低下するという問題がある．音声インターフェイスの利便性の向上のためにも，周囲雑音に対してロバストな音声認識の実現が急務である．

従来，音声認識の前処理として，様々な雑音抑圧手法が提案されている．一般に，これらの手法の有効性は雑音条件に強く依存しており，多種多様な雑音を広範囲のSNRに渡って効果的に抑圧する手法は存在しないことが知られている．このような中，様々な雑音条件下でロバストな音声認識を実現するため

には，雑音条件に適した手法を適宜選択する，複数の雑音抑圧手法を統合する，ことなどが有効であると考えられる．

本稿では，4つの雑音抑圧手法（時間方向スムージングを用いたスペクトルサブトラクション法[1]，時間領域SVDに基づく音声強調[2]，GMMに基づく音声信号推定[2]，ピッチ同期KLT[3]）とそれらの組合せの有効性を，AURORA-2J[4]を用いて調べる．

2 雑音抑圧手法

2.1 時間方向スムージングを用いたスペクトルサブトラクション法

スペクトルサブトラクション法[5]をパワースペクトル領域における減算として定義する．音響分析における第tフレームの離散フーリエ変換の第i成分について，観測信号のパワーを$|X_i(t)|^2$，事前に推定した雑音を$|\tilde{N}_i|^2$とする．このとき，雑音除去された音声のパワー$|\tilde{S}_i(t)|^2$は，

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

として推定される．ここで，$|X_i(t)|^2$は，真の音声$S_i(t)$，真の雑音$N_i(t)$，これらの位相差$\theta_i(t)$を用いて，

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

と表現できる．式(2)の最後の項が相互相関項である．パワースペクトル領域でのスペクトルサブトラクションは，$\cos\theta_i(t)$の期待値が0となることに基づいている．しかし，特定のフレームに関する場合，これが0付近の値をとる確率は小さく，$\cos\theta_i(t)$が0であるという仮定は不適切である．

ここで，スペクトルの時間方向のスムージング

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

を考える．ただし，$\tau=0,1,\ldots,T-1$，$\sum_\tau \beta_\tau = 1$である．右辺に式(2)を代入すると，

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

となる．Tフレームの間，音声，及び雑音がほぼ定常であると仮定すると，

$$\sum_\tau \beta_\tau |S_i(t-\tau)|^2 \approx |S_i(t)|^2 \quad (5)$$

$$\sum_\tau \beta_\tau |N_i(t-\tau)|^2 \approx |N_i(t)|^2 \quad (6)$$

$$\begin{aligned} &\sum_\tau \beta_\tau |S_i(t-\tau)||N_i(t-\tau)|\cos\theta_i(t-\tau) \\ &\quad \approx |S_i(t)||N_i(t)|\sum_\tau \beta_\tau \cos\theta_i(t-\tau) \end{aligned} \quad (7)$$

と考えられ，

$$\begin{aligned} \overline{|X_i(t)|^2} &\approx |S_i(t)|^2+|N_i(t)|^2 \\ &\quad +2|S_i(t)||N_i(t)|\sum_\tau \beta_\tau \cos\theta_i(t-\tau) \end{aligned} \quad (8)$$

となる．ここで，隣接するフレーム間で音声と雑音の位相差が無相関であると仮定し，$\beta_\tau = 1/T$とすると，$\phi = \sum_\tau \beta_\tau \cos\theta_i(t-\tau)$の確率分布の標準偏差は$1/\sqrt{2T}$となり，$T$が大きくなるほど$\phi$が0に近い値をとる確率が高くなる．つまり，$T$が大きければ，式(4)の第3項が0に近い値と仮定できる．これらから，式(4)は，

$$\overline{|X_i(t)|^2} \approx |S_i(t)|^2+|N_i(t)|^2 \quad (9)$$

となり，式(1)の$|X_i(t)|^2$を$\overline{|X_i(t)|^2}$に置き換えることにより，

$$\begin{aligned} |\tilde{S}_i(t)|^2 &= \overline{|X_i(t)|^2} - \alpha|\tilde{N}_i|^2 \\ &\approx |S_i(t)|^2+|N_i(t)|^2-\alpha|\tilde{N}_i|^2 \end{aligned} \quad (10)$$

となることから，雑音の推定がより正確になる．

2.2 時間領域SVDに基づく音声強調

信号$a(t)$を間隔N，及び最大$M-1$の遅延を用いて表すことにより，$N\times M$次元のToeplitz行列$\mathbf{A}$を以下のように構成することができる[6]．

$$\mathbf{A} = \begin{pmatrix} a(M-1) & \cdots & a(0) \\ \vdots & \ddots & \vdots \\ a(M+N-2) & \cdots & a(N-1) \end{pmatrix} \quad (11)$$

次に，i番目の短時間フレームにおいて，雑音重畳音声$x_i(t)$は，クリーン音声$s_i(t)$と雑音$n_i(t)$により以下のように表現できる．

$$x_i(t) = s_i(t)+n_i(t) \quad (12)$$

このとき，式(12)は，式(11)のToeplitz行列を用いて式(13)のように表すことができる．

$$\mathbf{X}_i \quad = \quad \mathbf{S}_i + \mathbf{N}_i \tag{13}$$

$\mathbf{X}_i$に対してSVDを適用することにより，$\mathbf{X}_i$は$\mathbf{X}_i = \mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i^T$というように3つの行列に分解され，結果として特異値行列$\mathbf{\Sigma}_i = diag\left(\sigma_m^{\mathbf{X}_i}\right)$が得られる（$m = 0, \ldots, M-1$）．ここで，特異値$\sigma_m^{\mathbf{X}_i}$は，$s_i(t)$と$n_i(t)$が無相関と見なすことにより，式(14)のように表される．

$$\sigma_m^{\mathbf{X}_i} \quad = \quad \sigma_m^{\mathbf{S}_i} + \sigma_m^{\mathbf{N}_i} \tag{14}$$

式(14)において，$n_i(t)$が白色性の雑音であれば，$\sigma_m^{\mathbf{N}_i}$は全ての特異値$\sigma_m^{\mathbf{X}_i}$に一様に分布すると仮定できる．従って，$\sigma_m^{\mathbf{S}_i}$は式(15)のように推定できる．

$$\hat{\sigma}_m^{\mathbf{S}_i} \quad = \quad \sigma_m^{\mathbf{X}_i} - \bar{\sigma}^{\mathbf{N}_i} \tag{15}$$

ここで，$\bar{\sigma}^{\mathbf{N}_i}$は$\mathbf{N}_i$の特異値の平均値である．推定された$\hat{\sigma}_m^{\mathbf{S}_i}$を用いて，Toeplitz行列$\hat{\mathbf{S}}_i$は式(16)のように推定される[6].

$$\hat{\mathbf{S}}_i \quad = \quad \mathbf{U}_i\mathbf{W}_i\mathbf{\Sigma}_i\mathbf{V}_i^T \tag{16}$$

$$\mathbf{W}_i \quad = \quad diag\left(\frac{\sigma_m^{\mathbf{X}_i} - \bar{\sigma}^{\mathbf{N}_i}}{\sigma_m^{\mathbf{X}_i}}\right) \tag{17}$$

式(14)において，音声成分の特異値$\sigma_m^{\mathbf{S}_i}$が次元R以上の高次元で消失すると仮定すると，高次元の特異値は雑音成分の特異値に相当すると仮定できる[6].

$$\sigma_m^{\mathbf{N}_i} \quad \simeq \quad \sigma_m^{\mathbf{X}_i} \quad (m \geq R) \tag{18}$$

このことより，雑音の特異値の平均値$\bar{\sigma}^{\mathbf{N}_i}$は，以下のように推定できる．

$$\bar{\sigma}^{\mathbf{N}_i} \quad = \quad \frac{1}{M-R}\sum_{m=R}^{M-1}\sigma_m^{\mathbf{X}_i} \tag{19}$$

なお，式(11)のToeplitz行列の次元を決定するパラメータには，$M = 28$及び$N = 173$を与えた．また，特異値の打ち切り次元Rには，式(20)に示す特異値の累積寄与率$ACR(r,i)$を90%以上にする最小の値rを設定した．

$$ACR(r,i) \quad = \quad \frac{\sum_{m=0}^{r}\sigma_m^{\mathbf{X}_i}}{\sum_{m'=0}^{M-1}\sigma_{m'}^{\mathbf{X}_i}} \times 100 \tag{20}$$

$$R \quad = \quad \underset{r}{\mathrm{argmin}}\left\{ACR(r,i) > 90\right\} \tag{21}$$

2.3 GMMに基づく音声信号推定

2.3.1 信号モデル

第i番目の短時間フレームにおいて，雑音重畳音声，音声，雑音のメルフィルタバンク出力の対数値を要素に持つ，J次元ベクトルをそれぞれ$\mathbf{X}(i)$，$\mathbf{S}(i)$，$\mathbf{N}(i)$とすると，各ベクトルの要素間の独立性を仮定することにより，$\mathbf{X}(i)$は以下のように表される．

$$\begin{aligned}\mathbf{X}(i) &= \log\left[\exp(\mathbf{S}(i)) + \exp(\mathbf{N}(i))\right] \\ &= \mathbf{S}(i) + \log\left[1 + \exp(\mathbf{N}(i) - \mathbf{S}(i))\right] \\ &= \mathbf{S}(i) + \mathbf{G}(i) \qquad (22)\end{aligned}$$

$$\mathbf{G}(i) \quad = \quad \log\left[1 + \exp(\mathbf{N}(i) - \mathbf{S}(i))\right] \tag{23}$$

式(22)において，$\mathbf{G}(i)$は，$\mathbf{X}(i)$における雑音成分，すなわち$\mathbf{S}(i)$とのミスマッチ成分に相当する．

2.3.2 GMMを用いた$\mathbf{G}(i)$の推定

式(24)に示す$\mathbf{S}(i)$のK混合分布GMMを用いて，$\mathbf{G}(i)$を推定する．

$$p(\mathbf{S}(i)) \quad = \quad \sum_{k=1}^{K} P(k)\mathcal{N}(\mathbf{S}(i), \mu_{S,k}, \mathbf{\Sigma}_{S,k}) \tag{24}$$

上式において，$p(\mathbf{S}(i))$は$\mathbf{S}(i)$の出力確率である．また，$P(k)$，$\mu_{S,k}$，$\mathbf{\Sigma}_{S,k}$は，それぞれ要素分布kにおける混合重み，平均ベクトル，対角分散行列である．次に，式(24)のようなGMMが与えられたときに，$\mathbf{X}(i)$をLog-add compensation法[7]を用いて，$\mathbf{S}(i)$と同じようにK混合分布のGMMを用いてモデル化することを考える．ここで，雑音重畳音声の開始10フレームを雑音のみが存在する区間であるとして推定した，$\mathbf{N}(i)$の平均ベクトルをμ_Nとすると，$\mathbf{X}(i)$のGMMの要素分布kにおける平均ベクトル$\mu_{X,k}$は，式(22)を用いて，

$$\begin{aligned}\mu_{X,k} &\simeq \mu_{S,k} + \log\left[1 + \exp(\mu_N - \mu_{S,k})\right] \\ &= \mu_{S,k} + \mu_{G,k} \qquad (25)\end{aligned}$$

と近似できる．また，対角分散行列$\mathbf{\Sigma}_{X,k}$は，

$$\mathbf{\Sigma}_{X,k} \quad \simeq \quad \mathbf{\Sigma}_{S,k} \tag{26}$$

として近似する．式(25)において，$\mu_{G,k}$は要素分布kにおける雑音成分$\mathbf{G}(i)$の平均ベクトルに相当し，$\mu_{G,k}$を式(27)のように，$\mathbf{X}(i)$の事後確率$P(k|\mathbf{X}(i))$を用いて重み付け平均することにより，フレームiにおける$\mathbf{G}(i)$の推定値$\hat{\mathbf{G}}(i)$を推定する．

$$\hat{\mathbf{G}}(i) \quad = \quad \sum_{k=1}^{K} P(k|\mathbf{X}(i))\mu_{G,k} \tag{27}$$

$$P(k|\mathbf{X}(i)) = \frac{P(k)\mathcal{N}(\mathbf{X}(i),\mu_{X,k},\mathbf{\Sigma}_{X,k})}{\sum_{k'=1}^{K} P(k')\mathcal{N}(\mathbf{X}(i),\mu_{X,k'},\mathbf{\Sigma}_{X,k'})} \quad (28)$$

2.3.3 時間領域フィルタリング

式(22)において，雑音成分$\mathbf{G}(i)$は，対数メルスペクトル（メルフィルタバンク出力値）領域にて定式化されていたが，線形メルスペクトル領域では，式(22)を用いて以下のように表現できる．

$$\begin{aligned}\exp(\mathbf{X}(i)) &= \exp(\mathbf{S}(i))\exp(\mathbf{G}(i)) \\ &= \exp(\mathbf{S}(i))\frac{\exp(\mathbf{S}(i))+\exp(\mathbf{N}(i))}{\exp(\mathbf{S}(i))} \quad (29)\end{aligned}$$

上式を変形すると，

$$\begin{aligned}\exp(\mathbf{S}(i)) &= \exp(\mathbf{X}(i))\frac{\exp(\mathbf{S}(i))}{\exp(\mathbf{S}(i))+\exp(\mathbf{N}(i))} \\ &= \exp(\mathbf{X}(i))\exp(-\mathbf{G}(i)) \quad (30)\end{aligned}$$

が得られる．式(29)，(30)より，線形メルスペクトル領域で表現された雑音成分$\exp(\mathbf{G}(i))$の逆数は，ウィナーフィルターのフィルタゲインに相当することが分かる．よって，推定されたウィナーフィルターゲイン$\exp(-\hat{\mathbf{G}}(i))$をMel-warped IDCT[8]を用いてインパルス応答に変換し，雑音重畳音声に各フレーム毎に畳み込むことによって，クリーンな音声波形信号を得ることができる．

2.4 ピッチ同期KLT

ピッチ同期KLTでは，クリーン音声$s(t,i)$のt番目のフレームの各サンプルは，t番目のフレームにおける$(2T+1)$次元のベクトル$S_p(t,i)$の推定値から再構成される．ここで，

$$\begin{aligned}S_p(t,i) &= (s((t-T-1)K+i), \\ &\quad \ldots, s((t+T-1)K+i))^T \quad (31)\end{aligned}$$

であり，$i=1,\ldots,L$（L：フレーム長）である．雑音が加法性であると仮定すると，入力信号は次式で表される．

$$X_p(t,i) = S_p(t,i) + N_p(t,i) \quad (32)$$

ここで，$N_p(t,i)$は$(2T+1)$次元のベクトルである．入力信号からクリーン音声を推定する$(2T+1)\times(2T+1)$次元の行列Hを考える．

$$\hat{S_p} = HX_p \quad (33)$$

このとき，推定誤差は次式で表される．

$$r = \hat{S_p} - S_p = (H-I)S_p + HN_p = r_s + r_n \quad (34)$$

ここで，$r_s=(H-I)S_p$は音声の歪み，$r_n=HN_p$は残留雑音である[9]．音声の歪みのエネルギー$\overline{\varepsilon_s^2}$と残留雑音のエネルギー$\overline{\varepsilon_n^2}$を，次のように定義する．

$$\overline{\varepsilon_s^2} = trE\{r_s r_s^T\} = tr\{(H-I)R_s(H-I)^T\} \quad (35)$$

$$\overline{\varepsilon_n^2} = trE\{r_n r_n^T\} = tr\{HR_nH^T\} \quad (36)$$

ここで，R_sとR_nは，クリーン音声と雑音の共分散行列である．R_sとR_nが既知のとき，Hは次式により求められる．

$$\min_H \overline{\varepsilon_s^2}, \quad \text{subject to}: \frac{1}{K}\overline{\varepsilon_n^2} \le \sigma_n^2 \quad (37)$$

ここで，σ_n^2は正の定数である．Hをラグランジュの未定乗数法により求める．

$$L_H(H,\mu) = \overline{\varepsilon_s^2} + \mu(\overline{\varepsilon_n^2} - K\sigma_n^2) \quad (38)$$

$$\mu(\overline{\varepsilon_n^2} - K\sigma_n^2) = 0 \quad \text{for } \mu \ge 0 \quad (39)$$

ここで，μはラグランジュ乗数である．$\nabla_H L(H,\mu)=0$，式(35)，式(36)より，次式を得る．

$$H = R_s(R_s+\mu R_n)^{-1} \quad (40)$$

R_sを次式のように固有値分解する．

$$R_s = U\Lambda_s U^T \quad (41)$$

ここで，Λ_sは$(2T+1)\times(2T+1)$次元の対角行列であり，その対角成分はクリーン音声の共分散行列の固有値である．また，Uはその固有ベクトルである．Uは逆KLT，U^TはKLTと呼ばれる．式(40)に式(41)を代入し，

$$H = U\Lambda_s(\Lambda_s+\mu U^T R_n U)^{-1}U^T \quad (42)$$

を得る．雑音が白色，すなわち$R_n \simeq \lambda_n I$であると仮定する．ここで，λ_nは白色雑音の分散である．このとき，Hは次式で表される．

$$H = UGU^T \quad (43)$$

ここで，

$$G = \mathrm{diag}(g_t(1), g_t(2), \cdots, g_t(2T+1)) \quad (44)$$

$$g_t(i) = \lambda_s^i/(\lambda_s^i+\mu\lambda_n) \quad (45)$$

以上から，クリーン音声$\hat{S_p}=HX_p$を得る．

表 1: 雑音抑圧手法の組合せ

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

3 AURORA-2J を用いた認識実験

3.1 実験条件

雑音下連続数字認識タスクである AURORA-2J[4] を用いて，以下の 4 つの雑音抑圧手法とそれらの組合せの有効性を調べる．

(S) 時間方向スムージングを用いたスペクトルサブトラクション法

(T) 時間領域 SVD に基づく音声強調

(G) GMM に基づく音声信号推定

(K) ピッチ同期 KLT

雑音抑圧手法の組合せを表 1に示す．ここで，表中の (S-T) は，まず入力信号を (S) で処理し，次にその出力信号を (T) で処理することを意味している．他も同様である．

学習・認識においては，AURORA-2J の標準スクリプトを用いている．ここで，各数字モデルの状態数は 16，各状態の混合数は 20 である．また，特徴量は 39 次元であり，MFCC 係数（12 次元）と対数パワー，及びその Δ 係数・ΔΔ 係数からなる．標準スクリプトからの唯一の変更点は，CMN を行っていることである．なお，本実験では，学習データに対しても認識時と同様の雑音抑圧処理を行っている．

3.2 実験結果

各雑音条件において最適な手法を選択した場合の，Clean training と Multicondition training における単語正解精度を表 2に示す．ここで，表中の各セルには，単語正解精度と手法の名前を併記している．ただし，複数の手法が該当する場合は，その学習データセットにおける平均的な認識性能が高かったものを記載している．

表 2より，雑音の種類や SNR によって，最適な手法・組合せは異なっていることが分かる．また，Clean training の場合は，手法単体よりも組合せの方が単語正解精度が高いことを示している．一方，Multicondition training の場合は，手法単体の方が有利なケースが多くあり，雑音抑圧量を増やしても，必ずしも認識性能の改善につながらないことが分かる．

表 3: Relative performance

Relative performance				
	A	B	C	Overall
Clean Training	72.88%	71.13%	61.89%	70.11%
Multicondition training	15.77%	40.24%	34.94%	33.28%
Average	44.33%	55.68%	48.42%	51.69%

表 4: (G-K) の Relative performance

Relative performance				
	A	B	C	Overall
Clean Training	66.93%	66.09%	57.31%	64.79%
Multicondition training	-30.84%	22.43%	14.19%	7.94%
Average	18.05%	44.26%	35.75%	36.36%

表 5: (S) の Relative performance

Relative performance				
	A	B	C	Overall
Clean Training	18.02%	24.43%	9.59%	19.12%
Multicondition training	6.54%	35.09%	32.35%	27.66%
Average	12.28%	29.76%	20.97%	23.39%

次に，表 2の結果から求めた，ベースラインに対する Relative performance を表 3に示す．また，(G-K) と (S) のベースラインに対する Relative performance を表 4と表 5に示す．ここで，(G-K) と (S) は，各々 Clean training，Multicondition training の場合に，ベースラインに対する平均的な Relative performance が最も高かったものである．これらの結果より，各雑音条件において最適な手法を選択することにより，Clean training の場合は 5.32%（70.11% − 64.79%），Multicondition training の場合は 5.62%（33.28% − 27.66%）の性能改善を達成できることが分かる．

4 おわりに

本稿では，様々な雑音条件下でロバストな音声認識を実現するために，4 つの雑音抑圧手法とそれらの組合せの有効性を，AURORA-2J を用いて調べた．その結果，雑音条件によって最適な手法・組合せは異なっており，雑音条件に適したものを適宜選択することにより，認識性能を大幅に改善できることが明らかとなった．また，Multicondition training の場合は，雑音抑圧量を増やしても，必ずしも認識性能の改善につながらないことが分かった．今後は，各雑音条件における最適な手法の選択法について検討を進める予定である．

表 2: 単語正解精度（%Acc）

Clean Training (%Acc)

	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	(G-G) 99.97	(T-G) 99.97	(K-G) 100.00	(G-G) 99.94	99.97	(G-G) 99.97	(T-G) 99.97	(K-G) 100.00	(G-G) 99.94	99.97	(G-G) 99.97	(G) 99.97	99.97	99.97
20 dB	(T-K) 99.45	(G-K) 99.79	(K-G) 99.70	(G-K) 99.81	99.69	(G-K) 99.45	(K-G) 99.37	(G-K) 99.76	(G-K) 99.29	99.47	(G-K) 99.08	(G-K) 99.00	99.04	99.47
15 dB	(G-K) 98.46	(G-K) 98.70	(T-G) 99.02	(G-K) 98.12	98.58	(G-K) 98.83	(G-K) 98.00	(G-K) 98.87	(G-K) 97.53	98.31	(G-K) 97.38	(K-G) 97.79	97.59	98.27
10 dB	(G-K) 94.23	(G-K) 94.53	(K-G) 96.09	(G-K) 95.34	95.05	(G-K) 94.35	(K-G) 93.32	(G-K) 94.66	(S-G) 90.13	93.12	(T-K) 91.00	(K-G) 91.20	91.10	93.49
5 dB	(K-K) 80.66	(G-K) 77.45	(S-G) 84.37	(G-K) 85.28	81.94	(G-K) 77.68	(K-G) 78.33	(G-K) 78.80	(S-G) 80.16	78.74	(T-K) 72.24	(K-G) 74.73	73.49	78.97
0 dB	(K-K) 55.85	(G-K) 41.08	(S-G) 57.92	(T-G) 54.03	52.22	(G-K) 41.91	(K-G) 50.76	(S-G) 49.54	(S-G) 55.75	49.49	(K-K) 41.57	(T-G) 45.10	43.34	49.35
-5 dB	(K-K) 26.28	(G-K) 13.63	(S-G) 22.40	(T-G) 21.94	21.06	(G-K) 9.89	(K-K) 20.83	(G-S) 18.70	(S-G) 20.83	17.56	(K-K) 19.83	(T-G) 18.77	19.30	19.31
Average	85.73	82.31	87.42	86.52	85.49	82.44	83.96	84.33	84.57	83.82	80.25	81.56	80.91	83.91

Multicondition Training (%Acc)

	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	(S) 99.88	(S-G) 99.88	(T) 99.91	(S-G) 99.88	99.89	(S) 99.88	(S-G) 99.88	(T) 99.91	(S-G) 99.88	99.89	(S) 99.88	(G-G) 99.91	99.90	99.89
20 dB	(G-G) 99.82	(G-G) 99.61	(S-G) 99.88	(S) 99.60	99.73	(S-T) 99.48	(S-G) 99.55	(S-T) 99.46	(S-T) 99.66	99.54	(S) 99.72	(S-G) 99.55	99.64	99.63
15 dB	(T) 99.45	(S-G) 99.37	(S) 99.64	(S) 99.04	99.38	(T-S) 98.50	(G-G) 98.91	(S) 98.45	(T-G) 98.43	98.57	(G) 99.36	(G-S) 98.97	99.17	99.01
10 dB	(G-K) 98.25	(S) 97.76	(S-G) 98.51	(G-G) 98.12	98.16	(K-S) 93.80	(K-G) 96.37	(T-S) 95.05	(S-T) 94.26	94.87	(G-G) 97.76	(G-G) 96.83	97.30	96.67
5 dB	(T-K) 93.25	(K-T) 89.27	(S-G) 94.81	(T-G) 92.81	92.54	(K-S) 81.30	(K-G) 88.06	(S-S) 85.30	(S-G) 87.10	85.44	(T-G) 91.99	(S) 87.06	89.53	89.10
0 dB	(T-K) 78.14	(S-T) 63.57	(S-T) 80.88	(T-G) 75.50	74.52	(K-S) 50.38	(K-G) 68.32	(S-G) 63.64	(S) 69.55	62.97	(S) 71.11	(S) 65.45	68.28	68.65
-5 dB	(K) 46.27	(S-S) 25.82	(S-T) 46.14	(K) 41.62	39.96	(S-S) 14.95	(S) 33.56	(S) 28.36	(S-S) 38.20	28.77	(T-G) 38.90	(S) 32.71	35.81	34.65
Average	93.78	89.92	94.74	93.01	92.86	84.69	90.24	88.38	89.80	88.28	91.99	89.57	90.78	90.61

謝辞

本研究の一部は，総務省戦略的情報通信研究開発推進制度，及び通信・放送機構の研究委託による．

参考文献

[1] 北岡教英, 赤堀一郎, 中川聖一, “スペクトルサブトラクションと時間方向スムージングを用いた雑音環境下音声認識,” 電子情報通信学会論文誌, Vol. J83-D-II, No. 2, pp. 500–509, 2000.

[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] 山本一公, 中村哲, 武田一哉, 黒岩眞吾, 北岡教英, 山田武志, 水町光徳, 西浦敬信, 藤本雅清, “AURORA-2J/AURORA-3J データベースとその評価ベースライン,” 情報処理学会研究報告, SLP-47-19, 2003.

[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] Y. Gong, “A comparative study of approzimations for parallel model combination of static and dynamic parameters,” Proc. ICSLP2002, Vol. III, pp. 1209–1032, 2002.

[8] D. Macho, L. Mauuary, B. Noé, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, F. Saadoun, “Evaluation of a noise-robust DSR front-end on Aurora databases,” Proc. ICSLP2002, Vol. I, pp. 17–20, 2002.

[9] 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.