音声言語情報処理 45－2
（2003．2．7）

時間領域SVDとGMMに基づく音声信号推定法の統合による雑音下音声認識

藤本 雅清　　有木 康雄

龍谷大学 理工学部

〒 520-2194 大津市瀬田大江町横谷 1-5　　Tel: 077-543-7427

E-mail: masa@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

あらまし 本研究では，時間領域 SVD に基づく音声強調法と，GMM に基づく音声信号推定法を併用した雑音に頑健な音声認識手法を提案する．GMM に基づく音声信号推定法において最も大きな問題点は，雑音の平均ベクトルの推定問題であり，本研究では，雑音の時間変動に追随して雑音の平均ベクトルを逐次更新することについて検討した．また，より高い音声認識精度を得るために，時間領域 SVD に基づく音声強調法を GMM に基づく音声信号推定法の前処理として用いた．提案手法を AURORA2 データベースを用いて評価した結果，全ての雑音環境で大幅な音声認識率の改善が得られた．

キーワード： 雑音に頑健な音声認識，GMM に基づく音声信号推定，時間領域 SVD に基づく音声強調，AURORA2 データベース

Speech Recognition in Noise by Combination of Temporal Domain SVD Based Speech Enhancement and GMM Based Speech Estimation Method

Masakiyo Fujimoto　　Yasuo Ariki

Faculty of Science and Technology, Ryukoku University

1-5 Yokotani, Oe-cho, Seta, Otsu-shi, 520-2194 Japan　　Tel: +81-77-543-7427

E-mail: masa@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

Abstract A noise robust speech recognition method by combination of temporal domain singular value decomposition(SVD) based speech enhancement and Gaussian mixture model(GMM) based speech estimation is proposed in this paper. The critical neck of GMM based approach is the noise estimation problem. For this noise estimation problem, we investigated the adaptive noise estimation in GMM based approach. Furthermore, in order to obtain higher recognition accuracy, we employed a temporal domain SVD based speech enhancement method as the pre-processing module of GMM based approach. In evaluation on the AURORA2 tasks, our method showed the significant improvement in recognition accuracy at all the noise conditions.

Keywords : noise robust speech recognition, GMM based speech estimation, temporal domain SVD based speech enhancement, AURORA2 database

1　はじめに

近年，音声認識技術の飛躍的な進歩に伴い，音声認識システムの実用化が進められている．しかし，現行のシステムの多くは，実環境で背景雑音の影響が大きい場合，認識精度が著しく低下するという問題を抱えている．これを受けて，雑音に頑健な音声認識システムを確立するために，様々な研究が行われている [1].

雑音に頑健な音声認識システム確立のためのアプローチとして，認識システムを雑音に適応させる雑音適応法 [2, 3] と，雑音が重畳した音声から雑音成分を取り除く雑音除去法 [4, 5, 6] の 2 種類が考えられる．

雑音除去の方法として従来，SS(Spectral Subtraction) 法 [4] がよく用いられている．SS 法等を用いて雑音除去を行う際には，雑音重畳音声に含まれる雑音成分を何らかの方法で推定する必要がある．一般に，雑音が定常的である場合には，入力信号の開始数フレームを雑音のみが存在する区間であるとして，その区間の平均スペクトルを雑音重畳音声全体に含まれる雑音

成分と見なすことが多い．しかし，雑音が定常的であっても，実際に観測される雑音成分には微少な時間変動があり，雑音の種類によっては，この時間変動が無視できないものになる．この様な場合，雑音の平均スペクトル等を用いて，雑音成分の時間変動を無視することは，雑音除去後の音声のスペクトル歪みを増大させる要因になり，音声認識精度に影響を与えてしまう．

以上のような問題においてSegura らは，クリーン音声のGMM(Gaussian Mixture Model) と雑音の平均スペクトルを用いて各短時間フレーム毎に雑音成分の期待値を推定し，推定された期待値を用いて雑音除去処理を行うことにより，高い音声認識精度が得られることを示している [7]．しかし，Segura らの方法においても，入力信号の開始数フレームで得た雑音の平均スペクトルがパラメータとして用いられているため，雑音の時間変動について十分に考慮されていない．この問題において本研究では，過去に推定された雑音の平均スペクトルと現在のフレームおける観測信号を用いて，雑音の平均スペクトルを逐次更新することについて検討した．

また，より高い音声認識精度を得るために，時間 (波形) 領域での特異値分解 (SVD: Singular Value Decomposition) による音声強調手法 [8] を，GMM に基づく音声信号推定法の前処理として用いた．この様な処理を用いて事前に SNR を改善しておくことにより，GMM に基づく音声信号推定法がより効果的に働くものと考えられる．

ここで，一般に雑音除去処理を行うと，残差雑音及び，推定誤差等による音声信号の歪みが生じ，音声認識率に影響を与えるという問題がある．この問題を解決するために，教師無し MLLR 適応 [9] を用いることにより，推定誤差により生じるスペクトル歪みに音響モデルを適応させた．

提案手法の評価には，AURORA2[10] と呼ばれる雑音下音声認識の評価用データベースを用いており，評価の結果，AURORA2 データベースに含まれる全ての雑音環境において，大幅な認識率の改善が得られた．

2　処理概要

図 1 に提案手法の処理概要を示す．図 1 において，まず最初に時間領域 SVD に基づく音声強調法により，SNR を改善させる．次に，クリーン音声で学習した GMM を用いて，クリーン音声信号の推定を行う．最終的に，推定されたクリーン音声のメルフィルタバンク出力の対数値に対して DCT を適用して MFCC に変換し，CMS(Cepstral Mean Subtraction) を行った後に音声認識を行っている．また，この際，入力音声 1 文を用いて教師無し MLLR 適応を行う．以下，各処理の詳細について述べる．

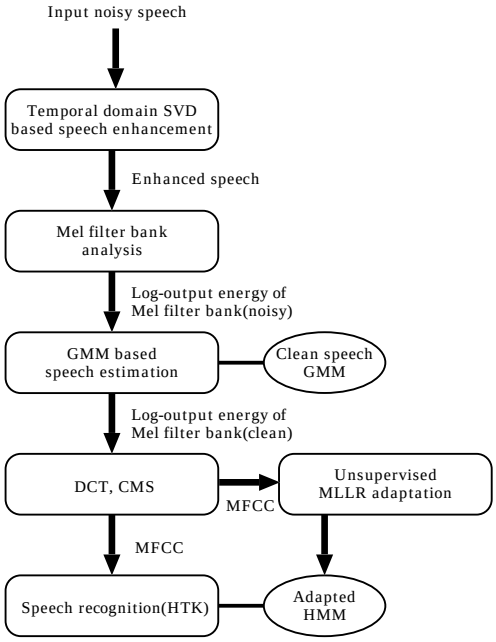

図 1: 提案手法の処理概要

3　時間領域 SVD に基づく音声強調

3.1　時間領域 SVD

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．

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

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．

$$x_i(t) = s_i(t) + n_i(t) \quad (2)$$

この時，式 (2) は，式 (1) の Toeplitz 行列を用いて式 (3) のように表すことができる．

$$\mathbf{X}_i = \mathbf{S}_i + \mathbf{N}_i \quad (3)$$

$\mathbf{X}_i$ 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 (4) のように表される．

$$\sigma_m^{\mathbf{X}_i} = \sigma_m^{\mathbf{S}_i} + \sigma_m^{\mathbf{N}_i} \quad (4)$$

式 (4) 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 (5) のように推定できる．

$$\hat{\sigma}_m^{\mathbf{S}_i} = \sigma_m^{\mathbf{X}_i} - \bar{\sigma}^{\mathbf{N}_i} \tag{5}$$

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 (6) の様に推定される．

$$\hat{\mathbf{S}}_i = \mathbf{U}_i \mathbf{W}_i \mathbf{\Sigma}_i \mathbf{V}_i^T \tag{6}$$

$$\mathbf{W}_i = diag\left(\frac{\sigma_m^{\mathbf{X}_i} - \bar{\sigma}^{\mathbf{N}_i}}{\sigma_m^{\mathbf{X}_i}}\right) \tag{7}$$

式 (4) 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．

$$\sigma_m^{\mathbf{N}_i} \simeq \sigma_m^{\mathbf{X}_i} \quad (m \geq R) \tag{8}$$

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．

$$\bar{\sigma}^{\mathbf{N}_i} = \frac{1}{M-R} \sum_{m=R}^{M-1} \sigma_m^{\mathbf{X}_i} \tag{9}$$

なお，本研究では，式 (1) 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 (10) 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．

$$ACR(r,i) = \frac{\sum_{m=0}^{r} \sigma_m^{\mathbf{X}_i}}{\sum_{m'=0}^{M-1} \sigma_{m'}^{\mathbf{X}_i}} \times 100 \tag{10}$$

$$R = \arg\min_r \{ACR(r,i) > 90\} \tag{11}$$

3.2 雑音の平均特異値の適応的減算

時間領域SVD に基づく音声強調法において，雑音の影響をより多く取り除くために，SS 法と同様にして，以下のように雑音の平均特異値 $\bar{\sigma}^{\mathbf{N}_i}$ の減算量を制御する係数 α を導入することを試みた．

$$\hat{\sigma}_m^{\mathbf{S}_i} = \sigma_m^{\mathbf{X}_i} - \alpha\bar{\sigma}^{\mathbf{N}_i} \tag{12}$$

ここで，α の値が大きな値に設定された場合，より多くの雑音成分を取り除くことができる．しかしこの場合，高 SNR の区間では過剰な減算により，信号歪みを発生させてしまう．一方，α の値を小さくした場合は信号歪みをおさえることができるが，低 SNR の区間では雑音成分を大きく残してしまう．

これらの問題を解決するためには，SNR に応じて適応的に α の値を設定する必要がある．ここで，一般に言われる SNR とは，音声全体での平均値 (Global SNR) のことであり，雑音が比較的定常であっても，1 フレーム単位で見た，局所的な SNR(Local SNR) はクリーン音声のパワーに応じて常に変化している．よって，1 フレーム単位の Local SNR($SNR(i)$ と定義する) に応じて，係数 α の値を式 (13) のような決定関数 g を定義して設定すれば，$\hat{\sigma}_m^{\mathbf{S}_i}$ のより高い推定精度が得られるものと考えられる [5].

$$\alpha(i) = g(SNR(i)) \tag{13}$$

次に，$SNR(i)$ の推定法について述べる．雑音重畳音声の短時間 RMS(Root Mean Square) パワーを $Pow_x(i)$，クリーン音声の推定短時間 RMS パワーを $Pow_s(i)$，雑音の平均推定短時間 RMS パワーを $\overline{Pow}_n$ としたとき，$SNR(i)$ は以下のように推定される．

$$SNR(i) = \begin{cases} 20\log_{10}\dfrac{\hat{Pow}_s(i)}{\overline{Pow}_n} & \hat{Pow}_s(i) > 0 \\ \gamma \quad (=-10) & \hat{Pow}_s(i) \leq 0 \end{cases} \tag{14}$$

$$\hat{Pow}_s(i) = Pow_x(i) - \overline{Pow}_n \tag{15}$$

式 (15) において，$\overline{Pow}_n$ は，観測信号の最初の 100ms が雑音のみの区間であると仮定して推定する．また，$Pow_s(i)$ が負の値を持つとき，$SNR(i)$ を計算できないので，定数 γ を代入する．

$SNR(i)$ により $\alpha(i)$ を与える決定関数 g として，本研究では図 2 に示すような関数を与えた．なお，この関数 g の形状は実験的に求めたものである．

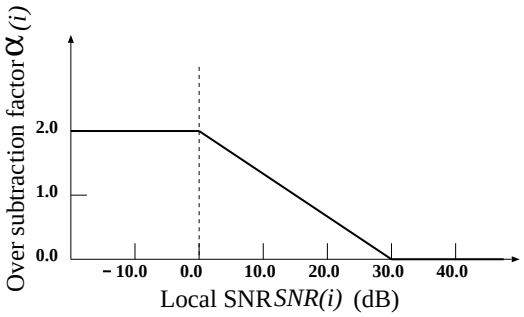

図 2: 減算制御係数の決定関数 $g(SNR(i))$

4 GMM に基づく音声信号推定

4.1 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．

$$\begin{aligned}\mathbf{X}(i) &= \log\left[\exp(\mathbf{S}(i)) + \exp(\mathbf{N}(i))\right] \\ &= \mathbf{S}(i) + \log\left[1 + \exp(\mathbf{N}(i) - \mathbf{S}(i))\right] \\ &= \mathbf{S}(i) + \mathbf{G}(i) \end{aligned} \tag{16}$$

$$\mathbf{G}(i) = \log\left[1 + \exp(\mathbf{N}(i) - \mathbf{S}(i))\right] \tag{17}$$

式 (16) において，$\mathbf{G}(i)$ は $\mathbf{X}(i)$ における雑音成分（$\mathbf{S}(i)$ とのミスマッチ成分）に相当する．

4.2 GMM を用いた $\mathbf{G}(i)$ の期待値の推定

式 (18) に示す $\mathbf{S}(i)$ の K 混合分布 GMM を用いて，$\mathbf{G}(i)$ の期待値を推定する．

$$p(\mathbf{S}(i)) = \sum_{k=1}^{K} P(k)\mathcal{N}(\mathbf{S}(i), \mu_{S,k}, \mathbf{\Sigma}_{S,k}) \tag{18}$$

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 (18) のような GMM が与えられたときに，$\mathbf{X}(i)$ を Log-add compensation 法 [3] 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 (16) を用いて，

$$\begin{aligned}\mu_{X,k} &\simeq \mu_{S,k} + \log\left[1 + \exp(\mu_N - \mu_{S,k})\right] \\ &= \mu_{S,k} + \mu_{G,k} \end{aligned} \tag{19}$$

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

$$\mathbf{\Sigma}_{X,k} \simeq \mathbf{\Sigma}_{S,k} \tag{20}$$

として近似する．

式 (19) 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 (21) 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 $\mathbf{G}(i)$ の期待値 $\hat{\mathbf{G}}(i)$ を推定する．

$$\hat{\mathbf{G}}(i) = \sum_{k=1}^{K} P(k|\mathbf{X}(i))\mu_{G,k} \tag{21}$$

$$P(k|\mathbf{X}(i)) = \frac{P(k)\mathcal{N}(\mathbf{X}(i), \mu_{X,k}, \mathbf{\Sigma}_{X,k})}{\sum_{k'=1}^{K} P(k')\mathcal{N}(\mathbf{X}(i), \mu_{X,k'}, \mathbf{\Sigma}_{X,k'})} \tag{22}$$

以上の手法により得られた $\hat{\mathbf{G}}(i)$ を用いて，$\mathbf{S}(i)$ の推定値 $\hat{\mathbf{S}}(i)$ は，次式により得られる [7]．

$$\hat{\mathbf{S}}(i) = \mathbf{X}(i) - \hat{\mathbf{G}}(i) \tag{23}$$

4.3 雑音平均ベクトルの逐次更新

4.2 では，音声信号の推定の際に，雑音のみであると見なされる区間で推定された雑音の平均ベクトル μ_N を，全てのフレームにおいて用いていた．しかし，雑音が時間変動することを考えた場合，雑音平均ベクトルの推定値にこのような時間不変の値を用いることは好ましくない．従って，本研究では，式 (24) に示すように，雑音平均ベクトルをスムージングにより各周波数帯域毎に更新することを試みた．

$$\mu_{Nj}(i) = \rho\mu_{Nj}(i-1) + (1-\rho)X_j(i) \tag{24}$$

ここで，j はベクトル $\mathbf{X}(i)$，μ_N の要素番号 (周波数帯域の番号)，$X_j(i)$ は $\mathbf{X}(i)$ の第 j 要素，$\mu_{Nj}(i)$ は更新されたフレーム i での μ_N の第 j 要素である．

雑音の推定値の更新は，雑音が比較的緩やかな時間変化をすると仮定し，式 (25) が満たされる場合にのみ行う．

$$\exp(X_j(i)) < \eta \cdot \exp(\mu_{Nj}(i)) \tag{25}$$

なお，本研究では式 (24)，(25) で用いられるパラメータは，それぞれ $\rho = 0.97$，$\eta = 2$ としている．

5 教師無し MLLR 適応

一般に雑音除去を行うと，推定誤差等による残差雑音及び，音声スペクトルの歪みが生じてしまい，音声認識率に影響を与えるという問題がある．この問題を解決するために，本研究では教師無し MLLR 適応 [9] を用いることにより，推定誤差により生じるスペクトル歪みに音響モデルを適応させた．教師無し MLLR 適応を数字 HMM に対して行うためには，適応データの数字ラベルが必要となる．本研究では，適応データを適応前の HMM により認識した結果を数字ラベルとして用いている．また，適応データには入力音声 1 文章のみを用いており，MLLR 適応における HMM 内の正規分布クラスタ数は 1 とした．

6 実験

以上に述べた手法を用いて，AURORA2 データベースによる評価を行った．

6.1 AURORA2 データベース

本研究で使用した AURORA2 データベースは，ELRA [11] より配布されている雑音下音声認識の評価用データベースである．AURORA2 データベース内の雑音重畳音声データは，TI-Digits(連続英語数字音声) データベースに種々の雑音を人工的に重畳することにより生成されており，表 1 に示すような 3 種類のテストセットが用意されている [10]．

表 1: AURORA2 データベースの雑音環境

	加算性雑音	フィルタ特性
SetA	Subway, Babble, Car, Exhibition	G.712
SetB	Restaurant, Street Airport, Station	G.712
SetC	Subway, Street	MIRS

表 1 において，SetA，SetB ではそれぞれ 4 種類，SetC では SetA，SetB から 1 種類ずつ選択した加算性雑音が用いられ，SNR は-5〜20dB(5dB 刻み) 及びクリーン環境が用意されている．全ての音声データには，電話回線を模擬したフィルタ特性が畳み込まれており，SetA，SetB では G.712，SetC では MIRS と呼ばれるフィルタ特性になっている [10]．なお，G.712 のフィルタ特性は，全ての学習データにも畳み込まれている．このため，SetA，SetB は加算性雑音のみが存在する環境での評価であり，SetC は，加算性雑音に加えて，乗法性歪みが存在する環境での評価となる．また，各雑音，SNR 毎に 1001 文章の音声データ (男女混在) がテストデータとして用意されており，各音声データの標本化周波数は 8kHz(16bit) である．

次に認識システムと，評価方法について述べる．HMM の学習及び認識は，HTK[12] により行われる．認識時の語彙数は 13(数字 1〜9，oh，zero，無音，ショートポーズ) であり，各語彙毎に Whole Word HMM を学習する．AURORA2 データベース標準の HMM の構造は表 2 の通りである．

表 2: AURORA2 データベース標準 HMM の構造

	状態数	混合分布数
数字 (1〜9, oh, zero)	16	3
無音	3	6
ショートポーズ	1	6

HMM の学習データは，クリーン音声のみのデータ (Clean) と，雑音重畳音声を含んだデータ (Multi) の 2 種類のデータセットが用意されており，それぞれのデータセットを用いて HMM を学習する．Multi に含まれる雑音重畳音声データには，テストセット SetA に含まれる 4 種類の雑音が重畳しており，SNR はクリーン及び，5〜20dB のみである．認識は，クリーン音声により学習された HMM と，雑音重畳音声により学習された HMM それぞれを用いて行う．

6.2 フロントエンド処理による実験結果

まず，フロントエンド処理部にあたる，以下の 3 種類の雑音除去処理の評価を行った．また，AURORA2 データベースでの評価において，本研究では，クリーン音声で学習された HMM を用いて評価している．

手法 1 ： 時間領域 SVD に基づく音声強調

手法 2 ： GMM に基づく音声信号推定

手法 3 ： 手法 1 + 手法 2(提案手法)

提案手法の評価における音響分析条件は表 3 の通りであり，全てのデータに CMN 処理を行っている．また，GMM に基づく音声信号推定法において必要となるクリーン音声の GMM には，学習データに含まれる全てのクリーン音声から学習した 256 混合分布の GMM を用いている．

表 3: 音響分析条件

標本化周波数	8kHz(16bits)
高域強調	$1 - 0.97z^{-1}$
特徴パラメータ	13 次 MFCC(0 次含む) + Δ + ΔΔ
分析区間長	25ms
分析周期	10ms
時間窓	Hamming window

ここで，AURORA2 データベースのテストデータには各雑音環境において，-5〜20dB 及び，Clean の 7 段階の雑音レベルが用意されているが，一般に AURORA2 データベースにおいて評価の対象となるのは，0〜20dB の環境である [13]．このため，本研究においても，0〜20dB の環境を対象として評価を行っている．

表 4 に，それぞれの手法による SetA，SetB，SetC の平均認識率を示す．

表 4: 単語正解精度 (%)

SNR	Baseline	手法 1	手法 2	手法 3
20dB	93.98	96.83	**98.09**	97.93
15dB	83.52	93.14	**96.41**	96.11
10dB	62.83	83.02	**92.13**	91.59
5dB	35.39	61.87	79.30	**80.81**
0dB	14.56	34.61	52.06	**58.08**
Average	58.06	73.90	83.60	**84.90**

手法 1，2 と手法 3 を比較した結果，低 SNR 環境で大きな改善が得られ，時間領域 SVD に基づく音声強調手法が，GMM に基づく音声信号推定法の前処理として，効果的に働いたことが確認できる．

しかし，高 SNR 環境では，手法 2 と比べて認識率が僅かに低下している．この認識率の低下の原因として，時間領域 SVD に基づく音声強調手法により，雑音の定常的な成分が抑圧されたが，非定常的な成分が十分に抑圧されなかったため，非定常性が強調された残差雑音が残留してしまったことが考えられる．このことにより，雑音平均ベクトルの逐次更新が手法 2 の場合に比べて有効に動作せず，認識率に影響を与えてしまったと考えられる．低 SNR 環境においても，このような非定常性の強調による影響は存在していると考えられるが，低 SNR 環境では，認識率に大きな影響を与えていた，雑音成分の主となる定常的な成分を時間領域 SVD により事前に取り除くことができたため，結果として認識率の改善が得られたと考えられる．

以上のような問題を解決するために，今後，雑音の非定常成分の抑圧手法および，より高精度な雑音平均ベクトルの更新手法について検討する必要がある．

6.3 教師無し MLLR 適応による実験結果

次に，6.2 の手法 3 により得られた推定音声信号を適応データとして，教師無し MLLR 適応を適用した．表 5 に，教師無し MLLR 適応による SetA，SetB，SetC の平均認識率を示す．

表 5: 単語正解精度 (%)

SNR	手法 3	手法 3+MLLR
20dB	97.93	**98.03**
15dB	96.11	**96.35**
10dB	91.59	**92.88**
5dB	80.81	**83.43**
0dB	58.08	**58.89**
Average	84.90	**85.97**

表 5 の結果より，教師無し MLLR 適応を適用することにより，適応無しの場合に比べて，平均で約 1%の認識精度の改善が得られた．

今回の実験では，適応データとして入力音声 1 文のみを用いているが，このような非常に少量な適応データでは，MLLR 適応の性能を十分に発揮できていないと考えられる．このため今後，非常に少量な適応データであっても，高精度に音響モデルの適応を行うことのできる手法について検討を行う必要がある．

7 おわりに

本研究では，時間領域 SVD に基づく音声強調法と GMM に基づく音声信号推定法を用いた，雑音に頑健な音声認識手法を提案した．提案手法を AURORA2 データベースを用いて評価した結果，全ての雑音環境で大幅な音声認識率の改善が得られ，提案手法の有効性が確認できた．今後，雑音の非定常的な成分の抑圧手法及び，より高精度な雑音平均ベクトルの更新手法について検討する予定である．また，時間領域 SVD では，雑音の白色性の前提をおいていたが，雑音が有色性であっても効果的に雑音成分を抑圧できる手法についても検討する予定である．

謝辞

本研究を行うにあたり多大な助言を頂いた，SLP 雑音下音声認識評価ワーキンググループの皆様方に深く感謝致します．

参考文献

[1] 中村 哲: “実音響環境に頑健な音声認識を目指して”, 信学技報, EA2002-12, pp.31-36(2002).

[2] M.J.F.Gales and S.J.Young: “Robust Continuous Speech Recognition Using Parallel Model Combination”, IEEE Trans. Speech and Audio Processing, Vol.4, No.5, pp.352-359, Sep.(1996)

[3] Y.Gong: “A Comparative Study of Approzimations for Parallel Model Combination of Static and Dynamic Parameters”, ICSLP'02, Vol.III, pp.1209-1032(2002).

[4] S.F.Boll: “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, IEEE Trans. Acoustic Speech Signal Processing, Vol.27, No.2, pp.113-120, (1979)

[5] 山本 寛樹, 山田 雅章, 小森 康弘, 大洞 恭則: “推定 Segmental SNR に基づく適応的 Spectral Subtraction 法による音声認識”, 信学技報, SP94-50, pp.17-24(1994).

[6] M.Fujimoto and Y.Ariki: “Evaluation of Noisy Speech Recognition Based on Noise Reduction and Acoustic Model Adaptation on the AURORA2 Tasks”, ICSLP'02, Vol.I, pp.465-468(2002).

[7] J.C.Segura, A.de la Torre, M.C.Benitez and A.M.Peinado: “Model-Based Compensation of the Additive Noise for Continuous Speech Recognition. Experiments Using AURORA II Database and Tasks”, EuroSpeech'01, Vol.I, pp.221-224(2001).

[8] C.Uhl and M.Lieb: “Experiments with an Extend Adaptive SVD Enhancement Scheme for Speech Recognition in Noise”, ICASSP'01(2001).

[9] C.L.Leggetter and P.C.Woodland: “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models”, Computer Speech and Language, Vol.9, pp.171-185(1995).

[10] H.G.Hirsch and D.Pearce: “The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condition”, ISCA ITRW ASR2000, pp.18-20(2000).

[11] ELRA Web site: http://www.icp.inpg.fr/ELRA/home.html

[12] HTK Web site: http://htk.eng.cam.ac.uk/

[13] AURORA2 Spread sheet: http://icslp2002.colorado.edu/special_sessions/aurora/