音声言語情報処理 44－5
(2002. 12. 19)

GMMに基づく音声信号推定法を用いた雑音下音声認識

藤本 雅清† 有木 康雄†

† 龍谷大学 理工学部
〒520-2194 滋賀県大津市瀬田大江町横谷1-5
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する．本手法の主となる部分には，GMMに基づく音声信号推定法を用いている．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による音声強調手法をGMMに基づく音声信号推定法の前処理として用いた．さらに，時間領域SVD法において，雑音の影響をより多く取り除くために，雑音成分の減算制御係数を導入し，この値を適応的に決定することについても検討した．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，AURORA2データベース

Noise Robust Speech Recognition Using GMM Based Speech Estimation Method

Masakiyo FUJIMOTO† and Yasuo ARIKI†

† Faculty of Science and Technology, Ryukoku University
1-5 Yokotani, Oe-cho, Seta, Otsu, 520-2194 Japan
E-mail: †masa@arikilab.elec.ryukoku.ac.jp, ††ariki@rins.st.ryukoku.ac.jp

Abstract In this paper, a noise robust speech recognition method is proposed, by combining temporal domain singular value decomposition(SVD) based speech enhancement and Gaussian mixture model(GMM) based speech estimation. The critical neck of the GMM based approach is the noise estimation problem. For this noise estimation problem, we investigated the adaptive noise estimation in the 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 the GMM based approach. In addition, to reduce more influence of the noise included in the noisy speech, we introduce an adaptive over subtraction factor into the temporal domain SVD based speech enhancement. In evaluation on the AURORA2 tasks, our method showed the significant improvement in the recognition accuracy at all the noise conditions.

Key words noise robust speech recognition, GMM based speech estimation, temporal domain SVD, 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，それらの多くは比較的静かな環境を想定したものが大半を占めており，実環境で背景雑音の影響が大きい場合，認識率が極端に低下してしまうという問題があり，完全な実用化には至っていないのが現状である．これを受けて，背景雑音に頑健な音声認識システムを確立し，音声認識システムの実用化を実現するために，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方法(雑音適応)[2]-[5]と，雑音が重畳した音声から雑音成分を取り除き，クリーンな音声を抽出して認識を行う方法(雑音除去)[6]-[8]の2種類が考えられる．

雑音適応の方法として，PMC(Parallel Model Combination)

法[2]やNOVO(VOice mixed with NOise)法[3]に代表されるHMM合成法が提案されており，その有効性が報告されている．また，非定常雑音下での認識率を改善するために，初期の合成HMMを雑音の時間変動に応じて逐次的に適応させていく手法が提案されている[4]．しかし，HMM合成法は合成のための計算量が比較的多く，大語彙連続音声認識に使われるTriphoneモデルのHMMのようにモデル数，混合数の多いHMMに対して合成を行うと，膨大な時間がかかってしまうという問題がある．

一方，雑音除去の方法では従来，SS(Spectral Subtraction)法[6]がよく用いられており，少ない計算量の割には有効な手法であることが知られている．ここで，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，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，雑音除去後のスペクトル歪みを増大させる要因になり，最終的な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[9]．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．この問題に対して本研究では，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[10]を，GMMに基づく音声信号推定法の前処理として用いた．この様な処理を加えて事前に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．さらに，時間領域SVDによる音声強調手法において，雑音の影響をより多く取り除くために，SS法等と同様に雑音成分の減算量を制御する係数を導入し，この係数を適応的に決定することについても検討した．

提案手法の評価には，AURORA2[11], [12]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に基いて音声強調を行うことにより，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．

以下，3.では時間領域SVDに基づく音声強調法，4.ではGMMに基づく音声信号推定法について述べる．

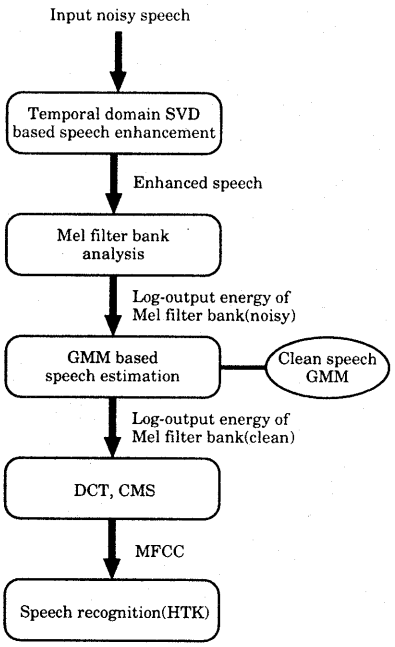

図1　提案手法の処理概要

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．

$$\sigma_m^{\mathbf{X}_i} = \sigma_m^{\mathbf{S}_i} + \sigma_m^{\mathbf{N}_i} \tag{4}$$

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．

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

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．

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

3.2 雑音の平均特異値の適応的減算

3.1では，時間領域SVDによる音声強調法について述べた．本研究では，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．

$$\hat{\sigma}_m^{\mathbf{S}_i} = \sigma_m^{\mathbf{X}_i} - \alpha\bar{\sigma}^{\mathbf{N}_i} \tag{10}$$

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(11)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．

$$\alpha(i) = g(SNR(i)) \tag{11}$$

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．

$$SNR(i) = \begin{cases} 20\log_{10}\dfrac{\hat{Pow}_s(i)}{\overline{Pow}_n} & \hat{Pow}_s(i) > 0 \\ \gamma \ (=-10) & \hat{Pow}_s(i) \leqq 0 \end{cases} \tag{12}$$

$$\hat{Pow}_s(i) = Pow_x(i) - \overline{Pow}_n \tag{13}$$

式(13)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．

次に，得られたLocal SNRである $SNR(i)$ を用いて，減算制御係数 $\alpha(i)$ を与える決定関数 g を構成する．本研究では減算制御係数の決定関数 g として図2に示すような関数を与えており，$SNR(i)$ が0dB以下の場合は減算量が最大(2.0倍)となり，30dB以上の場合には減算を行っていない．尚，この関数 g の形状は実験的に求めたものである．

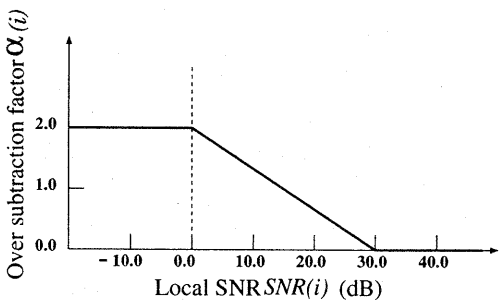

図2 減算制御係数の決定関数 $g(SNR(i))$

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．

$$\begin{aligned} \mathbf{X}(i) &= \log\left[\exp(\mathbf{S}(i)) + \exp(\mathbf{N}(i))\right] \\ &= \log\left[\exp(\mathbf{S}(i))\left(1 + \frac{\exp(\mathbf{N}(i))}{\exp(\mathbf{S}(i))}\right)\right] \\ &= \mathbf{S}(i) + \log\left[1 + \exp(\mathbf{N}(i) - \mathbf{S}(i))\right] \\ &= \mathbf{S}(i) + \mathbf{G}(i) \end{aligned} \tag{14}$$

$$\mathbf{G}(i) = \log\left[1 + \exp(\mathbf{N}(i) - \mathbf{S}(i))\right] \tag{15}$$

式(14)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G(i)の期待値の推定

式(16)に示すS(i)のK混合分布GMMを用いて，G(i)の期待値を推定する．

$$p(\mathbf{S}(i)) = \sum_{k=1}^{K} P(k)\mathcal{N}(\mathbf{S}(i), \mu_{S,k}, \mathbf{\Sigma}_{S,k}) \tag{16}$$

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(16)のようなGMMが与えられたときに，$\mathbf{X}(i)$をLog-add compensation法[5]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(14)を用いて，

$$\begin{aligned}\mu_{X,k} &\simeq \mu_{S,k} + \log\left[1 + \exp(\mu_N - \mu_{S,k})\right] \\ &= \mu_{S,k} + \mu_{G,k}\end{aligned} \tag{17}$$

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

$$\mathbf{\Sigma}_{X,k} \simeq \mathbf{\Sigma}_{S,k} \tag{18}$$

として近似する．

式(17)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(19)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．

$$\hat{\mathbf{G}}(i) = \sum_{k=1}^{K} P(k|\mathbf{X}(i))\mu_{G,k} \tag{19}$$

$$P(k|\mathbf{X}(i)) = \frac{P(k)\mathcal{N}(\mathbf{X}(i), \mu_{X,k}, \mathbf{\Sigma}_{X,k})}{\sum_{k'=1}^{K} P(k')\mathcal{N}(\mathbf{X}(i), \mu_{X,k'}, \mathbf{\Sigma}_{X,k'})} \tag{20}$$

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[9]．

$$\hat{\mathbf{S}}(i) = \mathbf{X}(i) - \hat{\mathbf{G}}(i) \tag{21}$$

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(22)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．

$$\mu_{Nj}(i) = \rho\mu_{Nj}(i-1) + (1-\rho)X_j(i) \tag{22}$$

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(23)が満たされる場合にのみ行う．

$$\exp(X_j(i)) < \eta \cdot \exp(\mu_{Nj}(i)) \tag{23}$$

5. 実　　験

3.，4.で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．

5.1 AURORA2データベース

本研究で使用したAURORA2データベースの詳細について述べる．AURORA2データベースは仏国ELRA(European Language Resources Association)[13]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データベースに含まれる雑音重畳音声データは，米国LDC(Linguistic Data Consortium)[14]より配布されているTI-Digits(英語連続数字音声)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[11]．

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[11]．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(Hidden Markov Model Toolkit)[15]により行われており，学習，認識を行うためのスクリプトが提供されている．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HMMを学習する(Whole Word HMM)．AURORA2データベース標準のHMMの構造は表2の通りであり，ショートポーズのHMMは無音HMMの第3状態を共有している．また，全てのHMMにおいて状態のスキップは考慮されていない．

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)	18状態16ループ	3
無音	5状態3ループ	6
ショートポーズ	3状態1ループ	6

HMMの学習データセットとしては，クリーン音声のみの学習データセット(Clean Condition Training)と，雑音重畳音声を含んだ学習データセット(Multi Condition Training)の2種類の学習データセットが用意されており，それぞれの学習データセットを用いてHMMを学習する．Multi Condition

Training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のみが既知である．学習データの量は，Clean Condition Training，Multi Condition Training共に8440文章であり，Multi Condition Trainingでは，各雑音環境毎に422文章ずつ(422 × 4(雑音の種類) ×5(SNR) = 8440)という内訳になっている．

提案手法の評価の際の特徴パラメータには，表3に示すように，0次MFCCを含む13次元のMFCC(AURORA2標準の特徴抽出では12次元MFCCとLog-Energy)と1次，2次の回帰係数を含めた39次元の特徴ベクトルを用いており，各文章に対してCMS処理を行っている(AURORA2標準の特徴抽出ではCMS処理は行わない)．

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

5.2 実験結果

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(Clean Condition Training)を用いて，以下の5種類の方法を評価した．

手法1：時間領域SVDに基づく音声強調

手法2：手法1 + 減算制御係数

手法3：GMMに基づく音声信号推定

手法4：手法3 + 雑音平均ベクトルの逐次更新

手法5：手法2 + 手法4(提案手法)

5.2.1 手法1と手法2の比較

時間領域SVDに基づく音声強調において，式(1)のToeplitz行列の次元を決定するパラメータには，$M = 28$および$N = 173$を与えた．また，特異値の打ち切り次元Rには，式(24)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．

$$ACR(r,i) = \frac{\sum_{m=0}^{r} \sigma_m^{\mathbf{X}_i}}{\sum_{m'=0}^{M-1} \sigma_{m'}^{\mathbf{X}_i}} \times 100 \qquad (24)$$

$$R = \arg\min_{r} \{ACR(r,i) > 90\} \qquad (25)$$

表4に，手法1及び2によるSetA，SetB，SetCの平均認識率を示す．

手法1と手法2を比較した結果，手法2により全ての環境において，認識率の改善が得られ，3.2で述べた雑音の平均特異値の減算制御係数による効果が確認できた．

表4　単語正解精度(%)

SNR	Baseline	手法1	手法2
Clean	99.07	98.21	**99.14**
20dB	93.98	94.74	**96.83**
15dB	83.52	90.13	**93.14**
10dB	62.83	78.54	**83.02**
5dB	35.39	56.05	**61.87**
0dB	14.56	30.14	**34.61**
-5dB	6.95	14.90	**16.54**
Ave.(20~0dB)	58.06	69.92	**73.90**
Ave.(overall)	56.61	66.13	**69.31**

5.2.2 手法3と手法4の比較

GMMに基づく音声信号推定法において，本研究では，学習データに含まれる全てのクリーン音声から，メルフィルタバンク出力の対数値を特徴量とした64混合分布のGMMを学習した．また，雑音平均ベクトルの更新の際に用いるパラメータは，それぞれ$\rho = 0.97$，$\eta = 2$とした．

表5に，手法3及び4によるSetA，SetB，SetCの平均認識率を示す．

手法3と手法4を比較した結果，手法4により全ての環境において認識率の改善が得られたが，全体的に改善量は小さい．雑音平均ベクトルの逐次更新による，認識率改善量が僅かであった理由として，以下の2つのことが考えられる．

(I): 比較的強い非定常性を持つ雑音であったため，雑音平均ベクトルの逐次更新が有効に作用しなかった．

(II): 式(23)の条件を満たさない場合は，音声＋雑音，もしくは突発性雑音が発生したとみなされるため，更新が行われない．

以上のような理由により，手法4による改善が小さかったと考えられる．

表5　単語正解精度(%)

SNR	Baseline	手法3	手法4
Clean	**99.07**	99.06	99.06
20dB	93.98	**98.03**	**98.03**
15dB	83.52	96.39	**96.54**
10dB	62.83	91.44	**91.78**
5dB	35.39	77.16	**77.84**
0dB	14.56	47.55	**49.63**
-5dB	6.95	20.61	**21.45**
Ave.(20~0dB)	56.61	82.11	**82.76**
Ave.(overall)	56.61	75.75	**76.33**

5.2.3 手法2，4と手法5の比較

表6に，手法2，4及び5によるSetA，SetB，SetCの平均認識率を示す．

手法2，4と手法5を比較した結果，低SNRで大きな改善が得られ，SVDに基づく音声強調手法が，GMMによる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では，手法4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，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，非定常性の強い残差雑

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．このような問題を解決するために，今後，雑音の非定常的な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．

表6　単語正解精度(%)

SNR	Baseline	手法2	手法4	手法5
Clean	99.07	**99.14**	99.06	99.08
20dB	93.98	96.83	**98.03**	97.80
15dB	83.52	93.14	**96.54**	96.04
10dB	62.83	83.02	**91.78**	91.20
5dB	35.39	61.87	77.84	**79.53**
0dB	14.56	34.61	49.63	**55.47**
-5dB	6.95	16.54	21.45	**26.98**
Ave.(20~0dB)	56.61	73.90	82.76	**84.01**
Ave.(overall)	56.61	69.31	76.33	**78.01**

5.2.4　GMMの混合分布数の認識率への影響

GMMによる音声信号推定法を行う際に用いる，GMMの混合分布数の認識率への影響についての調査を行った．今回の実験では，混合分布数が64，128，256のGMMを用い，手法5を用いて評価を行っている．

表7に，3種類のGMMによるSetA，SetB，SetCの平均認識率を示す．

表7の結果より，GMMの混合分布数を増加させるにつれ，認識率が改善されていることがわかる．このことから，GMMによる音声信号推定法では，より多くの要素分布を持つGMMを利用することにより，高い音声信号の推定精度が得られると言える．しかし，GMMの学習に利用できる音声データの量は有限であるため，あまりに多くの要素分布を持つGMMでは，Data sparsenessの問題により学習に要するデータの欠乏が生じ，GMMのパラメータ推定の精度が劣化すると考えられる．このようなパラメータの推定精度が劣化したGMMを用いて音声信号の推定を行った場合，音声信号の推定精度もまた劣化すると考えられる．

このようなGMMの学習における問題を回避するため，変分ベイズ学習法[16], [17]等を用いて，与えられた有限の学習データを最適に表現するモデルを学習することについて検討する必要がある．

表7　単語正解精度(%)

SNR	Baseline	64混合分布	128混合分布	256混合分布
Clean	99.07	99.08	**99.11**	99.10
20dB	93.98	97.80	97.88	**97.93**
15dB	83.52	96.04	96.05	**96.11**
10dB	62.83	91.20	91.20	**91.59**
5dB	35.39	79.53	80.06	**80.81**
0dB	14.56	55.47	56.68	**58.08**
-5dB	6.95	26.98	27.77	**28.94**
Ave.(20~0dB)	56.61	84.01	84.37	**84.90**
Ave.(overall)	56.61	78.01	78.39	**78.94**

6. 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タスク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．今後，雑音の非定常的な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[12]の皆様方に深く感謝致します．

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] F.Martin, K.Shikano, Y.Minami and Y.Okabe: “Recognition of Noisy Speech by Composition of Hidden Markov Models”, 信学技報, SP92-96, pp.9-16(1992).
[4] K.Yao, K.K.Paliwal and S.Nakamura: “Sequential Noise Compensation by A Sequential Kullback Proximal Algorithm”, EuroSpeech'01, Vol.II, pp.1139-1142(2001).
[5] Y.Gong: “A Comparative Study of Approzimations for Parallel Model Combination of Static and Dynamic Parameters”, ICSLP'02, Vol.III, pp.1209-1032(2002).
[6] 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)
[7] 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).
[8] 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).
[9] 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).
[10] C.Uhl and M.Lieb: “Experiments with an Extend Adaptive SVD Enhancement Scheme for Speech Recognition in Noise”, ICASSP'01(2001).
[11] 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).
[12] 中村 哲, 武田一哉, 黒岩眞吾, 山田武志, 北岡教英, 山本一公, 西浦敬信, 藤本雅清, 水町光徳: “SLP雑音下音声認識評価ワーキンググループ活動報告”, 情報処理学会研究報告, SLP-42-11, pp.65-70(2002).
[13] ELRA Web site: http://www.icp.inpg.fr/ELRA/home.html
[14] LDC Web site: http://www.ldc.upenn.edu/index.html
[15] HTK Web site: http://htk.eng.cam.ac.uk/
[16] 上田修功: “ベイズ学習[III] — 変分ベイズ学習の基礎 —”, 電子情報通信学会誌, Vol.85, No.7, pp.504-509(2002).
[17] 上田修功: “ベイズ学習[IV・完] — 変分ベイズ学習の応用例 —”, 電子情報通信学会誌, Vol.85, No.8, pp.633-638(2002).