社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－SLP－54（41）
2004／12／21

GMMによる雑音抑圧手法選択に基づく雑音下音声認識

濱口　早太†　　北岡　教英†　　中川　聖一†

† 豊橋技術科学大学　〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘1-1
E-mail: †{hamaguchi,kitaoka,nakagawa}@slp.ics.tut.ac.jp

あらまし　様々な雑音条件下では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ている。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）の統合を行う。予めいくつかの雑音を重畳した学習用音声データに対して、それらの雑音条件に最適な抑圧手法を調べ、これを基にGMMで自動的に抑圧手法を選択する手法を提案する。本手法をAURORA-2Jを用いて評価した。その結果、雑音条件に適した雑音抑圧手法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示された。

キーワード　雑音下音声認識, 雑音抑圧手法選択, GMM, AURORA-2J

Robust speech recognition under noisy environments based on selection of multiple noise suppression methods using GMMs.

Souta HAMAGUCHI†, Norihide KITAOKA†, and Seiichi NAKAGAWA†

† Toyohashi University of Technology　Hibarigaoka 1-1, Tempaku-cho, Toyohashi, Aichi, 441-8580 Japan
E-mail: †{hamaguchi,kitaoka,nakagawa}@slp.ics.tut.ac.jp

Abstract　To achieve high recognition performance for a wide variety of noise and for a wide range of signal-to-noise ratio, this report 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. In this report, we investigated the optimal suppression method for each noise condition, and then also developed the method of choosing the optimal method automatically for unknown noise. Recognition results on the AURORA-2J task show the effectiveness of our proposed method .

Key words　noisy speech recognition, selection of noise suppression methods, GMM, AURORA-2J

1. はじめに

音声認識システムの実用化を目指した研究が広く行われている。しかし、音声認識の利用が想定される環境においては周囲雑音が存在することが多く、これらは音声認識システムの認識精度を極端に低下させる。認識システムの精度向上のためにはロバストな雑音下音声認識を実現することが必須である[1]。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。これらの手法の有効性は雑音条件に強く依存していることが多く、雑音の種類やＳＮＲにより得手不得手があ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と一般的に考えられている。すなわち、一種類の雑音抑圧手法の適用では効果に限界がある。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ことなどが有効と考えられる。

この問題に対し、最近、雑音を含む音響環境の変動に対応する複数の音響モデルを用意して並列に動作させ尤度最大確率で仮説を選択する方法[8][9]、ＳＮ比別のマルチパスモデルを用いて雑音のＳＮ比の多様性に対処する手法[10]、複数の雑音抑圧手法からそれぞれ得ていた特徴量を用いて並列に認識を行い得られた仮説を組み合わせる手法[11]などが提案されているが演算コストが大きい。また分散音声認識を考えた場合、バックエンドの処理が必要となるため現実的でない。福田らの手法[12]はフロントエンド処理であるが特殊な特徴量を使用する。

本稿ではGMMを用いて入力雑音に対し最適な抑圧手法を適宜選択する手法を提案する。各種の雑音に対する抑圧手法の効果を事前に調査し、それを基に抑圧手法別のGMMを学習する。認識時にはGMMで入力雑音の種類を認識前に判別し、最適な雑音抑圧手法を選択することで認識精度の向上を試みる。また、雑音重畳、抑圧処理後の音声で音響モデルを学習するマルチコ

ンディション学習ではある雑音条件で最適と思われた抑圧手法が、音響モデルの学習後にはそうでなくなり認識性能の低下を引き起こす。これにはGMMおよび音響モデルの反復学習を行うことによって対処する。実験には連続数字音声データベースである AURORA-2J [6] を使用する。AURORA-2J は雑音環境下連続英語数字音声認識タスクの共通評価フレームワークである AURORA-2 [2] の日本語版である。

2. 雑音抑圧手法

本稿では、既存の雑音抑圧手法として、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 [3], 時間領域 SVD に基づく音声信号推定 [4],GMM に基づく音声信号推定 [4], ピッチ同期 KLT [5] を用いる。上記の 4 つの手法を単独、もしくはこれら 2 つを逐次的に組み合わせた手法の計 20 種類を雑音重畳音声に適用する。

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 [3]

スペクトルの時間方向のスムージングを考慮したスペクトルサブトラクション法である。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、観測信号のパワースペクトルを $|X_i(t)|^2$, 推定した雑音のパワースペクトルを $|\tilde{N}_i|^2$ とし、雑音除去した音声のパワースペクトルを $|\tilde{S}_i(t)|$ とすると,

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

となる。これらの位相差を $|\theta_i(t)|$, 真の音声, 真の雑音のパワースペクトルを $|S_i(t)|$,$|N_i(t)|$ とすると、

$$|X_i(t)|^2 = |S_i(t)|^2 + |N_i(t)|^2 + 2|S_i(t)||N_i(t)|cos\theta_i(t) \tag{2}$$

の様に関係を表現できる。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、最後の項の期待値が 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い。ここで、スペクトルの時間方向のスムージング

$$\overline{|X_i(t)|^2} = \sum_r \beta_r |X_i(t-\tau)|^2 \tag{3}$$

を考える。ただし、$\tau = 0, 1, \cdots, T-1, \sum_r \beta_r = 1$ である。式 (3) の右辺に式 (2) を代入すると

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

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、右辺の第一項が真の音声、第二項が真の雑音に近似すると考えられる。よって T が大きければ式 (4) の第三項が 0 に近い値と仮定でき、

$$\overline{|X_i(t)|^2} \approx |S_i(t)|^2 + |N_i(t)|^2 \tag{5}$$

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、

$$|\tilde{S}_i(t)|^2 \approx |S_i(t)|^2 + |N_i(t)|^2 - \alpha|\tilde{N}_i|^2 \tag{6}$$

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 [4]

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

$$x_i(t) = s_i(t) + n_i(t) \tag{7}$$

このとき、式 (7) は Toeplitz 行列を用いて、以下の様に表すことができる。

$$X_i = S_i + N_i \tag{8}$$

X_i に対して SVD を適用することにより $X_i = U_i\Sigma_i V_i^T$ 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が得られ、各々が独立しているとすると

$$\sigma_m^{X_i} = \sigma_m^{S_i} + \sigma_m^{N_i} \tag{9}$$

となる。上式において、$n_i(t)$ が白色性の雑音であれば、$\sigma_m^{S_i}$ は下式のように推定できる。

$$\hat{\sigma}_m^{S_i} = \sigma_m^{X_i} - \bar{\sigma}^{N_i} \tag{10}$$

推定された $\hat{\sigma}^{S_i}$ を用いて Toeplitz 行列 $\hat{S}_i$ は

$$\hat{S}_i = U_i W_i \Sigma_i V_i^T \tag{11}$$

$$W_i = diag\left(\frac{\sigma_m^{X_i} - \bar{\sigma}^{N_i}}{\sigma_m^{X_i}}\right) \tag{12}$$

のように表される。式 (12) において、音声成分の特異値 $\sigma_m^{S_i}$ は次元 R 以上の高次元で消失すると仮定する。このことから、高次元の特異値は雑音成分に近似すると仮定できる。

$$\sigma_m^{N_i} \cong \sigma_m^{X_i} \quad (m \leqq R) \tag{13}$$

以上のことをまとめると、

$$\bar{\sigma}^{N_i} = \frac{1}{M-R}\sum_{m=R}^{M-1} \sigma_m^{X_i} \tag{14}$$

のように雑音の特異値平均を推定できる。

2.3 GMM に基づく音声信号推定 [4]

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の出力の対数値を要素に持つ J 次元ベクトルを $X_{\log}(i), S_{\log}(i), N(i)$ とする。

$X_{\log}(i)$ は以下のようにして表される。

$$\begin{aligned} X_{\log}(i) &= \log[\exp(S_{\log}(i)) + \exp(N_{\log}(i)))] \\ &= S_{\log}(i) + \log[1 + \exp(N_{\log}(i) - S_{\log}(i))] \\ &= S_{\log}(i) + G_{\log}(i) \end{aligned} \tag{15}$$

$$G_{\log}(i) = \log[1 + \exp(N_{\log}(i) - S_{\log}(i))] \tag{16}$$

雑音重畳音声と音声のミスマッチ成分 $G_{\log}(i)$ とする。$S_{\log}(i)$ の K 混合 GMM で $G_{\log}(i)$ の期待値を推定する。

$$p(S_{\log}(i)) = \sum_{k=1}^{K} P(k) N(S_{\log}(i); \mu_{S,k}, \sigma_{S,k}) \tag{17}$$

上式において、$p(S_{\log}(i))$ は $S_{\log}(i)$ の出現確率である。また、$P(k)$, $\mu_{S,k}$, $\sigma_{S,k}$ はそれぞれ要素分布 k における混合重み、平均ベクトル、対角共分散行列である。$S_{\log}(i)$ のGMMが与えられたときに、$X_{\log}(i)$ を $S_{\log}(i)$ と同じように、K混合GMMを用いてモデル化する。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 $N(i)$ の平均ベクトルを μ とする。$X_{\log}(i)$ のGMMの要素分布 k の平均ベクトルは

$$\begin{aligned}\mu_{X,k} &\simeq \mu_{S,k} + \log[1+\exp(\mu_N - \mu_{S,k})] \\ &= \mu_{S,k} + \mu_{G,k} \end{aligned} \tag{18}$$

と近似できる。また、対角共分散行列 $\Sigma_{X,k}$ は、

$$\Sigma_{X,k} \simeq \Sigma_{S,k} \tag{19}$$

として近似する。$\mu_{G,k}$ は要素 k における雑音成分の平均ベクトルに相当し、以下の式のように $X_{\log(i)}$ の事後確率 $P(k|X_{\log}(i))$ 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 $G_{\log}$ の推定値 $\hat{G}_{\log}(i)$ を推定する。

$$\hat{G}_{\log}(i) = \sum_{k=1}^{K} P(k|X_{\log}(i))\mu_{G,k} \tag{20}$$

$$P(k|X(i)) = \frac{P(k)N(X_{\log}(i);\mu_{X,k},\Sigma_{X,k})}{\sum_{k'=1}^{K} P(k')N(X(i);\mu_{X,k'},\Sigma_{X,k'})} \tag{21}$$

得られた $\hat{G}_{\log}(i)$ を用いて $S_{\log}(i)$ の推定値 $S'_{\log}(i)$ は

$$S_{\log}(i) = X_{\log}(i) - \hat{S}_{\log}(i) \tag{22}$$

2.4 ピッチ同期ＫＬＴ [5]

ピッチ同期ＫＬＴ（以下ＫＬＴ）では、クリーン音声 $s(tK+i)$ 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。$i=1,\cdots,L$（L:フレーム長）、K=1 ピッチ内のサンプル数である。ここで

$$\begin{aligned}S_p(t,i) = (&s((t-T-1)K+i), \\ &\ldots, s((t+T-1)K+i))^T\end{aligned} \tag{23}$$

となる。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。

$$\hat{S}_p = HX_p \tag{24}$$

このとき推定誤差は次式で表される

$$r = \hat{S}_p - S_p = (H-I)S_p + HN_p = r_S + r_n \tag{25}$$

ここで、r_S は音声の歪み、r_n は残留雑音である。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。

$$\overline{\varepsilon_S^2} = trE\{r_S r_S^T\} = tr\{(H-I)R_S(H-I)^T\} \tag{26}$$

$$\overline{\varepsilon_n^2} = trE\{r_n r_n^T\} = tr\{HR_nH^T\} \tag{27}$$

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、

$$\min_H \overline{\varepsilon_S^2},\ subject\ to: \frac{1}{K}\min_H \overline{\varepsilon_n^2} \leqq \sigma_n^2 \tag{28}$$

によって求める。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。

$$L_H(H,\mu) = \overline{\varepsilon_S^2} + \mu(\overline{\varepsilon_n^2} - K\sigma_n^2) \tag{29}$$

$$\mu(\overline{\varepsilon_n^2} - K\sigma_n^2) = 0 \quad for\ \mu \geqq 0 \tag{30}$$

ここで、μ はラグランジュ乗数である。式 (26), 式 (27) より、

$$H = R_S(R_S + \mu R_n)^{-1} \tag{31}$$

を得る。R_S を次式のように固有値分解する。

$$R_S = U\Lambda_S U^T \tag{32}$$

式 (31) に式 (32) を代入し

$$H = U\Lambda_S(\Lambda_S + \mu U^T R_n U)^{-1}U^T \tag{33}$$

を得る。雑音を白色と仮定する。このとき、H は次式で表される。

$$H = UGU^T \tag{34}$$

ここで、

$$G = diag(g_t(1), g_t(2), \cdots, g_t(2T+1)) \tag{35}$$

$$g_t(i) = \lambda_S^i/(\lambda_S^i + \mu\lambda_n) \tag{36}$$

以上から、クリーン音声 $\hat{S}_p = HX_p$ を得る。

3. 実験環境

本研究の評価には雑音環境下連続日本語数字音声認識タスクの共通フレームワーク AURORA-2J [6] を使用した。認識音声は 1〜７桁の連続数字である。AURORA-2J では２つの音響モデル学習セット（クリーン/マルチ）と３つのテストセット（セットA/B/C）がある。

クリーン学習セットは雑音を重畳しない音声で全８４４０発話である。クリーン音声には雑音を重畳していないため雑音抑圧手法は適用しない。

マルチ学習ではクリーン音声、雑音重畳音声を併せて学習する。マルチ学習セットは４種類の雑音を５段階のＳＮＲ (clean,20dB,15dB,10dB,5dB) で重畳する。クリーン学習データと同じ全８４４０発話である。マルチ学習データにはテストデータと同様に雑音抑圧手法を適用する。

テストセットは雑音を７段階のＳＮＲ (clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB) で重畳する。テストセットAはマルチ学習で使用された雑音を重畳する。テストセットBは学習データで使用する雑音とは異なる４種類の雑音を重畳する。テストセットCはテストセット A,B それぞれの雑音にさらに乗算性雑音が加わる。合計でテスト音声は全７００７０発話となる。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。分析条件は 16kHz サンプリング, フレーム周期１０ ms,

表 1　各雑音条件で最適な手法を手動で選択した場合の 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 2　GMM-KLT 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 3　SS 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

２５ msハミング窓である。各数字モデルの状態数は１８であり、各状態の混合数は２０である。評価の際、音響モデルはHMMを使用し、学習、認識時に３９次元の音声特徴量 (ＭＦＣＣ１２次元＋対数パワーとそのΔ, $\Delta\Delta$) を使用した。

4. 雑音抑圧手法自動選択

4.1 雑音抑圧手法の可能性

山田ら [7] は、同じ手法の組合せにおいて、雑音種類毎に最適な抑圧手法が異なり、最適な手法を選ぶことで認識率が向上できる可能性を示している。まず、各認識結果を雑音条件別（雑音種類、ＳＮＲ）にそれぞれ比較し、最も高い認識率とそれをもたらす雑音抑圧手法を選択する。評価は誤り改善率で行う。誤り改善率は比較対象手法の誤り改善率であり、比較対象手法の単語正解精度を X_m[%]、ベースラインの単語正解精度を X_b[%] とすると、

$$\text{誤り改善率} = \frac{X_m - X_b}{100.0 - X_b} \times 100\ \% \qquad (37)$$

のように定義できる。表 1-3 はモデルの認識性能を示す。表の上段はクリーン学習, 下段はマルチ学習の結果である。表中のA, B, Cはそれぞれテストセットに対応する。セットA, Bは４種類、セットCは２種類の雑音を重畳した音声である。表中の値はテストセット内全雑音種類においてSNR20dB～0dBまでの誤り改善率の平均値である。各雑音条件におけるテストデータに最適な手法を選択した場合のクリーン学習、マルチ学習における誤り改善率を表１に示す。また、クリーン学習の場合に

表 4　学習データの雑音条件

	Subway	Babble	Car	Exhibition
Clean				
20 dB				
15 dB				
10 dB				
5 dB				

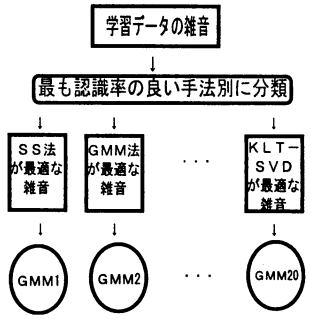

図 1　学習時：各抑圧手法別の GMM を学習

誤り改善率が最も高かった手法 (GMM-KLT) の誤り改善率を表 2, マルチ学習の場合に誤り改善率が最も高かった手法 (SS) を表 3 に示す。各性能を比較すると手法を単独で適用する場合より、雑音条件において最適な手法を選択する場合が性能向上を得る。

4.2 雑音抑圧手法自動選択に基づく音声認識

学習データの各雑音条件を表 4 に示す。4 種類の雑音を 5 段階のＳＮＲで重畳していることを示している。表中の 1 要素が雑音条件となり、全 20 種類となる。4.1 節の結果から、各雑音条件に最適な抑圧手法を表 4 にあてはめる。

雑音重畳音声に適用する抑圧手法はＧＭＭを用いて判定する。

図１にＧＭＭ学習時の流れ図を示す。本実験では各音声ファイルの開始 10 フレームは雑音のみがあるとして、その部分を用いた。学習データの各雑音条件に対応する雑音を最適な抑圧手法別に手動で分類する。分類した雑音の音声特徴量で各抑圧手法別のＧＭＭを学習する。ＧＭＭは雑音抑圧手法の単独（４種類）、複数適用合わせた場合（４×４）の計２０種類に、抑圧手法を適用しない条件を加え全２１種類が作成される。

認識時には、入力音声直前の雑音を、各抑圧手法別のＧＭＭに入力する。各ＧＭＭから出力された入力雑音に対する尤度を比較し、入力雑音の雑音条件を判定する。最も高い尤度を出力したＧＭＭに対応する抑圧手法を選択し、選択した抑圧手法を

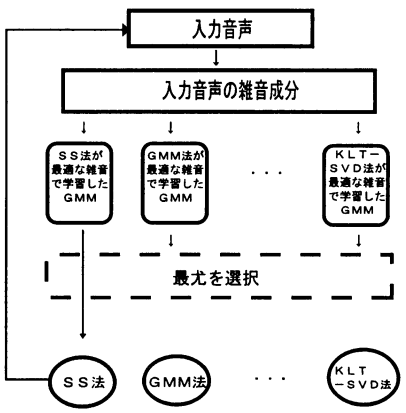

図 2　認識時：(例) 入力雑音に対し SS 法が最適と判定した場合の認識

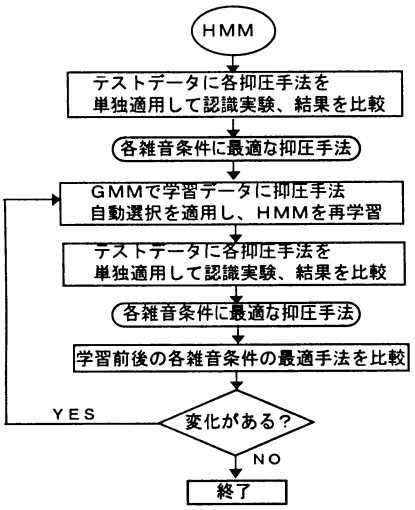

図３　反復学習アルゴリズムの流れ

入力音声に適用し認識を行う (図2)。GMMは対角共分散のガウス分布の６４混合である。学習、認識の際に学習、認識データから抽出する雑音成分は音声の直前の１０フレームを用い、分析条件は１３次元の音声特徴量 (ＭＦＣＣ１２次元＋対数パワー) を使用した。図２は入力雑音に最適な抑圧手法がＳＳ法であると判定された場合の処理の流れである。自動選択された抑圧手法を用いて雑音抑圧を行った入力音声を認識に用いる。このGMMを用いる手法のメリットはGMMにとって未知の雑音に対してもある程度適切な抑圧手法を自動選択できることである。抑圧手法を判定するためのGMMは学習データの雑音成分のみで学習するため、学習データにない入力雑音はGMMにとって未知雑音となる。GMMは未知雑音に対し特徴が似ている既知雑音として判定し、抑圧手法自動適用を行うので、未知の雑音に対してもある程度適切な雑音抑圧を実現できると期待される。

4.3　音響モデルの反復学習

4.3.1　反復学習の可能性

提案手法は、フロントエンドで特徴量を補正するための方法である。クリーン学習では学習データに抑圧手法を適用せずテストデータのみに適用するため、雑音条件毎に最適な抑圧手法が異なっても音響モデルに変化はない。しかし、マルチ学習では最適な抑圧手法を選択して適用した音声を用いて音響モデルを作り直すことができる。これにより認識率の向上が期待でいる一方で雑音条件ごとの最適な手法が異なってくる可能性もある。そこで、提案手法を適用した音響モデルの各雑音条件に最適な抑圧手法を再調査し、再度GMMを作成、提案手法を適用しモデルを再学習する。この操作を繰り返し、学習前後の雑音条件毎の最適な抑圧手法に変化がなくなるまで繰り返す。すなわち、GMMの識別目標と音響モデルが合致して認識精度が向上すると考えられる。

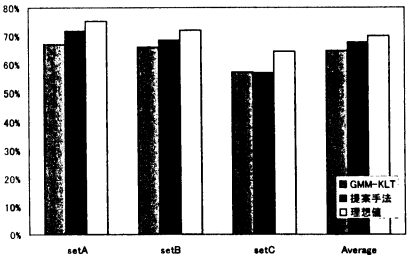

図４　クリーン学習における誤り改善率（%）

4.3.2　反復学習アルゴリズム

反復学習の流れを図３に示す。最初に、雑音抑圧手法を単独適用、もしくは雑音抑圧手法自動選択を適用した音響モデル (以下HMM) を準備し初期モデルとする。このHMMを用いて抑圧手法単独適用による認識実験を行い、各雑音条件に最適な抑圧手法を調査する。その結果から、抑圧手法別GMMを作成し抑圧手法自動選択を学習データに適用しHMMを学習する。再学習したHMMを用いて再び認識実験を行い、各雑音条件に最適な抑圧手法を調査する。学習前後の結果を比較し変化があれば再学習したHMMを先ほどの初期モデルと同様に使用して同様の作業を行う。学習前後の変化が完全になくなったときに反復学習を終了する。

5.　認識実験

４節の手法を認識実験によって評価する。なお、本手法は全てフロントエンド処理であり、特徴量の条件などもカテゴリー０[6] の条件を満たす。

5.1　クリーン学習モデルによる認識実験結果

入力音声に３つの抑圧手法を適用して認識実験を行った。１つ目は抑圧手法の単独適用において認識精度が最も高い手法 (GMM-KLT)、２つ目は提案手法を適用する。３つ目は各雑音条件で最適な抑圧手法を手動で選択し適用する。これは本実験における理想値となる。認識実験結果を図４に示す。図中では順に"GMM-KLT","提案手法","理想値"と記す。

評価は前述した誤り改善率を用いる。認識実験の結果から、全体的に"GMM-KLT"より認識精度が改善していることがわかる。まず、ベースライン（雑音抑圧手法を用いない場合）と比較して誤り改善率67.7%の性能改善を達成した。"GMM-KLT"と比較して誤り改善率で3%の精度改善が得られ単独手法適用よりも高い精度を得ることが確認できる。

既知雑音を重畳したテストセットＡの認識精度の改善と比較して、未知雑音を重畳したテストセットＢの改善は劣っていない。すなわち、未知雑音に対しても提案手法は頑健に動作し認識精度改善をもたらしていることがわかる。

5.2　マルチ学習モデルによる認識実験結果

マルチ学習における認識実験では、単独で平均単語正解精度が最も高い"SVD-GMM"を初期モデルとして反復学習を行ったモデルを用いた。反復学習を繰り返す毎に、そのモデルに対する雑音条件毎に最適な手法の変化は減少し、繰り返し５回目

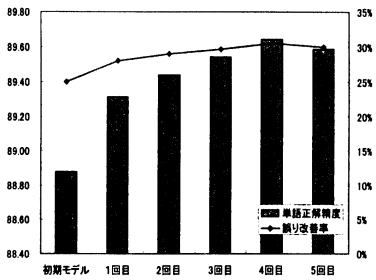

図 5 反復学習による認識精度の推移 (%)

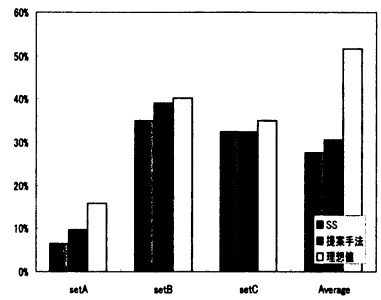

図 6 マルチ学習における誤り改善率 (%)

で変化はなくなった。認識実験は、反復学習の各繰り返しで得られた音響モデルを用いて行った。認識結果を図 5 に示す。棒グラフが単語正解精度、折れ線グラフが誤り改善率の推移を示している。繰り返しによって認識精度が改善し最高となった繰り返し 4 回目において単語正解精度、誤り改善率は最高となった。初期モデルの認識精度と比較して単語正解精度で 0.8%, 誤り改善率 5.4%の性能改善を達成している。手法は 5.1 節と同様である。マルチ学習において、最も高い誤り改善率をもたらす単独手法は"SS"であるため、これを使用する。また、提案手法の実験は、繰り返し 4 回目で得られた音響モデルを使用して行った。認識実験結果を図 6 に示す。これより、ベースラインと比較して誤り改善率 30.5%の性能改善を達成していることがわかる。"SS"と比較すると、誤り改善率で 2.9%の精度改善が見られ単独手法適用よりも高い精度を得ている。これらの結果から、提案手法は単独手法の適用よりも高い認識精度を得られることがわかる。テストセット別に考察するとクリーン学習時と同様に、未知雑音に対しても効果的な雑音抑圧を行い認識精度改善をもたらすことがわかる。

6. ま と め

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 つの雑音抑圧手法及びそれらの組合せた手法から、雑音条件に応じて GMM で自動的に手法を選択及び適用する手法について検討した。また、マルチ学習モデルの認識傾向のために反復学習手法を用い効果を考察した。本研究の評価用タスクとして雑音環境下連続日本語数字音声データベース AURORA-2J を使用した。その結果、GMM にとって未知の雑音に対しても効果的に雑音抑圧を行えることが示され、認識精度を改善できることが明らかとなった。今後は更なる認識精度向上のために GMM 学習データの雑音種類数の増加を試みる。

謝辞

本研究を進めるにあたって、各雑音抑圧手法のプログラムをご提供いただいた、武田一哉氏 (名古屋大学), 藤本雅清氏 (ATR) に深く感謝いたします。また多くの助言を頂いた山田武志氏 (筑波大学) に深く感謝いたします。本研究は情報処理学会 SIG-SLP 雑音下音声認識評価 WG の雑音下音声認識評価環境 (AURORA-2J) を使用して行いました。

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 SP 2002-12,2002.

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

[3] N.Kitaoka,S.Nakagawa,"Evaluation of spectral subtraction with smoothing of time direction on the AURORA2 task," Proc.ICSLP2002,pp.465-468,2002.

[4] 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.

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

[6] Satoshi Nakamura, Kazuya Takeda, Kazumasa Yamamoto, Takeshi Yamada, Shingo Kuroiwa, Norihide Kitaoka, Takanobu Nishiura, Akira Sasou, Mitsunori Mizumachi, Chiyomi Miyajima, Masakiyo Fujimoto, Toshiki Endo. "AURORA-2J: An Evaluation Framework for Japanese Noisy Speech Recognition," IEICE Trans. Inf. & Syst., (to appear), 2005.

[7] 山田武志, 岡田治郎, 武田一哉, 北岡教英, 藤本雅清, 黒岩眞吾, 山本一公, 西浦敬信, 水町光徳, 中村哲"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"情報処理学会研究報告,SLP-47-18,2003

[8] 松田繁樹, 實廣貴敏,Konstantin MARKOV, 中村哲 "雑音や発話スタイルの変動に頑健な日本語大語彙連続音声認識"情報処理学会研究報告 Vol.2004-SLP-50, pp.37-44, 2004

[9] 篠崎隆宏, 古井貞熙 "超並列デコーダによる話し言葉音声認識" 第 3 回話し言葉の科学と工学ワークショップ 講演予稿集,pp.67-72, 2004

[10] 伊田政樹, 中村哲,"雑音 GMM の適応化と SN 比別マルチパスモデルを用いた HMM 合成による高速な雑音環境適応化"電子情報通信学会論文誌,Vol.J86-D-II,No.2,pp.195-203,2003 年 2 月

[11] 岡田治郎, 山田武志, 北脇信彦 "複数の雑音抑圧手法による認識結果の統合の検討" 日本音響学会春季研究発表会 pp.157-158, 2004 年 3 月

[12] Takashi Fukuda and Tsuneo Nitta, "Canonicalization of Feature Parameters for Automatic Speech Recognition", Proc. ICSLP-2004, Vol. 4, pp.2537-2540, 2004.