最尤推定を用いた声道長線形変換による話者正規化

六井 淳[†] 中井 満[†]
下平 博[†] 嵯峨山 茂樹[††]

音声認識技術において性能劣化要因として話者性の違いや雑音などの使用条件の変化があげられる．近年，HMM（隠れマルコフモデル）のパラメータ推定に用いられる EM アルゴリズムに基づくケプストラム領域での声道長正規化手法が提案されている．従来法では，周波数領域において端点固定の非線形伸縮となるため，少量の適応データでは十分な精度が得られないという問題があった．本研究では声道長による特徴量の変化を周波数領域における線形伸縮ととらえ，ケプストラム空間へ変換する手法を提案する．従来のように複数の伸縮係数を用意するのではなく，最尤推定により伸縮係数を一意に求める．このため，実験的に本手法は少量の適応データにおいても良好な認識性能を与えることが確認された．

Speaker Normalization Using Linear Transformation of Vocal Tract Length Based on Maximum Likelihood Estimation

Jun Rokui,[†] Mitsuru Nakai,[†] Hiroshi Shimodaira[†]
and Shigeki Sagayama[††]

Vocal tract length normalization (VTLN) is one of the popular speaker adaptation techniques for speech recognition. The present study proposes a new VTLN algorithm in which expectation-maximization (EM) based parameter adaptation of HMM to vocal tract length is achieved in the mel-cepstral domain by utilizing a linear transformation model. Compared to other existing approaches based on bi-linear transformation for VTLN where a specific non-linear frequency warping function is employed in the spectrum domain and parameter adaptation of HMM is carried out in the cepstral domain, the proposed approach assumes a linear frequency warping with a single scaling factor and equivalent operation is modeled in the mel-cepstral domain by using a first order Taylor series approximation. The proposed scheme demonstrates significant improvement of recognition performance in a speaker independent word recognition task.

1. はじめに

現在の音声認識技術において認識性能を劣化させる主な要因として話者性の違いや雑音，回線などの使用条件の変化などがあげられる．本論文では声道長の違いに起因する話者性の問題に着目し，話者の声道長の変化を補正する新たな手法を提案する．

男性，女性，子供の発声音声の音響特徴は互いに大きく異なっており[1]，これは主に声道長に起因している．女性の声道長は男性の声道長よりも約 10%短く，子供の声道長は女性の声道長よりもさらに約 10%短いということが報告されている．観測された音声信号からこれら声道長の相違を推定することは容易ではない．音響学的には声道長はホルマントの位置と関係があることが知られている．このため，ホルマントの位置から声道長の補正を行う VTLN（Vocal Tract Length Normalization）が提案されている[2)~5)]．この手法は理想的な環境下では比較的少量のデータにより声道長が正確に求まることが確認されているが，実環境下では声道長の推定精度が低下する問題点が指摘されている．また，ホルマント位置の推定にともなう計算量が多いため実用的ではない．

実環境での声道長の推定精度の低下を解決するため，ML-VTLN（Maximum Likelihodd Vocal Tract Length Normalization）[6)~8)] が提案されている．ML-VTLN ではあらかじめ用意した複数の声道長パラメータの中から対象話者に最適なパラメータを選択する．

† 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology, Hokuriku
†† 東京大学
The University of Tokyo

そのため用意したパラメータの数だけ尤度計算が必要であり，VTLN 同様，計算量が多い．また，用意した有限個のパラメータ集合内でパラメータの探索を行うので，依然として推定精度の問題をかかえている．

計算量と推定精度の問題を解決するために，近年，VTLN-R（Vocal Tract Length Normalization using Rapid Maximum Likelihood Estimation）[9),10)] が提案された．VTLN-R はケプストラムを特徴量とする HMM（hidden Markov model）音響モデルを想定した声道長正規化の手法で，単一パラメータによって構成されるケプストラム空間上の線形変換によって HMM のパラメータ補正を行う．変換のパラメータがケプストラム空間上の最尤推定基準で求まるため，計算量が少なく，推定精度も高い．しかし，声道長の違いがパワースペクトルの周波数軸の線形伸縮で本来モデル化されるのに対して，VTLN-R が用いるケプストラム空間上の線形変換は，パワースペクトル上では周波数軸の非線形変換として現れる．そのため，パラメータの推定精度が高くても声道長正規化の効果が十分に得られない可能性がある．実際，周波数軸を線形に伸縮した方が非線形に伸縮した場合よりも高い性能を与えるということが報告されている[8)]．

そこで，本論文では VTLN-R と同様に，HMM の特徴量空間において最尤推定に基づく声道長正規化が可能で，かつ，得られた正規化がパワースペクトル領域では周波数軸の線形伸縮に相当する，新たな手法を提案する．特に，特徴量として近年の音声認識で広く利用されている MFCC（Mel-Frequency Cepstrum Coefficient）を想定した HMM のパラメータ補正の方式を示す．このようなケプストラム，あるいはメル・ケプストラム空間で HMM の適応を行う際に問題となるのが，周波数スペクトル上の周波数軸の線形伸縮が，（メル）ケプストラム空間では非線形な変換に対応する点である．このため，線形伸縮パラメータを HMM のパラメータ推定と同様の手法で効率良く求めることが困難である．本論文では同問題に対して線形近似による解決法を新たに示す．

2. 声道長正規化のモデル

前述したように，本論文では HMM における声道長正規化を少ない演算量で実現するために，声道長の正規化処理をパワースペクトル上ではなく，HMM の特徴量空間である（メル）ケプストラム空間で行う．以下ではメルケプストラム空間の特徴量である MFCC を用いた場合について正規化のモデル化を行う．ケプストラム空間上の特徴量 CC を用いた場合も考え方は同じである．

声道長の違いを周波数スペクトル上における周波数軸の線形伸縮としてとらえれば，伸縮係数 ν を用いて，$\tilde{\omega} = \nu\omega$ の形で定式化される．ここで，ω，$\tilde{\omega}$ は，それぞれ変換前と変換後の角周波数で，$\nu > 1$ は声道長の短縮を，$\nu < 1$ は伸長を意味する．一方，離散化された周波数軸における表現を考えると，周波数パワースペクトルはベクトルの形で表現されるので，周波数変換を行う前のパワースペクトルを N 次元空間上のベクトル $\boldsymbol{p}$，変換後のそれを $\tilde{\boldsymbol{p}}$ で記すと，両者は以下のような線形変換の形で表せる．

$$\tilde{\boldsymbol{p}} = W\boldsymbol{p} \tag{1}$$

ここで，変換行列 W の第 (i,j) 成分 w_{ij} は次式で与えられる．

$$w_{i,j} = \begin{cases} 1, & \text{if} \quad i = \mathrm{mod}(\lfloor \nu(j-1) \rfloor, N) + 1 \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

$$W_{N\times N} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,N} \\ & \vdots & \vdots & \\ w_{N,1} & \cdots & \cdots & w_{N,N} \end{bmatrix}$$

ここで $\lfloor x \rfloor$ は任意の実数 x について x より小さくない最小の整数，$\mathrm{mod}(m,n)$ は m を n で割ったときの剰余を意味する．

本手法の場合，全域通過フィルタを用いるなどして非線形伸縮を行ったものとは異なり，周波数軸の圧縮の際，高次成分の扱いが問題となる．本研究の場合，成分の入らない高次成分には周波数軸の低次成分が折り返しを行うようなアルゴリズムとなっている．これは周波数軸の高次成分に歪みを与えることになるが，音韻特性が周波数軸の低次に集中することから，認識に大きな影響はないと考えられる．むしろ，従来の端点固定な非線形伸縮では周波数軸の中央成分から高次成分にかけて大きく歪んでしまうため，認識に悪影響が出ると考えられる．

次に周波数領域での変換行列 W に基づき，メルケプストラム領域への変換方法を示す．パワースペクトルからメル周波数スペクトルへのフィルタバンク行列 $F_{n_c\times N}$ で記すと，メル周波数スペクトル（ベクトル）$\boldsymbol{s}$ は次式で与えられる☆．

☆

$$F_{n_c\times N_{n_c}} = \begin{bmatrix} f_{11} & \cdots & f_{1N_1} & 0 & \cdots & 0 \\ 0 & \cdots & f_{21} & \cdots & & 0 \\ \vdots & \vdots & & & & \vdots \\ 0 & & \cdots & & \cdots & f_{n_cN_{n_c}} \end{bmatrix}$$

ここで，f_{ij}, N_i は，それぞれ第 i 番目のフィルタの重みと帯域幅を意味し，n_c はフィルタの総数を表す．

$$\boldsymbol{s} = F\boldsymbol{p} \tag{3}$$

ここで，F は正方行列ではないため，式 (3) の逆変換，すなわち，$\boldsymbol{s}$ から $\boldsymbol{p}$ の完全な復元を求めることができない．そこで本論文では F のムーアペンローズ一般化逆行列 F^- を用いる．行列 F^- は $F\,F^-\,F = I$（恒等行列）となるような行列である．結果として，メル周波数スペクトル $\boldsymbol{s}$ と周波数伸縮後のメル周波数スペクトル $\tilde{\boldsymbol{s}}$ の間には次のような関係がある．

$$\tilde{\boldsymbol{s}} = T\boldsymbol{s} \tag{4}$$

ここで，T は次式で与えられる．

$$T = F\,W\,F^- \tag{5}$$

次に，MFCC 特徴量における関係式を求めるために，まず，離散コサイン変換（DCT）行列を C で，メルスペクトル $\boldsymbol{s}$ に対応するメルケプストラムを $\boldsymbol{c}$ で記すと，$\boldsymbol{c}$ と $\boldsymbol{s}$ の関係は次のように表される．

$$\boldsymbol{c} = C\log\boldsymbol{s} \tag{6}$$

ここで，log は自然対数である☆．式 (6) と式 (4) より，周波数伸縮前のメルケプストラム $\boldsymbol{c}$ と伸縮後の $\tilde{\boldsymbol{c}}$ との関係式が得られる．

$$\tilde{\boldsymbol{c}} = C\log\tilde{\boldsymbol{p}} = C\log\left\{T\exp(C^{-1}\boldsymbol{c})\right\} \tag{7}$$

3. HMM 適用のための拡張

前章で得られた MFCC 特徴量空間における声道長線形伸縮の関係式 (7) は，伸縮によってメルケプストラム・ベクトル空間上の 1 点がどのように移動するかを示している．この関係を，確率・統計モデルである HMM の出力確率密度関数のパラメータに反映させるには注意が必要である．たとえば，伸縮前の MFCC の平均ベクトルを $\boldsymbol{\mu}_c = E[\boldsymbol{c}]$ と置くと，伸縮後の平均ベクトル $\tilde{\boldsymbol{\mu}}_c = E[\tilde{\boldsymbol{c}}]$ との間には式 (7) の関係が成り立たない．すなわち，

$$E[\tilde{\boldsymbol{c}}] = E\{C\log\{T\exp(C^{-1}\boldsymbol{c})\}\} \tag{8}$$

$$\neq C\log\{T\exp(C^{-1}E[\boldsymbol{c}])\} \tag{9}$$

となって，平均ベクトルの線形変換の形で表現できない：

$$\tilde{\boldsymbol{\mu}}_c \neq C\log\{T\exp(C^{-1}\boldsymbol{\mu}_c)\} \tag{10}$$

この問題を解決するために本論文では式 (7) で与えられる変換を，テイラー展開の 1 次項による近似を利用して以下のように線形近似する．

$$\tilde{\boldsymbol{c}} = C\log\left\{T\exp(C^{-1}\boldsymbol{c})\right\} \tag{11}$$

ここで，$\exp(x+1) \approx x$ という線形近似を行う．

$$\approx C\log\left\{T(\boldsymbol{e} + C^{-1}\boldsymbol{c})\right\} \tag{12}$$

$$= C\log\left\{U\left(\boldsymbol{e} + U^{-1}TC^{-1}\boldsymbol{c}\right)\right\} \tag{13}$$

$$= C\log\left\{U\left(\boldsymbol{e}\cdot\left(\boldsymbol{e} + U^{-1}TC^{-1}\boldsymbol{c}\right)\right)\right\} \tag{14}$$

$$= C\log\left\{(U\boldsymbol{e})\cdot\left(\boldsymbol{e} + U^{-1}TC^{-1}\boldsymbol{c}\right)\right\} \tag{15}$$

$$= C\log U\boldsymbol{e} + C\log\left(\boldsymbol{e} + U^{-1}TC^{-1}\boldsymbol{c}\right) \tag{16}$$

ここで，$\log(x) \approx x+1$ という線形近似を行う．

$$\approx C\log U\boldsymbol{e} + CU^{-1}TC^{-1}\boldsymbol{c} \tag{17}$$

ここで，$\boldsymbol{e}$ は単位ベクトル，行列 U は次式で与えられる☆☆．

$$U = \begin{bmatrix} \sum_{j=1}^{n_c} t_{1j} & & 0 \\ & \ddots & \\ 0 & & \sum_{j=1}^{n_c} t_{n_c j} \end{bmatrix} \tag{18}$$

ただし，t_{ij} は行列 T の (i,j) 成分，すなわち $t_{ij} = (T)_{ij}$ である．

したがって，ベクトル $\boldsymbol{q}$，行列 B を以下のように定義し，

$$\boldsymbol{q} = C\begin{bmatrix} \log\left(\sum_{j=1}^{n_c} t_{1j}\right) \\ \vdots \\ \log\left(\sum_{j=1}^{n_c} t_{n_c j}\right) \end{bmatrix} \tag{19}$$

$$B = CU^{-1}TC^{-1} \tag{20}$$

これを式 (17) に代入すると最終的に以下の近似式が得られる．

$$\tilde{\boldsymbol{c}} \approx \boldsymbol{q} + B\boldsymbol{c} \tag{21}$$

線形近似の結果，平均値 $\boldsymbol{\mu}_c$，分散 Σ_c に関する変換も次のように簡潔になる．

$$\tilde{\boldsymbol{\mu}}_c \approx \boldsymbol{q} + B\boldsymbol{\mu}_c \tag{22}$$

$$\tilde{\Sigma}_c \approx B\Sigma_c B^t \tag{23}$$

上式の形から，本手法は声道長に関する制約を設けた MLLR（Maximum Likelihood Linear Regression）[11] と考えることもできる．

次に，i 番目フレームのデルタケプストラム $\Delta\boldsymbol{c}^{(i)}$ と動的尺度 $\Delta^2\boldsymbol{c}^{(i)}$ の変換に関しては次式が得られる．

$$\Delta\tilde{\boldsymbol{c}}^{(i)} = \tilde{\boldsymbol{c}}^{(i)} - \tilde{\boldsymbol{c}}^{(i-1)} \tag{24}$$

$$\approx B(\boldsymbol{c}^{(i)} - \boldsymbol{c}^{(i-1)}) = B\Delta\boldsymbol{c}^{(i)} \tag{25}$$

$$\Delta^2\tilde{\boldsymbol{c}}^{(i)} \approx B\Delta^2\boldsymbol{c}^{(i)} \tag{26}$$

したがって，平均値，分散に関する変換も次のように簡潔になる．

$$\tilde{\boldsymbol{\mu}}_{\Delta c} \approx B\boldsymbol{\mu}_{\Delta c} \tag{27}$$

$$\tilde{\boldsymbol{\mu}}_{\Delta^2 c} \approx B\boldsymbol{\mu}_{\Delta^2 c} \tag{28}$$

☆ 本論文では任意のベクトル $\boldsymbol{x}$ の各要素に対する演算の意味で，$\log\boldsymbol{x}$，$\exp(\boldsymbol{x})$ のような便宜的な表現を用いる．

☆☆ 演算子 $\cdot$ はベクトル $\alpha = (a_1, a_2, \cdots, a_n), \beta = (b_1, b_2, \cdots, b_n)$ が与えられた場合，$\alpha\cdot\beta = (a_1b_1, a_2b_2, \cdots, a_nb_n)$ を表す．

$$\tilde{\Sigma}_{\Delta c} \approx B\Sigma_{\Delta c}B^t \tag{29}$$
$$\tilde{\Sigma}_{\Delta^2 c} \approx B\Sigma_{\Delta^2 c}B^t \tag{30}$$

4. 伸縮係数の最尤推定

4.1 伸縮係数導出の定式化

本章では周波数スペクトルの周波数伸縮係数 ν の尤度最大化基準による推定方法（最尤推定法）について述べる．最尤法による ν の最適解 ν^* は次式で定義される．

$$\nu^* = \arg\max_{\nu} P(O|\Theta) \tag{31}$$

ここで，O は観測系列，$P(O|\Theta)$ は出現確率であり，$\Theta \equiv (\theta, \nu)$ で，θ は HMM のパラメータ集合である．

最適解 ν^* の推定は HMM のパラメータ推定で用いられる Baum-Welch アルゴリズムを適用する．すなわち，以下の目的関数の最大化問題として ν^* の推定を行う．

$$\Phi(\Theta', \Theta) = \sum_{j=1}^{J}\sum_{t=1}^{\mathcal{T}} P(O, p_t = j|\Theta') \log b_j(\tilde{\boldsymbol{c}}_t) \tag{32}$$

ここで，$\tilde{\boldsymbol{c}}_t$ は時刻 t の観測 MFCC である $\boldsymbol{c}_t$ に対して伸縮係数 ν による声道長補正を行った後の MFCC，$\mathcal{T}$ は観測系列の時間長，J は HMM の状態数，q_t は時刻 t における HMM の状態番号である．$b_j(\tilde{\boldsymbol{c}}_t)$ は状態 j における出力確率密度関数で，本論文では次式で与えられように平均 $\boldsymbol{\mu}_j$，共分散行列 Σ_j で与えられる M 次元ベクトル空間上の正規分布を仮定している．

$$b_j(\tilde{\boldsymbol{c}}_t) = \frac{1}{\sqrt{(2\pi)^M|\Sigma_j|}} \exp\left\{-\frac{1}{2}(\tilde{\boldsymbol{c}}_t - \boldsymbol{\mu}_j)^t\Sigma_j^{-1} (\tilde{\boldsymbol{c}}_t - \boldsymbol{\mu}_j)\right\} \tag{33}$$

ただし，本論文では共分散行列 Σ_j は対角行列を仮定している．

目的関数の極大点を与える ν^* について次式が成立する．

$$\frac{\partial\Phi(\Theta', \Theta)}{\partial\nu^*} = \sum_{j=1}^{J}\sum_{t=1}^{T} \frac{P(O, q_t = j|\Theta')\frac{\partial b_j(\tilde{\boldsymbol{c}}_t)}{\partial\nu^*}}{\partial b_j(\tilde{\boldsymbol{c}}_t)} = 0 \tag{34}$$

式 (21) を利用して変形すると次式が得られる，

$$\sum_{j=1}^{J}\sum_{t=1}^{\mathcal{T}} P(O, q_t = j|\Theta') \left[\sum_{m=1}^{M} \frac{1}{\sigma_{mj}^2} (\tilde{c}_{mt} - \mu_{mj})(-c_{mt})\right] = 0 \tag{35}$$

ここで，c_{mt}，$\tilde{c}_{mt}$，μ_{mj} はそれぞれベクトル $\boldsymbol{c}_t$，$\tilde{\boldsymbol{c}}_t$，$\boldsymbol{\mu}_j$ の第 m 成分，σ_{mj} は対角共分散行列 Σ_j の第 (m,m) 成分，すなわち $\sigma_{mj} = (\Sigma_j)_{mm}$ である．伸縮係数 ν^* について解くと最終的に次式が得られる．

$$\nu^* = \frac{C\log\left(\exp\left(C^{-1}\sum_{j=1}^{J}\sum_{t=1}^{\mathcal{T}}\gamma_t(j)\cdot \frac{\left[\sum_{m=1}^{M}\frac{1}{\sigma_{mj}^2}\mu_{mt}\mathbf{c}_{mt}\right]}{\left[\sum_{m=1}^{M}\frac{1}{\sigma_{mj}^2}\mathbf{c}_{mt}\right]}\right)\right)}{\sum_{j=1}^{J}\sum_{t=1}^{\mathcal{T}}\gamma_t(j)\cdot} \tag{36}$$

ここで，$\gamma_t(j)$ は次式で与えられる占有度数である．

$$\gamma_t(j) = \frac{P(O, q_t = j|\Theta)}{\sum_{j=1}^{J} P(O, q_t = j|\Theta)} \tag{37}$$

4.2 声道長正規化による音声認識の処理手順

4.2.1 声道長正規化学習処理

(1) Baum Welch アルゴリズムで学習音声の MFCC $\boldsymbol{c}^{(s)}$（s は話者を表すパラメータ）をすべて用い，占有度数 $\gamma_t(j)$ の計算を行う．
(2) 占有度数 $\gamma_t(j)$ より話者ごとの伸縮重み $\nu^{(s)}$ を計算する．
(3) 声道長補正 MFCC $\tilde{\boldsymbol{c}}^{(s)}$ を求める．
(4) 学習の話者数だけ (2)〜(3) の処理を行う．
(5) 占有度数 $\gamma_t(j)$ と $\tilde{\boldsymbol{c}}^{(s)}$ より HMM パラメータの再推定を行う．
(6) $\boldsymbol{c}^{(s)}$ を $\tilde{\boldsymbol{c}}^{(s)}$ に置き換え，占有度数 $\gamma_t(j)$ の計算を行う．
(7) (1) から繰り返す．

4.2.2 認 識 処 理

(1) 声道長正規化学習用の音声から伸縮重み $\nu^{(s)}$ を推定．
(2) 認識音声の MFCC $\boldsymbol{c}^{(s)}$ から適応時に計算される $\nu^{(s)}$ を用いて声道長補正された $\tilde{\boldsymbol{c}}^{(s)}$ を求める．
(3) すべての特徴量を用いて認識処理を行う．

5. 実 験

5.1 多数話者モデルを用いた認識実験

本節では，本手法による声道長正規化の効果を検証するため，適応前と適応後の効果について検証した．実験条件は**表 1** のとおりである．

適応に用いる単語数と認識性能の関係を男女別に集計した結果を**図 1** に示す．ここで適応単語数 0 は適

表 1 実験条件（1）
Table 1 Experimental condition (1).

分析条件	標本化周波数 12 kHz，ハミング窓 20 ms，フレーム間隔 10 ms
特徴量	MFCC 13 次元，MFCC＋ΔMFCC＋Δ^2MFCC
音響モデル	音素環境独立型 HMM（3 状態，3 混合，対角共分散），27 モデル
音声データベース	ATR データベース A セット（孤立単語）
話者	男性 5 話者（mht，mnm，msh，mmy，mms） 女性 5 話者（ffs，fms，fkn，fyn，faf）
学習データ	奇数番目 2620 単語/話者
評価データ	偶数番目 655 単語/話者
辞書	2620 単語
初期モデル	認識話者を除く 9 話者で学習
評価法	10 話者による交差検定☆

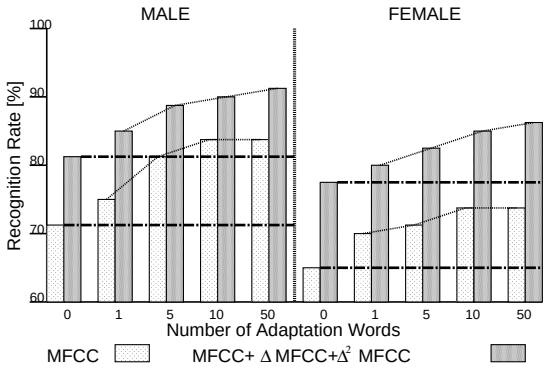

図 1 多数話者モデルによる認識実験結果
Fig. 1 Isolated word recognition results based on speaker independent model.

応を行う前の初期音響モデルによる認識結果を示している．認識率は評価話者 10 名による交差検定☆により求めた．

この結果より，適応前と比べ，MFCC13 次元の特徴量を用いた場合では，適応単語数 50 個で約 32%，適応単語数 1 個の場合でも約 15%の誤り削減率（10 話者平均）を実現している．また，MFCC＋ΔMFCC＋Δ^2MFCC 徴量を用いた場合，適応単語数 50 個で約 40%，適応単語数 1 個の場合でも約 13%の誤り削減率を実現している．

5.2 性別依存モデルによる実験結果

男性から女性，女性から男性，大人から子供のように声道長が大きく異なる音声に対する適応は容易ではない．本節では性別依存モデルを作成し，認識実験を行うことで，声道長が大きく異なった音声への適応が可能かについての検証を行った．実験条件は**表 2** のとおりである．その他の分析条件，音響モデルなどは表 1 と同じである．

表 2 実験条件（2）
Table 2 Experimental condition (2).

学習話者	男性 7 話者（mmy，mnm，msh，mtk，mtm，mtt，mxm） 女性 4 話者（ffs，fms，fkn，fyn）
評価話者	男性 1 話者（mms），女性 1 話者（faf）
男性初期モデル	男性 7 話者で学習
女性初期モデル	女性 4 話者で学習
特徴量	MFCC 13 次元

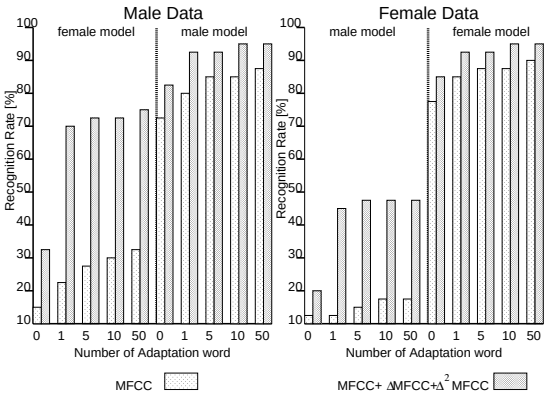

図 2 性別依存モデルによる認識実験結果
Fig. 2 Word accuracy rate based on the gender dependence model.

実験結果を**図 2** に示す．左側の図が音響モデルと評価データの性別が異なる場合，右側の図が性別が同一の場合の適応単語数に対する認識率を示している．

この結果から，男女のデータを混合させた初期モデルを用いた場合の認識性能（図 1）より劣るもののの，音響モデルと評価データの性別が異なる場合でも適応の効果が認められる．MFCC のみを特徴量としたとき，適応しない場合と比較して適応単語数 1 個で約 5%，50 個で約 12%の誤り削減率（男女平均）が得られている．さらに，ΔMFCC，Δ^2MFCC 特徴も併用した場合には 1 単語で約 42%，50 単語で約 48%の誤り削減率が得られている．

5.3 他手法との比較実験

5.3.1 多数話者モデルを用いた比較実験

提案手法の有効性を検証するために，VTLN-R と回帰クラス数を 1 とした場合の MLLR との比較を行った．MLLR（Maximum Likelihood Linear Regression）[11] は話者適応法として現在広く用いられている

☆ 交差検定：N 個のデータセットがある場合，$N-1$ 個のデータセットで学習し，残りの 1 セットで評価，これを評価データセットを変えてすべての組合せ（N 通り）について行い，その平均を求める検定法．

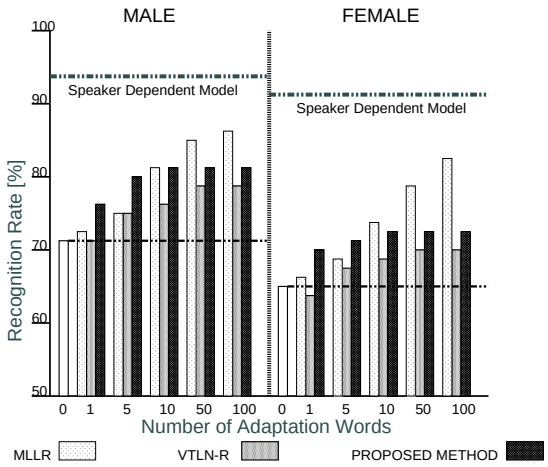

図 3 提案手法と MLLR，VTLN-R との比較
Fig. 3 Word accuracy rate of VTLN-R, MLLR, and the proposed method.

手法である．VTLN-R [9] は全域通過フィルタを周波数伸縮関数として用いる手法であり，高速かつ高い認識効果を与える声道長正規化手法である．

実験条件は表 1 と同じである．実験結果を**図 3** に示す．

図 3 から分かるように提案手法は適応単語数のいかんによらず VTLN-R よりも高い認識結果を示している．ただし，提案手法は認識率が 10 単語程度で飽和し，それ以上単語数を増やしても向上しないのに対して，VTLN-R の場合，適応単語数の増加に応じて認識率が向上しており，飽和傾向が見られない．MLLR は適応単語数が 5 単語以下の場合，提案手法より認識性能が低いが，10 単語を超えるあたりから提案手法より高い認識性能を示している．

以上の結果から，提案手法は適応単語数が非常に少ない状況における話者適応に適しているといえる．

5.3.2 特徴量の次元数に関する検討

VTLN-R の文献 9) においては，ケプトラム特徴量の次元数を 6 次元程度に落とした方が話者適応の効果が高いと報告されている．そこで，本論文でも同様の実験を行った．

実験条件は**表 3** のとおりである．MFCC13 次元の場合のフィルタバンクのチャネル数は 26，MFCC6 次元の場合のチャネル数は 12 である．

結果を**図 4** に示す．特徴量は 13 次元を用いた方が 6 次元の場合よりも高い認識性能が得られており，この結果は先の報告と一致していない．提案手法と VTLN-R との比較においては，MFCC13 次元の方が提案手法の優位性が出ており，MFCC6 次元の場合は

表 3 実験条件（5）
Table 3 Experimental condition (5).

学習話者	男性 2 話者（mau，mht），女性 2 話者（ffs，fms）
評価話者	男性 1 話者（mms），女性 1 話者（faf）
初期モデル	男性 2 話者，女性 2 話者で学習
特徴量	MFCC 6 次元，13 次元

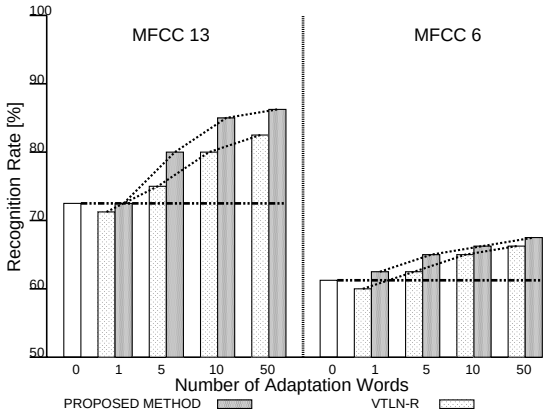

図 4 特徴量の次元数と認識率の関係
Fig. 4 Word accuracy rate obtained with MFCC dimensions.

表 4 実験条件（3）
Table 4 Experimental condition (3).

特徴量	MFCC 13 次元
学習話者	男性 2 話者（mau，mht） 女性 2 話者（ffs，fms）
評価話者	男性 1 話者（mau）
初期モデル	男性 2 話者，女性 2 話者で学習
非線形周波数正規化法[9]	$Z^{-1} = \frac{z^{-1}-a}{1-az^{-1}}$ ただし，$z = \exp(j\omega)$， $Z = \exp(j\nu\omega)$.

両者の差は小さく，特に，適応単語数が多い環境で両者の差はほとんどない．

5.4 周波数伸縮法の比較と推定精度

これまでの実験から，提案手法が VTLN-R よりも高い認識性能を示すことが分かったが，この原因について検討した．両手法の本質的な違いは提案手法が線形な周波数伸縮を行っているのに対して，VTLN-R は非線形な処理を行っている点にある．この違いが認識性能に及ぼす影響を調べるために，伸縮係数の推定を行わずに直接これを操作して認識率との関係を調べた．すなわち，線形周波数スペクトル領域において，伸縮係数の種々の値について提案手法と VTLN-R における周波数伸縮を行い，線形近似を用いずに特徴量を求めてモデルの学習および評価実験を行った．

実験条件を**表 4**，認識実験結果を**図 5** に示す．この図から，線形な周波数伸縮の方が非線形な伸縮よりも

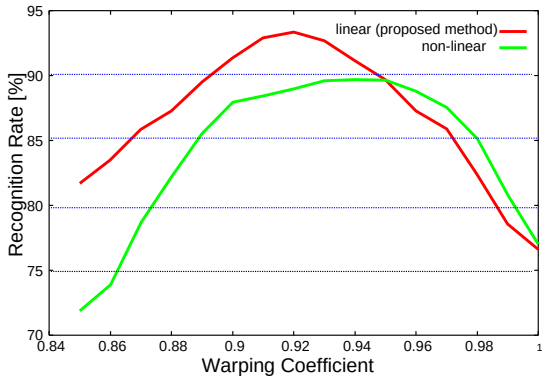

図 5 声道長伸縮係数と認識率の関係
Fig. 5 Word accuracy rate obtained with warping coefficient.

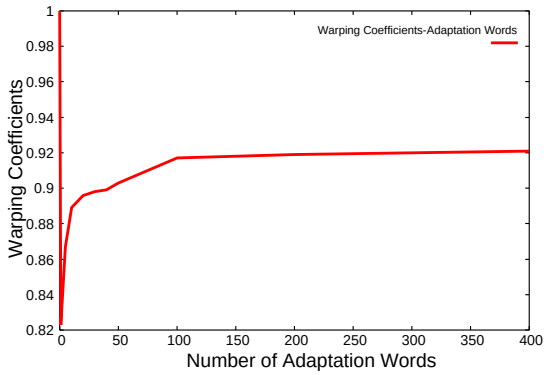

図 6 適応単語数と伸縮係数変化
Fig. 6 Warping coefficient values obtained with adaptation words.

高い適応効果があることが分かる．この結果は文献 8) における報告と一致している．

なお，図 5 から提案法における伸縮係数の最適値は $\nu = 0.92$ 付近にあることが分かる．一方，提案法の最尤推定によって求めた伸縮係数の値と，適応に用いた単語数の関係を**図 6** に示す．この図から，単語数 100 程度で推定結果が最適値に漸近しており，十分な量の適応データがあれば，ほぼ最適値を推定できることが分かる．単語数が 10 単語以下の場合は推定値が 0.86 〜0.88 となって推定誤差が大きくなるが，図 5 から分かるように，それでも伸縮を行わない場合（$\nu = 1.0$）よりも認識率が高いので提案法による話者正規化を行った方が良いといえる．

ところで，図 5 における最適な伸縮係数 0.92 における認識率は 93.6%であるのに対して，ほぼ同じ伸縮係数における提案法の認識率について調べたところ 92.5%であった．両者の差は，式 (21) で示される線形近似の有無によると考えられる．したがって，線形近似による認識率の低下は本実験では 1.1 ポイントの減少に収まっている．

6. ま と め

本論文では声道長パラメータを最尤推定によって決定し，線形変換によって声道長正規化を行う手法を提案した．提案手法同様に最尤推定によって声道長パラメータを求める手法（VTLN-R）と一般的適応手法である MLLR との比較実験から，適応単語数が 5 単語以下の場合には提案手法に優位性があることを確認した．また，周波数スペクトル領域における声道長正規化法について，提案法で用いている線形伸縮法と従来研究で広く用いられている非線形伸縮法による比較実験を行った結果，提案法で用いている線形伸縮の方が高い認識性能が得られることが分かった．

今後は雑音，伝達特性などの環境要因も含んだアルゴリズム構築を行い，実環境における有効性を評価したい．

参 考 文 献

1) Wakita, H.: Normalization of vowels by vocal tract length and its application to vowel identification, *IEEE Trans. Acoust, Speech, Signal Processing*, ASSP25:183 (1997).
2) Eide, E. and Gish, H.: A parametric approach to vocal tract length normalization, *ICASSP96*, Vol.1, pp.346–348 (1996).
3) Wakita, H.: Estimation of vocal tract shapes from acoustical analysis of the speech wave, *IEEE Acoust. Speech, Signal Processing*, ASSP27:281 (1979).
4) McDonough, J., Metze, F., Soltau, H. and Waibel, A.: Speaker compensation with sine-log all-pass transforms, *ICASSP2001* (May 2001).
5) Claes, T., Dologlou, J., Bosch, L.T. and Van Compernolle, D.: A novel feature transformation for vocal tract length normalization in automatic speech recognition, *IEEE Trans. Speech and Audio Processing*, Vol.6, No.6 (Nov. 1998).
6) Lee, L. and Rose, R.C.: Speaker normalization using efficient frequency warping procedure, *ICASSP96*, Vol.1, pp.353–356 (1996).
7) Kanthak, S., Welling, L. and Key, H.: Improved methods for vocal tract normalization, *ICASSP99*, p.1436 (1999).
8) Zhan, P. and Westohal, M.: Speaker normalization based on frequency warping, *ICASSP97*, pp.1039–1042 (1997).
9) 江森　正，篠田浩一：音声認識のための高速最ゆ

う推定を用いた声道長正規化，電子情報通信学会論文誌 DII，Vol.J83-DII, No.11, pp.2108–2117 (Nov. 2000).
10) Anastasakos, T., McDonough, J., Schwartz, R. and Makhoul, J.: A compact model for speaker-adaptive training, *ICSLP96*, Vol.2 (1996).
11) Legatter, 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).

(平成 13 年 11 月 16 日受付)
(平成 14 年 4 月 16 日採録)

六井　淳

1996 年信州大学理学部数学科卒業．1998 年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．現在，北陸先端科学技術大学院大学情報科学研究科博士後期課程在学中．複雑系理論，機械学習，ヒューマンインタフェース，音声認識に関する研究に従事．電子情報通信学会，ヒューマンインタフェース学会各会員．

中井　満

1991 年東北大学工学部情報工学科卒業．1993 年同大学大学院博士前期課程（情報工学）修了．1996 年同大学院博士後期課程（電気・通信工学）修了．1996 年北陸先端科学技術大学院大学情報科学研究科助手，現在に至る．工学博士．音声認識，文字認識に関する研究に従事．電子情報通信学会，日本音響学会各会員．

下平　博（正会員）

1982 年東北大学工学部電気工学科卒業．1984 年同大学大学院博士前期課程（情報工学）修了．1988 年同博士後期課程修了．同年同大学工学部情報工学科助手．1992 年北陸先端科学技術大学院大学情報科学研究科助教授，現在に至る．工学博士．音声，文字，画像の認識処理およびヒューマンインタフェースに関する研究に従事．電子情報通信学会，日本音響学会，IEEE 各会員．

嵯峨山茂樹（正会員）

1972 年東京大学工学部計数工学科卒業．1974 年同大学大学院工学系研究科計数工学専攻修士課程修了．同年日本電信電話公社に入社，武蔵野電気通信研究所にて音声情報処理の研究に従事．1990 年 ATR 自動翻訳電話研究所音声情報処理研究室長として自動翻訳電話プロジェクトを遂行．1993 年 NTT ヒューマンインタフェース研究所にて音声認識・合成・対話の研究開発に従事．1998 年北陸先端科学技術大学院大学情報科学研究科教授．2001 年東京大学大学院工学系研究科のち情報理工学系研究科教授．博士（工学）．1990 年発明協会発明賞，1994 年日本音響学会技術開発賞，1995 年情報処理学会山下記念研究賞，1996 年科学技術庁長官賞（研究功績者表彰）および電子情報通信学会論文賞等を受賞．日本音響学会，電子情報通信学会，IEEE，ヨーロッパ音声通信学会（ESCA），AVIRG 各会員．