Vol.2012-SLP-94 No.23
2012/12/21

特徴量強調における教師なし話者適応に関する検討

グェンドゥックズイ[1,2]　吉岡拓也[1]　峯松信明[2]　広瀬啓吉[2]

概要：近年，音声認識技術は様々なアプリケーションで使用されている．しかし，録音環境に含まれる雑音や残響等の音響的な歪みにより認識性能が大幅に低下する．この問題の解決策として，クリーン音声の GMM を用いて観測音声の特徴量から音響的歪みの影響を取り除く特徴量強調技術が知られている．一方，モバイルデバイスへの音声入力に代表される最近のアプリケーションの多くでは，多様な環境で録られた認識対象個人の音声データを蓄積しておくことが容易にできる．しかしながら，こうした個人データをどのように扱えば特徴量強調を含む認識システム全体の性能を効果的に向上できるかは明らかでない．本研究では，特徴量強調に用いるクリーン音声 GMM の MAP 適応と音声認識に用いる音響モデルの MLLR 適応のいくつかの組み合わせ方について，その効果を実験的に比較検討する．

キーワード：特徴量強調，VTS，MAP 適応，MLLR 適応，話者適応

A Study on Unsupervised Speaker Adaptation for Feature Enhancement

Duy Nguyen Duc[1,2]　Takuya Yoshioka[1]　Nobuaki Minematsu[2]　Keikichi Hirose[2]

Abstract: Speech recognition has been an active research area for many years, and nowadays, it is being used in many practical applications. However, the recognition performance is often seriously degraded by noise and reverberation present in recording environments. One promising approach to solve this problem is feature enhancement, which attempts to restore clean feature vectors using a GMM of clean speech. Meanwhile, in many recent applications including those to mobile devices, it is easy to collect target user's speech data recorded in various environments. However, how to exploit these data for improving the performance of a speech recognizer that performs feature enhancement is an open question. This study experimentally compares different methods for combining MAP adaptation of the clean speech GMM and MLLR adaptation of the recognizer's acoustic model.

Keywords: Feature enhancement, VTS, MAP adaptation, MLLR adaptation, Speaker adaptation

1. はじめに

近年，音声認識技術はモバイルデバイスへの入力，カーナビゲーションシステム，ディクテーション等の多様なシーン，デバイスで使われている．しかし，音響モデルの学習に用いる音声データと認識対象音声では録音環境や話者が異なるため，認識対象音声と音響モデルの間にはミスマッチが生じ，その結果認識精度が大幅に低下するという問題がある．録音環境及び話者の相違の影響を低減するために，様々な手法が提案されている．

録音環境の相違を補正する手段として，特徴量強調と呼ばれるアプローチが知られている．特徴量強調は，観測された劣化音声の特徴量からクリーン音声の特徴量へ変換する手法であり，その代表例として Vector Taylor Series (VTS)[1], [2] が挙げられる．この手法では，まずクリーン音声の特徴量の分布を混合正規分布（GMM: Gaussian Mixture Model）でモデル化し，事前に学習しておく．次

[1] NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation
[2] 東京大学情報理工学系研究科
Graduate School of Information and Technology, The University of Tokyo

に，劣化音声の特徴量から推定された歪み過程（クリーン特徴量から劣化特徴量が生成される過程）のモデルと学習しておいたクリーン音声の GMM を用いて，劣化特徴量をクリーン特徴量に変換する．

一方，話者の多様性に対処するために，話者適応化技術が古くから研究されている．話者適応化は認識対象話者個人の音声データ（適応データ）を用いて，様々な話者の音声から学習された不特定話者音響モデルのパラメータを適応データに適合するように変更する，つまり音響モデルを認識対象話者に適応させる技術である．話者適応化技術の代表例として最大事後確率 (MAP: Maximum A Posteriori) 推定による適応（MAP 適応）[3], [4] や最尤線形回帰 (MLLR: Maximum Likelihood Linear Regression) による適応（MLLR 適応）[5], [6] が広く知られている．

多くのアプリケーションでは，録音環境と話者の違いの 2 つの問題は同時に存在する．例えば，モバイルデバイスへの音声入力では，録音環境は多様で予期できない一方，ユーザがデバイスを使用する度に当該ユーザ個人の音声データを蓄積することができる．このような状況では，話者適応化を音響モデルだけでなく特徴量強調用のクリーン GMM にも実施することが考えられるので，さらなる認識精度の向上が見込まれる．そこで，本研究では，クリーン GMM の MAP 適応と音響モデルの MLLR 適応のいくつかの組み合わせ方について，認識システム全体の性能の観点から実験的に比較検討する．

本稿では，2 章で VTS 特徴量強調および雑音モデルの推定方法を説明する．3 章で話者適応化技術である MAP 適応及び MLLR 適応を簡単に解説する．4 章ではいくつかの実験とその結果を示す．最後に，5 章で本研究を通じて得られた知見をまとめる．

2. 特徴量強調

特徴量強調の目的は，劣化音声の特徴量 $\boldsymbol{x}_t$ が与えられた時，クリーン音声の特徴量 $\boldsymbol{s}_t$ を推定することである．ただし，t は短時間フレームのインデクスを表す．クリーン特徴量の推定値は通常，クリーン特徴量の事後確率分布の平均 $E(\boldsymbol{s}_t|\boldsymbol{x}_t)$ として計算される．その計算方法の一つとして，以下に述べる VTS が知られている．

2.1 VTS に基づく特徴量強調

VTS では，クリーン特徴量の事後確率分布を，次で説明するクリーン特徴量と劣化過程の 2 種類のモデルを用いて求める．クリーン特徴量モデルは，以下のような対角共分散行列を持つ GMM で表現される：

$$p(\boldsymbol{s}_t) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{s}_t; \boldsymbol{\mu}_k^s, \boldsymbol{\Sigma}_k^s) \tag{1}$$

ここで，K は GMM の混合数，π_k は重み，$N(\boldsymbol{s}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均 $\boldsymbol{\mu}$，分散 $\boldsymbol{\Sigma}$ を持つ正規分布の確率密度関数である．これらのパラメータはクリーン音声のコーパスから学習される．

劣化過程モデルは，クリーン特徴量 $\boldsymbol{s}$ から劣化特徴量 $\boldsymbol{x}$ が生成される過程をモデル化したものである．加法性雑音特徴量を $\boldsymbol{r}_t$，乗法性雑音特徴量（時不変）を $\boldsymbol{h}$ とする．特徴量として対数メル周波数スペクトルを用いる場合，劣化過程モデルは以下のように定義される：

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{s}_t + \boldsymbol{h} + \log(1 + \exp(\boldsymbol{r}_t - \boldsymbol{s}_t - \boldsymbol{h})) && (2)\\ &= \boldsymbol{s}_t + \boldsymbol{h} + f(\boldsymbol{s}_t, \boldsymbol{r}_t, \boldsymbol{h}) && (3)\end{aligned}$$

関数 f はミスマッチ関数と呼ばれる．加法性雑音 $\boldsymbol{r}_t$ は次のような正規分布でモデル化される．

$$\boldsymbol{r}_t = N(\boldsymbol{r}_t; \boldsymbol{\mu}^r, \boldsymbol{\Sigma}^r) \tag{4}$$

多くの従来研究では，平均 $\boldsymbol{\mu}^r$ と共分散行列 $\boldsymbol{\Sigma}^r$ は時間に依存しないと仮定される．ただし，雑音が非定常な場合，平均 $\boldsymbol{\mu}^r$，もしくは平均と共分散行列 $\boldsymbol{\Sigma}^r$ を t に依存させたほうが高い認識性能が得られる（2.2 節を参照）．劣化過程モデルを記述するパラメータ $\{\boldsymbol{\mu}^r, \boldsymbol{\Sigma}^r, \boldsymbol{h}\}$ は劣化音声特徴量から直接推定される．

以上で述べたクリーン特徴量モデルと劣化過程モデルに基づいて，劣化特徴量が与えられた時のクリーン特徴量は次のように推定される．まず，ミスマッチ関数 $f(\boldsymbol{s}_t, \boldsymbol{r}_t, \boldsymbol{h})$ を VTS によって以下のように近似する．

$$\begin{aligned} f(\boldsymbol{s}_t, \boldsymbol{r}_t, \boldsymbol{h}) \simeq\ & f(\boldsymbol{s}_0, \boldsymbol{r}_0, \boldsymbol{h}_0) + f_s(\boldsymbol{s}_0, \boldsymbol{r}_0, \boldsymbol{h}_0)(\boldsymbol{s}_t - \boldsymbol{s}_0) \\ &+ f_r(\boldsymbol{s}_0, \boldsymbol{r}_0, \boldsymbol{h}_0)(\boldsymbol{r}_t - \boldsymbol{r}_0) \\ &+ f_h(\boldsymbol{s}_0, \boldsymbol{r}_0, \boldsymbol{h}_0)(\boldsymbol{h} - \boldsymbol{h}_0) \end{aligned} \tag{5}$$

ここで，f_s，f_r，f_h はそれぞれミスマッチ関数 $f(\boldsymbol{s}, \boldsymbol{r}, \boldsymbol{h})$ の $\boldsymbol{s}$，$\boldsymbol{r}$，$\boldsymbol{h}$ に関する偏導関数，$(\boldsymbol{s}_0, \boldsymbol{r}_0, \boldsymbol{h}_0)$ はテーラー展開の中心である．これを用いて，劣化特徴量の GMM を次式のように計算できる．

$$p(\boldsymbol{x}_t) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_t|k) \tag{6}$$

$$p(\boldsymbol{x}_t|k) = N(\boldsymbol{x}_t; \boldsymbol{\mu}_k^x, \boldsymbol{\Sigma}_k^x) \tag{7}$$

但し，

$$\boldsymbol{\mu}_k^x \simeq \boldsymbol{\mu}_k^s + \boldsymbol{h} + f(\boldsymbol{\mu}_k^s, \boldsymbol{\mu}^r, \boldsymbol{h}) \tag{8}$$

$$\begin{aligned} \boldsymbol{\Sigma}_k^x \simeq\ & f_s(\boldsymbol{\mu}_k^s, \boldsymbol{\mu}^r, \boldsymbol{h}) \Sigma_k^s f_s(\boldsymbol{\mu}_k^s, \boldsymbol{\mu}^r, \boldsymbol{h})^T \\ &+ f_r(\boldsymbol{\mu}_k^s, \boldsymbol{\mu}^r, \boldsymbol{h}) \Sigma^r f_r(\boldsymbol{\mu}_k^s, \boldsymbol{\mu}^r, \boldsymbol{h})^T \end{aligned} \tag{9}$$

上記で近似した劣化特徴量の GMM を用いて，次のようにクリーン特徴量を推定する．

$$\begin{aligned} \hat{\boldsymbol{s}}_t &= E(\boldsymbol{s}_t|\boldsymbol{x}_t) && (10)\\ &= \sum_{k=1}^{K} (\boldsymbol{x}_t - f(\boldsymbol{\mu}_k^s, \boldsymbol{\mu}^r, \boldsymbol{h})) p(k|\boldsymbol{x}_t) && (11)\end{aligned}$$

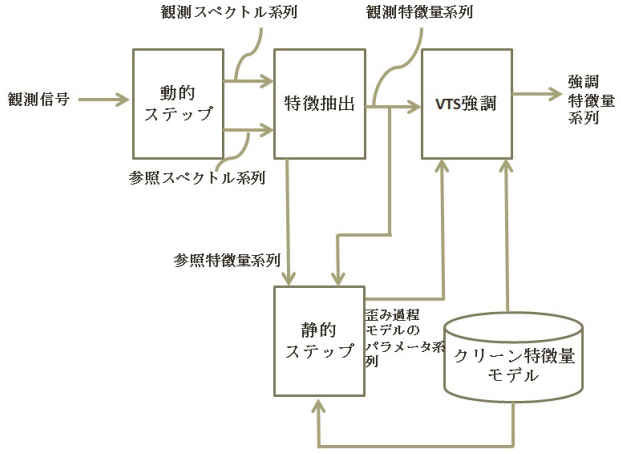

図 1 動的静的アプローチによる特徴量強調.
Fig. 1 Feature enhancement using dynamic-static approach.

但し，$p(k|\boldsymbol{x}_t)$ は $p(k|\boldsymbol{x}_t) = p(\boldsymbol{x}_t|k)p(k)/p(\boldsymbol{x}_t)$ として計算される．

2.2 歪み過程モデルパラメータの推定

歪み過程モデルのパラメータの推定方法としていくつかの手法が知られているが，本研究では非定常雑音に効果的な動的静的アプローチを用いる [7]．この手法では，加法性雑音の平均は時間に応じて変化する，すなわち短時間フレーム t に依存して $\boldsymbol{\mu}_t$ と表されるものとする．この手法では，特徴量領域よりもパワースペクトル領域での信号表現を用いたほうが雑音の変化を捉えやすいことに着目し，加法性雑音スペクトルを暫定的に推定した後（動的ステップ），これを特徴量領域に尤度規準で最適変換する（静的ステップ）．

動的静的アプローチによる特徴量強調の流れを**図 1** に示す．まず，動的ステップで，劣化音声から加法性雑音のスペクトルを暫定的に推定し（参照スペクトルとする），同時に観測スペクトルも求める．次に，観測スペクトルと参照スペクトルから特徴量を抽出し，得られたものをそれぞれ観測特徴量 $\boldsymbol{x}_t$ と参照特徴量 $\hat{\boldsymbol{r}}_t$ とする．静的ステップでは，予め学習したクリーン GMM と 2 種類の特徴量系列 $(\boldsymbol{x}_t)_t, (\hat{\boldsymbol{r}}_t)_t$ を用いて，加法性雑音特徴量の平均系列 $(\boldsymbol{\mu}_t^r)_t$，共分散行列 $\boldsymbol{\Sigma}^r$，および乗法性雑音特徴量 $\boldsymbol{h}$ を最尤推定する．具体的には，推定すべきパラメータ数を減らす目的で，平均 $\boldsymbol{\mu}_t^r$ と参照特徴量 $\hat{\boldsymbol{r}}_t$ の間に次式の関係を仮定する．

$$\boldsymbol{\mu}_t = \hat{\boldsymbol{r}}_t + \boldsymbol{b} \tag{12}$$

その上で，バイアス $\boldsymbol{b}$ を $\boldsymbol{\Sigma}$，$\boldsymbol{h}$ とともに推定する．この推定には，二重 EM アルゴリズム [8], [9] の変種が用いられる．詳しいアルゴリズムについては [7] の 3 節を参照されたい．最後に，VTS 強調ステップでは，クリーン GMM と静的ステップで推定した歪み過程モデルのパラメータを用いて，強調特徴量を求める．

3. 話者適応化

本章では，3.1 節で一般的な GMM の MAP 適法方法について説明する．3.2 節では，音響モデルの MLLR 適応について説明する．

3.1 GMM の MAP 適応

MAP 推定法では，適応データ $X = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ が与えられた時，次式に示すように，モデルパラメータ θ の事後確率密度関数を最大にする点を求める．

$$\hat{\theta} = \arg\max_{\theta} p(\theta|X) \tag{13}$$
$$= \arg\max_{\theta} p(X|\theta)p(\theta) \tag{14}$$

ただし，$p(X|\theta)$ と $p(\theta)$ は，それぞれパラメータ θ の尤度関数と事前確率密度関数である．GMM を用いる場合，尤度関数 $p(X|\theta)$ は次式で与えられる．

$$p(X|\theta) = \prod_{t=1}^{n}\sum_{k=1}^{K}\omega_k N(x_t; m_k, r_k) \tag{15}$$

モデルパラメータ $\theta = \{\omega_1, ..., \omega_K; m_1, ..., m_K; r_1, ..., r_K\}$ の事前分布は次式で与えられる．

$$p(\theta) = \prod_{k=1}^{K}\omega_k{}^{\nu_k-1}|r_k|^{\frac{(\alpha_k-p)}{2}}\exp[-\frac{\mathrm{tr}(u_k r_k)}{2}] \times \exp[-\frac{\tau_k}{2}(m_k-\mu_k)^T r_k(m_k-\mu_k)] \tag{16}$$

$\{\tau_k, \mu_k, \alpha_k, u_k, \nu_k\}$ は事前分布を規定する超パラメータである．適応後のパラメータ θ の値は，EM アルゴリズムを用いて推定される．EM アルゴリズムは，(20) で与えられる E-step と (17–19) で構成される M-step を，収束するまで交互に繰り返し実行する．

$$\omega_k^{'} = \frac{\nu_k - 1 + \sum_{t=1}^{n} c_{kt}}{n - K + \sum_{k=1}^{K}\nu_k} \tag{17}$$
$$m_k^{'} = \frac{\tau_k\mu_k + \sum_{t=1}^{n} c_{kt}\nu_k}{\tau_k + \sum_{t=1}^{n} c_{kt}} \tag{18}$$
$$r_k^{'-1} = \frac{u_k + \tau_k(\mu_k - m_k^{'})(\mu_k - m_k^{'})^T}{\alpha_k - p + \sum_{t=1}^{n} c_{kt}} + \frac{\sum_{t=1}^{n} c_{kt}(x_t - m_k^{'})(x_t - m_k^{'})^T}{\alpha_k - p + \sum_{t=1}^{n} c_{kt}} \tag{19}$$
$$c_{kt} = \frac{\omega_k f(x_t|\theta_k)}{f(x_t|\theta)} \tag{20}$$

3.2 音響モデルの MLLR 適応

MLLR では，音響モデルを構成する隠れマルコフモデル（HMM）の各ガウス分布の平均ベクトルや共分散行列を，共通のアフィン変換行列を用いて変換することで，音響モデルを入力音声データに適応させる．本稿で述べる実験では，平均ベクトルのみを適応させた．すなわち，行列 $\boldsymbol{A}$ と

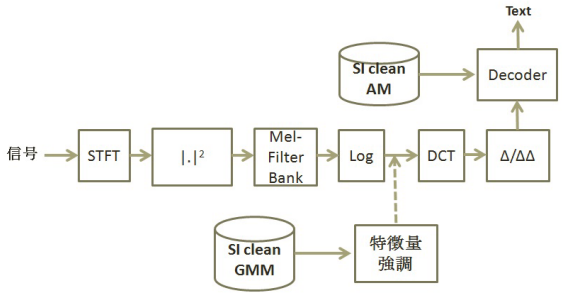

図 2 ベースラインの音声認識システム
Fig. 2 Baseline speech recognition system

ベクトル $\boldsymbol{b}$ を用いて，各正規分布の平均ベクトルを次式によって更新する．

$$\hat{\boldsymbol{\mu}}_k = \boldsymbol{A}\boldsymbol{\mu}_k + \boldsymbol{b} \tag{21}$$

但し，$\boldsymbol{\mu}_k$ と $\hat{\boldsymbol{\mu}}_k$ はそれぞれ適応前，適応後の平均ベクトルである．式中の変換行列 $\boldsymbol{A}$ とベクトル $\boldsymbol{b}$ は最尤推定により求められる．詳しくは [5], [6] を参照されたい．

4. 実験

本研究では，VTS 特徴量強調を含む音声認識システムをベースラインとして用いる．このシステムの構成を**図 2** に示す．このシステムでは，クリーン音声のコーパスから学習された不特定話者クリーン GMM (SI-clean GMM) と不特定話者クリーン音響モデル (SI-clean AM) を用いる．劣化音声が入力されると，その対数メルフィルタバンク特徴量に対して SI-clean GMM を用いた VTS 強調を行う．その結果得られた特徴量から MFCC とその $\Delta, \Delta\Delta$ 特徴量を抽出し，認識を行う．

本研究では，クリーン GMM と音響モデルの適応化方法として，それぞれ MAP 適応を MLLR 適応を用いる．各適応化処理をどのような条件で組み合わせればよいかについて，残響音声認識を例にとって実験的に検証したので，以下にその結果を報告する．

4.1 共通の実験条件

学習データとして英語発話コーパス WSJ（2 万語）を使う．評価データは異なる 8 話者の残響を含む発話（合計 333 発話）である．残響音声はテストセットに含まれるクリーン音声に会議室内で測定されたインパルス応答を畳み込んで作成した．本研究の実験では，残響時間が 0.78 秒であり，話者とマイクロホンとの距離が 2m である．

SI-clean GMM は WSJ の学習データから抽出された対数メルフィルタバンク特徴量を用いて，学習した．GMM の混合数は 256 とした．音声認識で用いる特徴量には，12 次元の MFCC（C1-C12）とその $\Delta, \Delta\Delta$ 特徴量，及び ΔC0, $\Delta\Delta$C0 の合計 38 次元のベクトルを用いた．

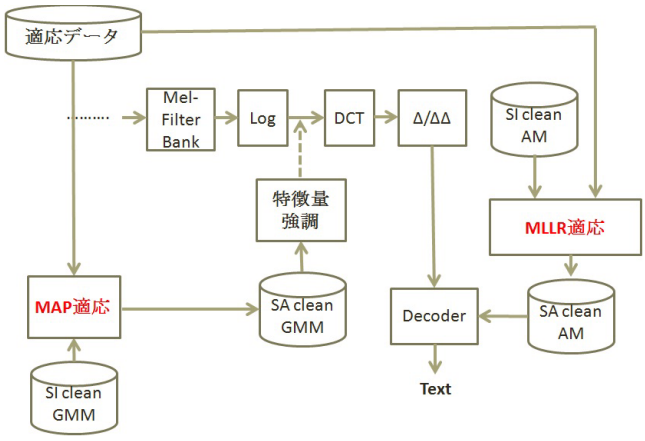

図 3 実験 1 の認識システム：同一のクリーン適応データを用いてクリーン GMM とクリーン音響モデルを適応させる．
Fig. 3 Speech recognition system used for Experiment 1, where clean GMM and clean acoustic model are adapted using the same clean data.

4.2 実験 1

この実験では，適応データとして，各話者の評価セットに含まれないクリーン音声（それぞれ 40 発話）を使い，クリーン GMM を MAP 適応，クリーン音響モデルを MLLR 適応する．この実験における認識システムは**図 3** のように構成される．本実験では，次の 4 つの条件について検討する．

(1) 適応なし (SI-GMM SI-AM：ベースライン)
(2) クリーン GMM のみ MAP 適応 (SA-GMM SI-AM)
(3) クリーン AM のみ MLLR 適応 (SI-GMM SA-AM)
(4) クリーン GMM を MAP 適応，かつクリーン音響モデルを MLLR 適応 (SA-GMM SA-AM)

実験結果を**表 1** に示す．クリーン GMM のみ適応させた場合，5.64%の認識誤りが削減された．これは，認識対象話者に適合したクリーン GMM を用いることで，特徴量強調性能が向上したことを示している．一方，クリーン音響モデルのみを適応させた場合，認識性能は逆に低下した．この結果は，クリーン音響モデルのみを話者適応させた場合，特徴量強調の結果得られる特徴量と適応された音響モデルのミスマッチがかえって大きくなる場合があることを示している．ミスマッチが大きくなる理由として，この条件では特徴量強調は不特定話者のデータから学習された GMM を用いて行われるため，強調された特徴量は認識対象話者とは異なる音響的な特性をもつことが考えられる．このことは，クリーン GMM とクリーン音響モデルの両方を適応させた場合の性能が最も高かった（6.51%の誤り削減率）ことからも示唆される．

4.3 実験 2

この実験では，クリーン GMM の適応データとして実験 1 で用いたデータを使う．一方，クリーン音響モデルの適

表 1　Word Error Rate (WER) of Experiment 1

	WER	WER relative improvement
SI-GMM SI-AM	54.93%	N/A
SA-GMM SI-AM	51.83%	5.64%
SI-GMM SA-AM	58.45%	-6.42%
SA-GMM SA-AM	51.35%	6.51%

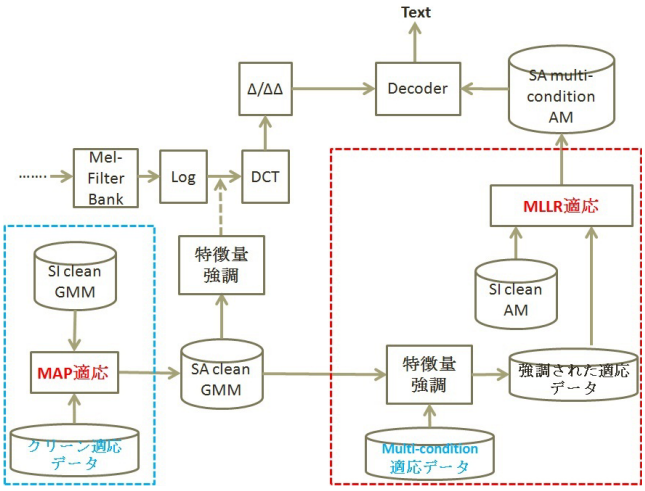

図 4　実験 2 の認識システム：クリーン適応データを用いてクリーン GMM を，マルチコンディション適応データを用いて音響モデルを適応させる．

Fig. 4　Speech recognition system used for Experiment 2. Clean GMM is adapted using speaker-specific clean data while clean acoustic model is adapted using speaker-specific multi-condition data.

応データはマルチコンディションデータを使う．すなわち，音響モデルは話者と環境の両方に適応される．それぞれの話者のマルチコンディション適応データはクリーン音声の 8 発話，4 種類の異なる残響環境における 32 発話の計 40 発話からなる．これらの適応データセットを使って，**図 4** に示す構成で実験を行った．まず，クリーン適応データを用いて，クリーン GMM（SI-clean GMM）の MAP 適応を行うと，話者適応化クリーン GMM（SA-clean GMM）が得られる．音響モデルの MLLR 適応の段階では，SA-clean GMM を使って，マルチコンディション適応データを強調する．強調して得られたデータを用いて，クリーン AM（SI-clean AM）に対して MLLR 適応を施す．この実験では，次の 3 つの条件について検討する．

(1) 適応なし (SI-GMM SI-AM)
(2) クリーン AM のみ MLLR 適応 (SI-GMM ES-AM)
(3) クリーン GMM を MAP 適応，クリーン音響モデルを MLLR 適応 (SA-GMM ES-AM)

実験結果を**表 2** に示す．クリーン GMM の MAP 適応をせずに，クリーン AM の MLLR 適応だけを行った場合，ベースラインシステムと比較して 4.89%の誤りが削減された．一方，クリーン GMM に対して話者適応をし，かつクリーン音響モデルに対して話者・環境適応を施すことによって，認識誤り率を 13.84%削減することが出来，大幅な改善が見られた．実験 1 の (SI-GMM SA-GMM) と (SA-GMM SA-GMM) の比較，及び実験 2 の (SI-GMM ES-GMM) と (SA-GMM ES-GMM) の比較から，特徴量強調を含むシステムにおいて音響モデルを適応させる場合，特徴量強調に用いる GMM も話者適応させることが重要であることが示された．

表 2　Word Error Rate (WER) of Experiment 2

	WER	WER relative improvement
SI-GMM SI-AM	54.93%	N/A
SA-GMM ES-AM	52.24%	4.89%
SA-GMM ES-AM	47.33%	13.84%

表 3　Word Error Rate (WER) of Experiment 3

	WER	WER relative improvement
SI-GMM SI-AM	54.93%	N/A
SA-GMM ES-AM	48.78%	11.20%

4.4　実験 3

本実験では，適応データとして実験 2 で用いたマルチコンディション適応データのみを使う．この適応データセットを使って，**図 5** に示す構成で実験を行った．音声や雑音は非定常な信号なので，マルチコンディションデータはクリーンに近いフレームを多く含む．この実験では，このようなフレームのみを用いてクリーン GMM を適応させる．ただし，本稿の範囲では，クリーンな発話を手動でマルチコンディションデータから選択した．すなわち，マルチコンディション適応データセットからクリーン音声データだけを抽出し，それを用いてクリーン GMM を MAP 適応させる．MAP 適応を行って得られた話者適応化クリーン GMM（SA-clean GMM）を用いて，マルチコンディション適応データを VTS 強調する．強調されたデータを使って，音響モデルの MLLR 適応を行う．今回の実験では，次の 2 つの条件について比較する．

(1) 適応なし (SI-GMM SI-AM)
(2) クリーン GMM に MAP 適応かつクリーン AM に MLLR 適応 (SA-GMM ES-AM)

実験 2 との主な違いは，クリーン GMM の適応に用いるデータが少ないことである．

実験結果を**表 3** にまとめる．実験 2 の SA-GMM ES-AM と比べると，特徴量強調用のクリーン GMM の MAP 適応データが少ないにもかかわらず，ほとんど同等の認識精度改善（11.20%）が達成された．一般に，音響モデルの MAP 適応は多量の適応データを要することが知られているが，クリーン GMM のパラメータ数は少ないため，数発話程度

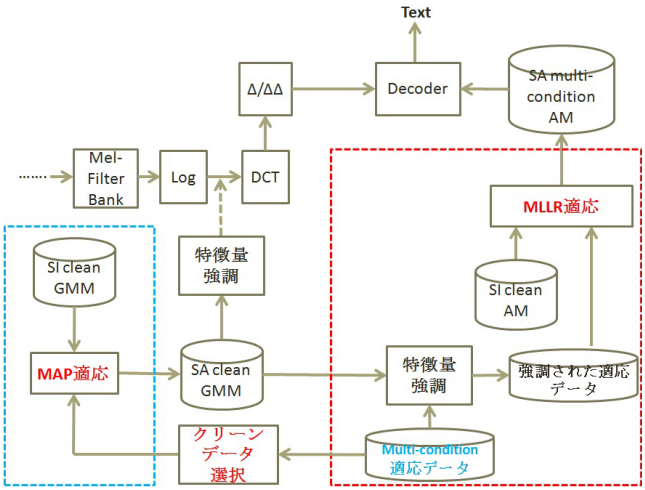

図 5 実験 3 の認識システム : 同一のマルチコンディション適応データを用いてクリーン GMM とクリーン音響モデルを適応させる.

Fig. 5 Speech recognition system used for Experiment 3, where clean GMM and clean acoustic model are adapted using the same multi-condition data.

のデータで十分な適応ができることを示している.

5. おわりに

5.1 まとめ

本研究では特徴量強調を行う音声認識システムにおいて，個人の音声データを用いてシステム全体の性能を向上させる手段について検討した．すなわち，個人データを適応データとして扱い，特徴量強調に用いるクリーン GMM の MAP 適応および音響モデルの MLLR 適応を行ういくつかの方法を比較検討した．具体的には，次のような 4 つの条件について検討した.

(1) 適応なし (ベースライン)
(2) クリーン GMM のみ MAP 適応
(3) クリーン AM のみ MLLR 適応
(4) クリーン GMM を MAP 適応，かつクリーン AM を MLLR 適応

適応データとして，クリーン音声だけからなる場合とマルチコンディションデータを用いる場合の 2 種類検討した．実験の結果，以下の 3 つのことが見出された.

(1) 特徴量強調に用いる GMM を話者適応化させると，常に認識性能が改善した.
(2) 音響モデルを話者適応化させる場合，同時に特徴量強調に用いる GMM も話者適応化させなければ，十分な認識性能の改善が得られない.
(3) 本研究では特徴量強調に用いる GMM を MAP 適応の方法を用いて適応させたが，適応データの量が少なくても十分高い効果が得られた.

5.2 今後の展望

最後に残された課題について述べる．まず，今回クリーン GMM に対する適応はクリーン適応データだけを使ったが，劣化した音声データも GMM の適応データとして直接用いることのできる方法を検討する必要がある．劣化音声データはクリーンデータよりも容易に収集できるので，これによって適応データを増やすことができ，その結果更なる認識性能の改善が期待される．次に，本研究では音響的な劣化要因として残響を取り上げたが，様々な雑音環境についても評価実験を行うことが肝要である．最後に，クリーン GMM の適応方法として MAP 適応以外の方法を用いることも検討したい.

参考文献

[1] Stouten, V.: *Robust automatic speech recognition in time-varying environments*, KU Leuven, Ph.D. Thesis (2006).
[2] Droppo, J. and Acero, A.: *Environmental robustness*, Spinger Handbook of Speech Processing (Benesty, J., Sondhi, M.M. and Huang, Y., eds.), Springer, pp.653-679 (2008).
[3] Lee, C.H. and Lin, C.H. and Juang, B.H.: *A study on speaker adaptation of the parameters of continuous density hidden Markov models*, IEEE Trans. Signal Processing, vol.39, no.4, pp.806-814 (1991).
[4] Gauvain, J.L. and Lee, C.H.: *Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains*, IEEE Trans. Speech and Audio Processing, vol.2, no.2, pp.291-298 (1994).
[5] Leggetter, C.J. and Woodland, P.C.: *Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models*, Computer Speech and Language, vol.9, pp.171-185 (1995).
[6] Gales, M.J.F.: *Maximum likelihood linear transformations for HMM-based speech recognition*, Computer Speech and Language, vol.12, pp.75-98 (1998).
[7] 吉岡拓也，中谷智広: 高即応・高精度な歪み特徴量モデルの推定のための動的静的アプローチ，情報処理学会研究報告 (2011).
[8] Zhao, Y. and Juang, B.H.: *A comparative study of noise estimation algorithms for VTS-based robust speech recognition*, Proc. Interspeech, pp.2090-2093 (2012).
[9] Gales, M.J.F: *Model-based approaches to handling uncertainty*, Robust Speech Recognition of Uncertain or Missing Data (Kolossa, D. and Haeb-Umbach, R., eds), Springer, pp.101-125 (2011).