論文 / 著書情報
Article / Book Information

論題(和文)	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
Title(English)	
著者(和文)	江森 正, 篠田浩一
Authors(English)	Koichi Shinoda
出典(和文)	電子情報通信学会論文誌, Vol. J83-D-II, No. 11, pp. 2108-2117
Citation(English)	, Vol. J83-D-II, No. 11, pp. 2108-2117
発行日 / Pub. date	2000, 11
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2000 Institute of Electronics, Information and Communication Engineers.

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

江森　正[†]　篠田　浩一[†]

Vocal Tract Length Normalization Using Rapid Maximum-Likelihood Estimation for Speech Recognition

Tadashi EMORI[†] and Koichi SHINODA[†]

あらまし　近年，隠れマルコフモデル（HMM）を用いた大語彙音声認識システムにおいて，声道長正規化と呼ばれる話者による声道長の違いを補正する話者正規化の手法が提案されている．本論文では，声道長による特徴量の変化を，ケプストラム空間における声道長パラメータを用いた線形写像で近似し，そのパラメータを発声から最ゆう推定する手法を提案する．従来の複数の声道長パラメータをあらかじめ用意する手法に比べ，計算量が少なく，より話者に最適なパラメータが推定可能である．日本語 5000 単語認識を用いた評価実験において，本方式単独で，7.1%誤りが減少し，また，ケプストラム平均正規化（CMN）と組み合わせた場合に，14.6%誤りが減少した．

キーワード　音声認識，隠れマルコフモデル，話者正規化，声道長，最ゆう推定

1. まえがき

近年，音声認識においては，隠れマルコフモデル（Hidden Markov Model; HMM）を用いた認識手法が一般に用いられている．HMM は，様々な要因により生じる発声の揺らぎを同一の確率分布からの異なる出力として扱うことが可能であり，その確率分布を学習する効率的なアルゴリズムが存在する．この特徴を生かし，話者の違いを発声の揺らぎの要因の一つととらえ，事前に集めた多数話者の発声データを用いてモデルを学習することにより，だれの声でも認識可能な不特定話者認識システムの実用化が可能となっている．

しかしながら，このような不特定話者認識システムは，使用者の音声を事前に登録した特定話者認識よりも一般に性能が低い．また，極端に認識性能が低い話者（特異話者）の存在が知られている．これらは，学習データに含まれる話者数が限られており，すべての話者の発声の音響的な多様性を網羅するモデルを作成できないためと考えられる．この問題は完全に解決することは困難である．そこで，多くのシステムでは，話者の違いから生じる発声の揺らぎに対処するために，話者適応化，話者正規化と呼ばれる手法を導入している．

話者適応化は，パターンマッチングに用いるモデルのパラメータを，使用者の少量の発声を用いて再推定する手法である．多くの話者適応化では，不特定話者モデルからその話者のモデルへの特徴量空間における写像を作成する．例えば，特徴量空間におけるアフィン変換を用いる方法（e.g. Maximum Likelihood Linear Regression; MLLR [6]），特徴量空間の区分空間における平行移動を用いる方法（e.g. Automatic Model Complexity Control; AMCC [12]），などがある．

話者正規化は，特徴抽出の段階で，話者の違いにより生じる発声の揺らぎを取り除く手法である．代表的なものに，ケプストラム平均正規化（Cepstrum Mean Normalization; CMN）[2]，声道長正規化（Vocal Tract Length Normalization; VTLN）[3] が挙げられる．CMN は，入力データの特徴量であるケプストラムの長時間平均を入力データから差し引く手法であり，話者のみならず，周囲雑音，反響，回線の違いなどにより生じる，発声の音韻的特徴の変化に比べ十分長時間のスケールで変化する揺らぎを取り除く効果がある．VTLN は，話者の声道長の違いにより生じる揺らぎを取り除く方法である．話者の声道長には個人差があり，声道（Vocal Tract）の共鳴周波数（ホルマント

[†] NEC 情報通信メディア研究本部，川崎市
Computer & Communication Media Research, NEC Corporation, 4–1–1 Miyazaki, Miyamae-ku, Kawasaki-shi, 216–8555 Japan

周波数）が異なる．このことは，話者により音声スペクトルの形状が違うことの一因である．VTLN は，話者の発声のスペクトルから声道長を求め，ある「標準的な」声道長から生じるスペクトルに変換する方法であり，理想的な環境下では，一単語発声程度の少量のデータから話者の声道長が正確に求められ効果が確認されている．ただし，実環境では発声変形，周囲雑音の影響から声道長推定の精度が低いことが問題となっている．そこで，あらかじめ声道長パラメータを複数用意し，話者ごとに最適なパラメータを選択する手法（ML-VTLN）が提案され，多くの認識システムで用いられている [5], [14], [16], [17]．しかしながら，この手法では，学習時，認識時とも，用意されたパラメータの数だけ同一発声に対するゆう度計算が必要となり計算量が多いこと，また，用意されたパラメータに必ずしも話者に対し最適なパラメータが存在するとは限らないこと，が問題となる．

近年，話者適応化を前提とした学習（Speaker Adaptive Training; SAT）という概念が提唱されている [1], [4], [11], [15]．これは，話者適応化を必ず行うことを前提とした場合，話者適応化の初期モデルとして，必ずしも話者の違いによる発声の揺らぎを表現した不特定話者モデルを用いる必要はなく，ある仮想的な「標準話者」のモデルを用意すればよいはずである，という考え方に基づいている．SAT では，多数話者の発声データを用いた事前の学習の際に，各々の学習話者に対し話者適応化を行い，その結果作成された写像の逆写像を求める．そして，その逆写像を用いて変換された発声データを入力として「標準話者」のモデルが学習される．写像としては，前述のアフィン変換がしばしば用いられている．しかしながら，アフィン変換は比較的パラメータ数が多く，少量の発声（例えば一単語発声など）では写像の推定精度が低い．

話者正規化と SAT は，ともに，入力発声から話者の違いにより生じる揺らぎを取り除く点で，同一の範疇（ちゅう）に属する手法と見ることができる．揺らぎが少なくなる分，モデル化の対象となる音響空間がコンパクトになるため，モデルの規模を小さくできる，同じ大きさのモデルにより高い表現力をもたせることができるという利点をもつ．また，他の要因から生じる揺らぎに対し，より精緻なモデリングが可能になると期待される．話者正規化においては，揺らぎを表現する写像を特徴抽出の段階で求めるのに対し，SAT ではパターンマッチングと組み合わせて求めている．これらの手法において重要な課題は，認識時にも写像を作成する必要があるため，実環境下において比較的少量の発声で，推定可能な写像を選択することである．

本論文では，声道長正規化を一つのパラメータを用いた線形写像で行う手法を提案する．パラメータは，HMM を用いたパターンマッチングの段階で，最ゆう推定で求める．ML-VTLN に比べ，あらかじめ複数のパラメータを用意する必要はなく，パラメータ選択も必要ないため，計算量は少ない．また，アフィン変換を用いた SAT に比べると，推定すべきパラメータ数が少なく，より少量の発声でのパラメータ推定が可能である．

次章で，提案手法のアルゴリズムを述べ，**3.** で評価実験結果について述べる．

2. 最ゆう推定を用いた声道長正規化

2.1 ワーピング関数とケプストラム変換

声道長の変換は，通常周波数軸上のワーピング関数として表される．一方，ケプストラムを特徴量とした音声認識において，認識や音響モデルの学習は，ケプストラム空間で計算されるゆう度を基準として行われる．そのため，声道長正規化を行う場合，ワーピング関数の推定はケプストラム空間で行うことが望ましい．そこで，本研究では，次式 (1) で示されるようなケフレンシ軸上の 1 次全域通過フィルタを周波数軸変換の関数として用いることとする [7]〜[9]．

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{1}$$

ここで，α は，$|\alpha| < 1$ の実数とする．また，z と $\hat{z}$ は，ω と $\hat{\omega}$ をそれぞれ変換前後の周波数として，$z = e^{j\omega}$，$\hat{z} = e^{j\hat{\omega}}$ である．式 (1) により周波数軸は，図 1 に示されるように，$\alpha < 0$ の場合低域に変換され，$\alpha > 0$ の場合高域に変換される．以後，α をワーピングパラメータと呼ぶ．

次に，変換前のケプストラム c_n を用い，変換後のケプストラム $\hat{c}_n$ を表す式の導出を説明する．ケプストラム c_n の z 変換を $S(z)$ とし，

$$S(z) = \sum_{m=0}^{\infty} c_m z^{-m} \tag{2}$$

と表す [10]．一方，$\hat{c}_n$ についても，その z 変換 $\hat{S}(z)$ を定義することで，式 (2) と同様の式を得ることができる．

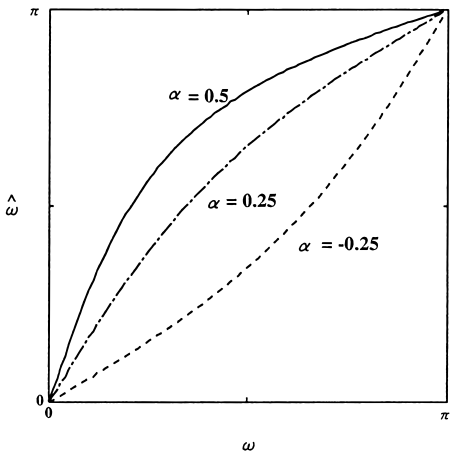


図 1　周波数ワーピング関数
Fig. 1　Frequency warping fuction.

$$\hat{S}(\hat{z}) = \sum_{m'=0}^{\infty} \hat{c}_{m'} \hat{z}^{-m'} \tag{3}$$

c_m の z 変換 $S(z)$ は，周波数 ω の関数である．$\hat{c}_m$ の z 変換 $\hat{S}(\hat{z})$ も同様に，周波数 $\hat{\omega}$ の関数である．ここで，$\hat{S}(e^{j\hat{\omega}}) \equiv S(e^{j\omega})$ とする．すなわち，周波数軸変換前の $S(e^{j\omega})$ と周波数軸変換後の $\hat{S}(e^{j\hat{\omega}})$ の値が，周波数軸上において $\omega, \hat{\omega}$ のときに等しくなるように，既知のパラメータである c_m を用いて $\hat{c}_m$ を決める．このとき，式 (2) と式 (3) の関係は次のようになる．

$$\sum_{m'=0}^{\infty} \hat{c}_{m'} \hat{z}^{-m'} = \sum_{m=0}^{\infty} c_m z^{-m} \tag{4}$$

このとき式 (4) の両辺に，$\hat{z}^{-(n+1)}/2\pi j$ をかけ，$\hat{z}$ についてコーシ積分

$$\frac{1}{2\pi j} \oint z^{n-1} dz = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases} \tag{5}$$

を行うと，c_n と $\hat{c}_n$ の関係を得ることができる．

$$\hat{c}_n = \sum_{m=0}^{\infty} c_m \frac{1}{2\pi j} \oint z^{-m} \hat{z}^{n-1} d\hat{z} \tag{6}$$

式 (1) を $\hat{z}$ について展開する．

$$\begin{aligned} z^{-1} &= \frac{\hat{z}^{-1} + \alpha}{1 + \alpha \hat{z}^{-1}} \\ &= (\hat{z}^{-1} + \alpha) \sum_{k=0}^{\infty} (-\alpha \hat{z}^{-1})^k \end{aligned} \tag{7}$$

式 (6) に式 (7) を代入し，各 c_n を抽出すると，$c_n, \hat{c}_n, \alpha$ だけの次式を得ることができる．(付録 **1.** 参照.)

$$\begin{aligned} \hat{c}_0 &= \sum_{m=0}^{\infty} \alpha^m c_m \\ \hat{c}_1 &= (1-\alpha^2) \sum_{m=1}^{\infty} m \alpha^{m-1} c_m \\ \hat{c}_2 &= c_2 + \alpha(-c_1 + 3c_3) \\ &\quad + \alpha^2(-4c_2 + 6c_4) \cdots \\ \hat{c}_3 &= c_3 + \alpha(-2c_2 + 4c_4) \\ &\quad + \alpha^2(c_1 - 9c_3 + 10c_5) \cdots \\ &\vdots \end{aligned} \tag{8}$$

式 (8) を，行列表現で表す．

$$\hat{\mathbf{c}} = \mathbf{A}_0 \mathbf{c} \tag{9}$$

ここで，$\hat{\mathbf{c}}$, $\mathbf{A}_0$, $\mathbf{c}$ は，次のとおりである．

$$\hat{\mathbf{c}} = \begin{pmatrix} \hat{c}_0 & \hat{c}_1 & \hat{c}_2 & \hat{c}_3 & \cdots \end{pmatrix}^t$$

$$\mathbf{A}_0 = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ 0 & 1-\alpha^2 & 2\alpha - 2\alpha^3 & \cdots & \cdots \\ 0 & -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

$$\mathbf{c} = \begin{pmatrix} c_0 & c_1 & c_2 & c_3 & \cdots \end{pmatrix}^t \tag{10}$$

式 (9) は，入力音声のケプストラム $\mathbf{c}$ をワーピングパラメータ α を用いて変換し，話者性によるホルマントのずれを補正したケプストラム $\hat{\mathbf{c}}$ を求める式であり，周波数軸のワーピングをケプストラム空間上の 1 次変換で表現している．式 (9) によるスペクトルの変換の様子を図 2 と図 3 に示す．$\alpha > 0$ の場合は高周波よりに，$\alpha < 0$ の場合は低周波よりに，変換されている．

2.2　ワーピングパラメータの最ゆう推定

本節では，データからワーピングパラメータ α を求める方法について説明する．最適な α を求める基準は，話者ごとに，観測系列 $\mathbf{O}$ の出現確率 $P(\mathbf{O}|\Theta)$ を最大にすることとする．なお，$\Theta \equiv (\theta, \alpha)$ であり，θ は，HMM のパラメータセットである．

$$\alpha = \arg\max_{\alpha} P(\mathbf{O}|\Theta) \tag{11}$$

従来，あらかじめ用意された複数の α の値に対

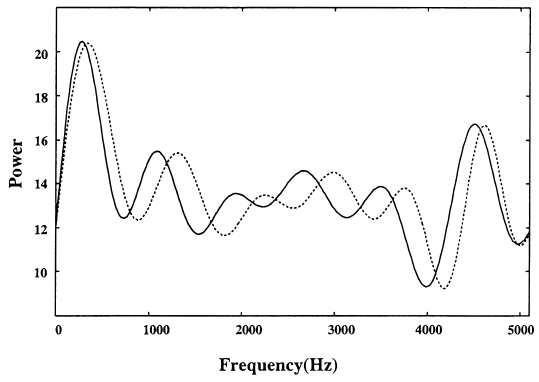


図 2　ワーピング関数によるスペクトル変換．$\alpha = 0.1$（実線が変換前のスペクトル，破線が変換後のスペクトル）
Fig. 2　Spectral transformation using frequency warping function. $\alpha = 0.1$. Original (thick line), transformed spectrum (dotted line).

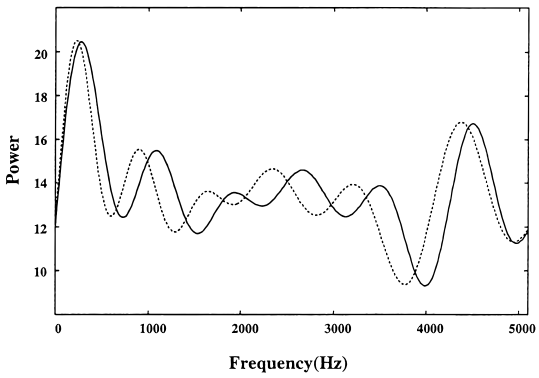


図 3　ワーピング関数による変換．$\alpha = -0.1$（実線が変換前のスペクトル，破線が変換後のスペクトル）
Fig. 3　Spectral transformation using frequency warping function. $\alpha = -0.1$. Original (thick line), transformed spectrum (dotted line).

し $P(\mathbf{O}|\Theta)$ を求め，$P(\mathbf{O}|\Theta)$ を最大とする α を選ぶ，ML-VTLN と呼ばれる方法が知られている [5], [14], [16], [17]．この方法では，α の値ごとに，$P(\mathbf{O}|\Theta)$ を計算する必要があるため，計算コストが膨大になる．また，演算量を抑えるために，あらかじめ用意する α の値を少なくした場合，選択された α と話者に最適なワーピングパラメータとの差が大きくなり，ホルマントの補正精度を十分に確保できないことも考えられる．

本節では，式 (9) を Baum-Welch アルゴリズムに組み入れ，ワーピングパラメータ α を推定するための定式化を行う．定式化にあたり，最大化すべき Q 関数（目的関数）を，

$$Q(\Theta', \Theta) = \sum_{j=1}^{J} \sum_{t=1}^{T} P(\mathbf{O}, q_t = j|\Theta') \log b_j(\hat{\mathbf{c}}_t) \tag{12}$$

とする．$P(\mathbf{O}, q_t = j|\Theta)$ は，モデル Θ が与えられたとき，時刻 t ($t = 1 \sim T$) に，状態が j ($q_t = j$) であり，観測系列 $\mathbf{O}$ が生成される同時確率を表す．ただし，$\mathbf{O}$ は，時刻 t に観測された特徴ベクトル $\mathbf{o}_t$ として，$\mathbf{O} = (\mathbf{o}_1 \mathbf{o}_2 \cdots \mathbf{o}_t \cdots)$ である．J は，全状態数を表す．式 (12) の，観測密度関数 $b_j(\hat{\mathbf{c}}_t)$ は，次の連続ガウス分布関数を仮定している．

$$b_j(\hat{\mathbf{c}}_t) = \frac{1}{\sqrt{(2\pi)^M |\mathbf{\Sigma}_j|}} \exp\left[-\frac{1}{2} (\hat{\mathbf{c}}_t - \mu_j)^t \mathbf{\Sigma}_j^{-1} (\hat{\mathbf{c}}_t - \mu_j) \right] \tag{13}$$

ここで，共分散行列 $\mathbf{\Sigma}_j$ は，σ_{mj}^2（m は次元，j は状態を表す）を対角成分にもち非対角成分は 0 とした対角行列，μ_j は，状態 j の平均ベクトルである．また，M は，ケプストラムの次元数とする．式 (12) を α について微分し，0 とおく．

$$\frac{\partial Q(\Theta', \Theta)}{\partial \alpha} = \sum_{j=1}^{J} \sum_{t=1}^{T} \frac{P(\mathbf{O}, q_t = j|\Theta') \dfrac{\partial b_j(\hat{\mathbf{c}}_t)}{\partial \alpha}}{b_j(\hat{\mathbf{c}}_t)} = 0 \tag{14}$$

式 (14) の解が，最適な α である．McDonough [7], [8] 等は，Newton 法を用いて α を求めている．しかし，式 (14) は，$\hat{\mathbf{c}}_t$ が α の多項式なため，複数の解が存在し（存在しないこともある），一意的に解くことはできない．ここで，学習の初期モデルとして用いられる HMM は，既に十分多数の話者の音声で学習されており，未知の話者に対しホルマントの位置が大きく逸脱することはないと仮定する．すなわち，α は十分小さい ($\alpha \ll 1$) とする．そして，式 (10) の $\mathbf{A}_0$ の 2 次以降の項を無視した行列

$$\mathbf{A} \equiv \begin{pmatrix} 1 & \alpha & 0 & 0 & \cdots \\ 0 & 1 & 2\alpha & 0 & \cdots \\ 0 & -\alpha & 1 & 3\alpha & \cdots \\ 0 & 0 & -2\alpha & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \tag{15}$$

を変換行列として用いることとする．式 (9) と式 (15) から式 (14) は，式 (16) のように変形することができる．

$$\sum_{j=1}^{J}\sum_{t=1}^{T}P(\mathbf{O},q_t=j|\Theta')\left[\sum_{m=1}^{M}\frac{1}{\sigma_{mj}^2}(\Delta c_{mjt}-\alpha\bar{c}_{mt})\bar{c}_{mt}\right]=0 \tag{16}$$

式 (16) を解くと，次のようになる．

$$\alpha=\frac{\sum_{j=1}^{J}\sum_{t=1}^{T}P(\mathbf{O},q_t=j|\Theta')\left[\sum_{m=1}^{M}\frac{1}{\sigma_{mj}^2}\Delta c_{mjt}\bar{c}_{mt}\right]}{\sum_{j=1}^{J}\sum_{t=1}^{T}P(\mathbf{O},q_t=j|\Theta')\left[\sum_{m=1}^{M}\frac{1}{\sigma_{mj}^2}\bar{c}_{mt}^2\right]} \tag{17}$$

このとき，次の記号を用いた．

$$\begin{aligned}&\Delta c_{mjt}=c_{mt}-\mu_{jm},\\&\bar{c}_{mt}=(m-1)c_{(m-1)t}-(m+1)c_{(m+1)t}\end{aligned} \tag{18}$$

c_{mt} は，時刻 t におけるケプストラムの第 m 次元の成分とする．ここで，Forward-Backward アルゴリズムより求めることができる占有度数 $\gamma_t(j)$（時刻 t，状態 j）を

$$\gamma_t(j)\equiv\frac{P(\mathbf{O},q_t=j|\Theta)}{\sum_{j=1}^{N}P(\mathbf{O},q_t=j|\Theta)} \tag{19}$$

のように定義する．式 (17) に式 (19) を代入することにより，

$$\alpha=\frac{\sum_{j=1}^{J}\sum_{t=1}^{T}\gamma_t(j)\left[\sum_{m=1}^{M}\frac{1}{\sigma_{mj}^2}\Delta c_{mjt}\bar{c}_{mt}\right]}{\sum_{j=1}^{J}\sum_{t=1}^{T}\gamma_t(j)\left[\sum_{m=1}^{M}\frac{1}{\sigma_{mj}^2}\bar{c}_{mt}^2\right]} \tag{20}$$

と記述できる．

式 (20) を HMM の学習と認識に組み入れることで，周波数軸上におけるホルマント位置の揺らぎを補正する声道長正規化を行うことができる．また，α は，HMM のパラメータの推定時に計算される占有度数 $\gamma_t(j)$ を用いるため，わずかな演算量の増加で計算を行うことができる．この手法を，VTLN-R（Vocal Tract Length Normalization using Rapid Maximum-Likelihood Estimation）と呼ぶ．ここでは，式 (20) の導出にケプストラムを用いているが，同様の計算でデルタケプストラムを用いた場合の定式化も可能である．

2.3 声道長正規化アルゴリズム

声道長正規化を用いた学習アルゴリズムは，図 4 に示すように次の四つのステップからなる．Step1 では，Baum-Welch アルゴリズムで，学習音声のケプストラム $\mathbf{C}_s$ を用い，占有度数の計算を行う．s は，話者を表すパラメータである．このとき，占有度数は不特定話者 HMM を用いて計算される．Step2 では，Step1 で求められた占有度数を用い，式 (20) を用いて各話者ごとの α_s を推定する．Step3 では，式 (9) を用い，声道長正規化されたケプストラム $\hat{\mathbf{C}}_s$ を求める．Step2 と Step3 は，学習の話者の数だけ繰り返し行われる．Step4 は，Step1 で計算された占有度数と Step3 で求められた $\hat{\mathbf{C}}_s$ を用いて HMM のパラメータの再推定を行う．更に，Step4 は，Step1 で用いる $\mathbf{C}_s$ を $\hat{\mathbf{C}}_s$ に置き換え，Step1 に戻る．2 回目以降の Step1 の処理は，Step4 で計算された HMM を用いて占有度数を計算する．

次に，認識アルゴリズムを説明する．認識アルゴリズムは，推定と，声道長正規化と認識の三つの動作に分かれる．推定では，図 4 の Step2 と同様に，推定用の音声から式 (20) を用いて α_s の推定が行われる．声道長正規化では，Step3 と同様に，認識音声のケプストラム $\mathbf{C}_s$ から，推定時に計算された α_s を用い，

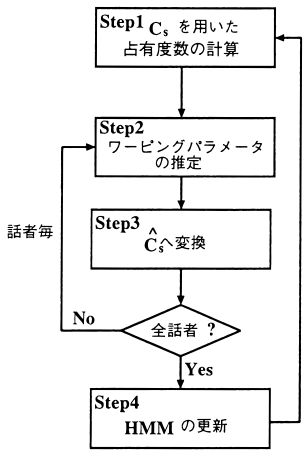


図 4　学習アルゴリズム
Fig. 4　Traning procedure.

声道長正規化されたケプストラム $\hat{\mathbf{C}}_s$ の計算が行われる．デルタケプストラムは，声道長正規化されたケプストラム $\hat{\mathbf{C}}_s$ を用いて計算される．認識では，すべての特徴量を用いて認識処理を行う．

3. 実　　験

3.1 実験条件

分析条件は，サンプリング周波数 11.025 kHz，帯域 300〜5000 Hz，フレーム間隔 16 ms で，メルケプストラム分析を用いた．特徴ベクトルは，正規化パワー差分，メルケプストラム 10 次元，メルケプストラムの変化量 10 次元の計 21 次元である．音響モデルは，半音節を認識単位とした不特定話者連続 HMM を用いた [13]．HMM の共分散行列は，対角成分のみを使用している．状態ごとの混合分布数は，2 である．学習には，音素バランス単語約 2000 単語，男性女性合わせて計 56 名の音声を用いた．声道長正規化の α_s の推定に用いた単語数と声道長正規化学習の有無，評価用音声の話者数と 1 話者当りの発声数，評価辞書について，表 1 に示す 2 種類の実験条件 A，B を用いた．まず，声道長正規化学習は，実験条件 A では行い，実験条件 B では行っていない．声道長正規化の α_s 推定用に用いた音声の単語数は，実験条件 A では，電子協 100 地名すべて用いた 100 単語とし，実験条件 B では，電子協 100 地名のうち 5 単語とした．ただし，α_s 推定用の単語の音声は，評価に用いた音声とは別の音声である．評価用音声の話者数は，実験条件 A では，男性女性各 45 の合計 90 名とし，実験条件 B では，男性女性各 14 の合計 28 名とした．評価用音声の 1 話者当りの音声数は，実験条件 A では，電子協 100 地名を 3 回ずつの合計 300 とし，実験条件 B では，電子協 100 地名のうち実験条件 B の推定用単語以外の 95 とした．認識辞書として，実験条件 A では，電子協 100 地名に任意に選択した 4900 単語を加えた 5000 単語の辞書を用い，実験条件 B では，電子協 100 地名に任意に選択した 1500 単語を加えた 1600 単語の辞書を用いた．認識の際の声道長正規化は，発声する内容が既知とした（教師あり）．

表 1　実 験 条 件
Table 1　Experimental condition.

	A	B
声道長正規化学習	あり	なし
推定単語数	100	5
評価話者数	90	28
1 話者の音声数	300	95
認識辞書	電子協 100 地名	
（加えた単語数）	4900	1500

3.2 従来方法との比較

本方式において，式 (15) の導出で用いた近似の妥当性を検証するために，従来の手法と比較を行った．従来の手法とは，複数の α_s を用意しておき，その中からゆう度が最大になるように選択する Zhan らによる手法 ML-VTLN である [17]．ただし，今回は比較のため，Zhan らの用いた関数を用いずに，式 (9) の変換を用いた．

本節の実験条件は，表 1 の A の実験条件を用いた．α_s の推定は，推定用の音声に現れるすべての音素を対象に 10 次元のケプストラムを用いて行った．声道長正規化学習の繰返し回数は 4 とした．表 2 に，比較実験における単語認識率を示す．値は，男性女性 45 名の平均と，男女 90 名の平均認識率である．表 2 の SI は，不特定話者 HMM の認識結果である．本節で行う，声道長正規化学習の初期 HMM として，SI の HMM を用いた．ML-VTLN の学習時における α_s は，±0.3 の範囲を 0.05 刻みで，合計 13 個の値からゆう度が最大になるものを選択した．認識時における α_s は，±0.5 の範囲を 0.05 刻で合計 21 個の値から，1 名につき 100 発声の推定用音声を用い，ゆう度を最大にするものを選んだ．VTLN-R1 は，前節で説明した学習アルゴリズムで声道長正規化を行い，認識時も声道長正規化を行った場合の実験結果である．

ML-VTLN における α_s 推定の刻幅は十分に小さく，式 (1) による周波数ワーピング関数を用いた声道長正規化での上限の認識性能とみなすことができる．表 2 から，ML-VTLN と VTLN-R1 による認識率は，それぞれ全体の平均で 79.3%と 79.2%であり，ML-VTLN と VTLN-R1 の結果はほぼ等しい．本方式の導出過程（式 (15)）で導入した近似による劣化はわずかなものであることがわかった．

3.3 パラメータの調整

声道長正規化の性能をより向上させるため予備実験を行った．実験は，α_s の推定に用いるケプストラム

表 2　従来法との比較（%）
Table 2　Word accuracy rate of ML-VTLN and VTLN-R1 (%).

話者	SI	ML-VTLN	VTLN-R1
男性	78.7	79.4	79.4
女性	78.8	79.1	79.0
平均	78.8	79.3	79.2

の次元数の調整と，α_s の推定に用いる音素の選別について，それぞれ行った．本節の実験条件は，表 1 の実験条件 B とした．

ワーピングパラメータ α_s の推定に用いる次元数についての実験結果を表 3 に示す．表 3 の SI は，認識時に声道長正規化を行わない場合の認識率である．10，6，5，4，3 は，認識時に，それぞれ 10 次元，6 次元，5 次元，4 次元，3 次元のケプストラムで α_s を推定した場合の結果である．ただし，本実験では，単語中の母音だけを α_s の推定に使用した．ここで，推定における次元数の設定方法を説明する．例えば，推定に使用するケプストラムの次元数を 5 次元とした場合，式 (20) は，式 (21) のようになる．

$$\alpha = \frac{\displaystyle\sum_{j=1}^{J}\sum_{t=1}^{T}\gamma_t(j)\left[\sum_{m=1}^{5}\frac{1}{\sigma_{mj}^2}\Delta c_{mjt}\bar{c}_{mt}\right]}{\displaystyle\sum_{j=1}^{J}\sum_{t=1}^{T}\gamma_t(j)\left[\sum_{m=1}^{5}\frac{1}{\sigma_{mj}^2}\bar{c}_{mt}^2\right]} \quad (21)$$

このとき，占有度数 $\gamma_t(j)$ の計算には，すべての特徴量（本実験の場合，メルケプストラム 1〜10 次元とメルケプストラムの変化量 1〜10 次元と正規化パワーの差分）を用いる．式 (21) で求められた α を用いて，入力された 1〜10 次元のケプストラムの声道長正規化を行う．

表 3 から，4 次元での認識性能が最も高くなっている．これから，ワーピングパラメータの推定においては，スペクトル包絡の細かな変化成分を含めて推定するよりも，スペクトル包絡の大局的な変化成分のみで推定する方が良いことがわかる．

上記の実験では，α_s の推定に母音を用いていたが，比較のため α_s の推定にすべての音素を用いた場合の評価を行った．実験結果を表 4 に示す．表 4 の VOWEL は，単語に含まれる母音に対応する HMM の状態だけを用いて α_s を推定した場合，ALL は，単語に含まれるすべての状態を用いて α_s を推定した場合の結果である．ここで，α_s の推定に使用するケプストラムの次元数を 4 次元とした．

実験の結果，表 4 より，母音のみを用いて推定する方が，認識性能が良くなることがわかった．

3.4 正規化の効果

本節では，正規化による学習の効果を検証するために，声道長正規化学習と CMN による正規化学習をそれぞれ単独で行った実験と，それらを組み合わせた実験を行った．本節の実験条件は，**3.2** と同じ実験条件 A を用いた．また，本節の声道長正規化における α_s の推定は，4 次元のケプストラムで母音だけを用いた．実験結果を表 5 に示す．表 5 は，各実験における単語認識率を示す．値は，男性女性 45 名と男女 90 名の平均の認識率である．

表 5 の SI は，**3.1** の実験条件で示される条件で学習を行った HMM の実験結果である．SA は，HMM に SI の実験で用いたものを使い，認識時にのみ声道長正規化を行った場合の実験である．VTLN-R2 は，前節で説明した学習アルゴリズムで声道長正規化学習を行い，認識時も声道長正規化を行った場合の実験結果である．表 2 の VTLN-R1 との違いは，α_s の推定を，VTLN-R1 が 10 次元のケプストラムとすべての音素を用いたのに対し，VTLN-R2 は 4 次元のケプストラムと母音だけを用いたことである．学習時の声道長正規化の初期 HMM として，SI の HMM を用いた．表 5 の CMN は，CMN を行った場合の実験結果である．学習の初期 HMM として，SI の HMM を用い，学習の入力として，話者ごとにすべての発声から求めた CM をケプストラムから差し引いたものを使用した．ここで，CM とは，学習音声の話者それぞれのケプストラムの長時間平均である．認識時の CM は，学習音声の全体から求めた CM を初期値とし，1 発声

表 3　ワーピングパラメータ α_s の推定に用いるケプストラム次元数と認識率（%）
Table 3　Word accuracy obtained with smaller cepstrum dimensions (%).

次元数	SI	10	6	5	4	3
男性	80.0	80.7	80.5	80.6	80.9	79.2
女性	77.1	77.5	77.4	78.3	78.7	78.9
平均	78.5	79.1	78.9	79.5	79.8	79.1

表 4　推定に用いる音素と認識率（%）
Table 4　Word accuracy of estimation using vowels (%).

音素	SI	VOWEL	ALL
男性	80.0	80.9	80.5
女性	77.1	78.7	78.4
平均	78.5	79.8	79.4

表 5　声道長正規化学習による認識率（%）
Table 5　Word accuracy rate of VTLN-R2 and its combination with CMN (%).

話者	SI	SA	VTLN-R2	CMN	V+C
男性	78.7	80.0	80.1	80.1	81.4
女性	78.8	76.6	80.4	81.4	82.3
平均	78.8	79.8	80.3	80.7	81.9

ごとにケプストラムの和を求め，総フレームの平均を計算し，逐次更新した．すなわち，$K+1$ 番目の発声に適用する CM は，式 (22) に示す，K 番目の発声までのケプストラムの総和の平均とする．

$$\begin{aligned}\mathbf{c}^{(K+1)} &= \bar{\mathbf{c}}^{(0)} + \sum_{k=1}^{K}\sum_{n=1}^{N^{(k)}} \mathbf{c}^{(k)}[n] \\ \bar{\mathbf{c}}^{(K+1)} &= \frac{\mathbf{c}^{(K+1)}}{\sum_{k=1}^{K} N^{(k)}}\end{aligned} \tag{22}$$

$\mathbf{c}^{(k)}[n]$ は，k 発声目の第 n フレームのケプストラムとする．K は発声回数，$N^{(k)}$ は，k 発声目のフレーム数とする．$\bar{\mathbf{c}}^{(K)}$ は，K 番目の発声のケプストラムから差し引く CM とする．V+C は，提案方式による声道長正規化と，CMN を組合せた場合の実験結果である．本実験では，声道長正規化の入力 $\mathbf{c}$ の代わりに，あらかじめ話者ごとに求めておいた CM を差し引いた $\mathbf{c}'$ を入力として用いることで，CMN との組合せを行った．認識は，ケプストラムから CM を差し引いた，$\mathbf{c}'$ を用いて α_s の推定を行い，α_s を用いて $\hat{\mathbf{c}}'$ を計算し，認識を行った．

表 5 の平均における不特定話者 HMM（SI）からの認識誤りの改善率は，声道長正規化による適応（SA）の場合 5.0%，声道長正規化（VTLN-R2）の場合 7.1%，CMN で 9.3%，声道長正規化と CMN の組合せ（C+V）の場合 14.6%である．本手法による声道長正規化において，学習時においても声道長正規化を行う方が，適応のみの手法より認識誤りの改善率が高い．また，CMN と組み合わせた場合，単独よりも高い認識誤りの改善率を得ることができた．

3.5 推定単語数と認識性能

ワーピングパラメータ α_s の推定に用いる単語数についての実験結果を表 6 に示す．実験に用いる実験条件は，**3.3** の条件と同じ，実験条件 B である．ただし，推定用音声は，電子協 100 地名の発声のうち，評価に使われていない 1〜5 発声について評価を行っている．推定は，用いるケプストラムの次元数を 4 とし，母音のみを使用している．表 6 の 1 から 5 は，1 発声から 5 発声を α_s の推定に用いた場合の結果である．SI は，認識時，学習時ともに声道長正規化を行わない場合である．ADAPT は，学習時には声道長正規化を行わない場合，VTLN-R2 は，前節の VTLN-R2 の HMM を用いた場合である．

表 6 に示すように，4 発声で認識性能が最も高くなるが，1 発声の推定単語数でも，最も高い認識性能に近い．これは，推定するべきパラメータが少ないため，1 発声程度でも十分であるためと考えられる．

表 6　α_s の推定に用いる単語数と認識率（%）
Table 6　Word accuracy obtained with small amount of adaptation data (%).

単語数	SI	1	2	3	4	5
ADAPT						
男性	80.0	80.7	80.9	80.9	81.0	80.9
女性	77.1	78.3	78.5	78.2	78.7	78.7
平均	78.5	79.5	79.7	79.5	79.8	79.8
VTLN-R2						
男性		80.5	81.4	81.1	81.4	81.1
女性		79.8	79.6	79.3	79.5	79.7
平均		80.2	80.5	80.2	80.5	80.4

4. むすび

ケプストラム空間上で声道長パラメータを最ゆう推定し，話者正規化を行う手法 VTLN-R を提案し，単独で 7.1%，従来の正規化の手法である CMN と組み合わせた場合 14.6%認識誤りが減少した．VTLN-R は，ML-VTLN など，あらかじめ声道長パラメータを複数用意し，そこから話者ごとに最適なパラメータを選択する手法に比べ，同等以上の性能をもち，かつ計算量が少い．

また，本手法は，ただ一つの自由パラメータをもつ制限された線形変換を用いた SAT とみなすことも可能である．アフィン変換を用いた手法に比べると，本手法では 1 単語発声でのパラメータ推定が可能であり，1 人の話者から極めて少量の発声しか得られない状況下で，特に効果的である．

本手法は，不特定話者認識の性能の著しく低い特異話者に対し，特に効果的と考えられ，今後は特異話者に対する性能評価を行いたい．また，実環境での使用を考慮し，ノイズ条件等の周囲環境が学習時と認識時に異なる場合における評価，及び，認識時に発声内容を指定しない教師なしの場合における評価を行いたい．

謝辞　本研究の機会を与えて頂きました NEC 情報通信メディア研究本部の渡辺部長と，吉田センター長，畑崎研究マネージャに感謝致します．また，本論文の内容について，御討論頂きました磯主任研究員，高木主任に感謝致します．最後に，有益なコメントをして頂きました匿名の査読者に感謝致します．

文　献

[1] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, “A compact model for speaker-adaptive

traning," Proc. ICSLP96, vol.2, FrP2L1.3, 1996.
[2] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am., vol.55, pp.1304–1312, 1974.
[3] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," Proc. ICASSP96, vol.1, pp.346–348, 1996.
[4] H. Jin, S. Matsoukas, R. Schwartz, and F. Kubaka, "Fast rubust inverse transform speaker adapted traning using diagonal transformations," Proc. ICASSP98, vol.2, pp.785–788, 1997.
[5] L. Lee and R.C. Rose, "Speaker normalization using efficient frequency warping procedures," Proc. ICASSP96, vol.1, pp.353–356, 1996.
[6] C.J. 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.
[7] J. McDonough and W. Byrne, "Speaker adaptation with all-path transforms," Proc. ICASSP99, no.2093, 1999.
[8] J. McDonough and W. Byrne, "Single-pass adapted traning with all-pass transforms," Proc. EUROSPEECH99, vol.6, pp.2737–2740, 1999.
[9] A.V. Oppenheim and D.H. Johnson, "Discrete representation of signals," Proc. IEEE, vol.60, pp.681–691, June 1972.
[10] A.V. Oppenheim and R.W. Shafer, Discrete-Time Signal Processing, Prentice-Hall, Englewood Cliffs, New Jersey, 1989.
[11] D. Pye and P.C. Woodland, "Experiments in speaker normalization and adaptation for large vocabulary speech recognition," Proc. ICASSP97, vol.2, pp.1047–1050, 1997.
[12] 篠田浩一，渡辺隆夫，"音声認識における自律的なモデル複雑度制御を用いた話者適応化，" 信学論（D-II），vol.J79-D-II, no.12, pp.2054-2061, Dec. 1996.
[13] 渡辺隆夫，磯谷亮輔，塚田　聡，"半音節を単位とするHMMを用いた不特定話者音声認識，" 信学論（D-II），vol.J75-D-II, no.8, pp.1281–1289, Aug. 1992.
[14] S. Wegmann, D. Macllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," Proc. ICASSP96, vol.1, pp.339–341, 1996.
[15] L. Welling, R. Haeb-Umbach, X. Aubert, and N. Harberland, "A study on speaker normalization using vocal tract normaliztion and speaker adaptive traning," ICASSP98, vol.2 , pp.797–800.
[16] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," Proc. ICASSP99, no.1436, 1999.
[17] P. Zhan and M. Westohal, "Speaker normalization based on frequency warping," Proc. ICASSP97, pp.1039–1042, 1997.

付　　録

1. 式 (8) の導出

本節では，式 (6) と式 (7) を用いて，式 (8) の導出を行う．式 (7) を式 (6) に代入すると，次の式になる．

$$\hat{c}_n = \sum_{m=0}^{\infty} c_m \frac{1}{2\pi j} \oint \left[(\hat{z}^{-1}+\alpha) \sum_{k=0}^{\infty} (-\alpha \hat{z}^{-1})^k \right]^m \hat{z}^{n-1} d\hat{z} \tag{A·1}$$

ここで，

$$F_{n,m}(\alpha) = \frac{1}{2\pi j} \oint \left[(\hat{z}^{-1}+\alpha) \sum_{k=0}^{\infty} (-\alpha \hat{z}^{-1})^k \right]^m \hat{z}^{n-1} d\hat{z} \tag{A·2}$$

とすると，式 (A·1) は，次の式になる．

$$\hat{c}_n = \sum_{m=0}^{\infty} c_m F_{n,m}(\alpha) \tag{A·3}$$

$n = 0, 1$ の場合の計算の手順を示す．任意の n 次元の計算も，同様の手順で求めることができる．

1.1 $n = 0$ の場合

$n = 0$ の場合，式 (A·3) は，次のようになる．

$$\hat{c}_0 = \sum_{m=0}^{\infty} c_m F_{0,m}(\alpha) \tag{A·4}$$

このとき，式 (A·4) の $F_{0,m}(\alpha)$ を $m = 0, 1, \cdots$ について計算することで，$\hat{c}_0$ の導出ができる．$m = 0$ の場合，

$$F_{0,0}(\alpha) = \frac{1}{2\pi j} \oint \hat{z}^{-1} d\hat{z} = 1 \tag{A·5}$$

となる．$m = 1$ の場合，

$$\begin{aligned} F_{0,1}(\alpha) &= \frac{1}{2\pi j} \oint \left[(\hat{z}^{-1}+\alpha) \sum_{k=0}^{\infty} (-\alpha \hat{z}^{-1})^k \right] \hat{z}^{-1} d\hat{z} \\ &= \frac{1}{2\pi j} \oint \left[(\hat{z}^{-1}+\alpha)(1-\alpha \hat{z}^{-1} \cdots) \right] \hat{z}^{-1} d\hat{z} \\ &= \frac{1}{2\pi j} \oint \left[\alpha + (1-\alpha^2)\hat{z}^{-1} \cdots \right] \hat{z}^{-1} d\hat{z} \\ &= \alpha \end{aligned} \tag{A·6}$$

となる．同様に，$m = 2, 3, 4, \cdot$ の場合も計算が可能であり，結局，式 (8) の $\hat{c}_0$ の導出ができる．

$$\hat{c}_0 = c_0 + \alpha c_1 + \alpha^2 c_2 + \cdots \tag{A·7}$$

1.2 $n=1$ の場合

$n=1$ の場合も，$n=0$ の場合と同様に計算ができる．

$$\hat{c}_1 = \sum_{m=0}^{\infty} c_m F_{1,m}(\alpha) \qquad \text{(A·8)}$$

ここで，$F_{1,m}(\alpha)$ が 0 にならない条件は，$[\cdots]^m$ の $\hat{z}$ の次数が -1 の場合である．$n=0$ の場合と同様に，$m=0,1,2,\cdots$ について計算を行う．$m=0$ の場合，$F_{1,0}(\alpha)=0$ となる．$m=1$ の場合，

$$\begin{aligned} F_{1,1}(\alpha) &= \frac{1}{2\pi j}\oint \left[(\hat{z}^{-1}+\alpha)\sum_{k=0}^{\infty}(-\alpha\hat{z}^{-1})^k\right] d\hat{z} \\ &= \frac{1}{2\pi j}\oint \left[(\hat{z}^{-1}+\alpha)(1-\alpha\hat{z}^{-1}\cdots)\right] d\hat{z} \\ &= \frac{1}{2\pi j}\oint \left[\alpha+(1-\alpha^2)\hat{z}^{-1}-\alpha\hat{z}^{-2}\cdots\right] d\hat{z} \\ &= (1-\alpha^2) \qquad \text{(A·9)} \end{aligned}$$

となる．$m=2$ の場合，

$$\begin{aligned} F_{1,2}(\alpha) &= \frac{1}{2\pi j}\oint \left[(\hat{z}^{-1}+\alpha)\sum_{k=0}^{\infty}(-\alpha\hat{z}^{-1})^k\right]^2 d\hat{z} \\ &= \frac{1}{2\pi j}\oint \left[\alpha+(1-\alpha^2)\hat{z}^{-1}+\cdots\right]^2 d\hat{z} \\ &= \frac{1}{2\pi j}\oint \left[\alpha^2+2\alpha(1-\alpha^2)\hat{z}^{-1}+\cdots\right] d\hat{z} \\ &= 2\alpha(1-\alpha^2) \qquad \text{(A·10)} \end{aligned}$$

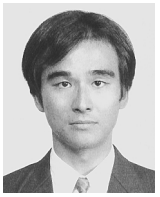
となる．同様に，$m=3,4,\cdot$ の場合も計算が可能であり，結局，式 (8) の $\hat{c}_1$ を導出することができる．

$$\hat{c}_1 = (1-\alpha^2)c_1 + 2\alpha(1-\alpha^2)c_2 + \cdots \qquad \text{(A·11)}$$

（平成 12 年 2 月 25 日受付，6 月 30 日再受付）

江森　正

平 5 東理大・理工・物理卒．平 7 慶大大学院物理学修士課程了．同年日本電気（株）入社．以来，音声認識の研究開発に従事．現在，NEC 情報通信メディア研究本部勤務．



篠田　浩一　（正員）

昭 62 東大・理・物理卒．平 1 同大大学院修士課程了．同年日本電気（株）入社．以来，音声認識の研究開発に従事．平 9〜10 ルーセントテクノロジー・ベル研究所客員研究員．平 9 日本音響学会粟屋学術奨励賞，平 10 本会論文賞各受賞．現在，NEC 情報通信メディア研究本部勤務．日本音響学会，IEEE 各会員．