音声言語情報処理 29－9
(1999. 12. 20)

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
NEC C&C メディア研究所

川崎市宮前区宮崎4丁目1-1
044-856-2304
t-emori@ccm.cl.nec.co.jp, shinoda@ccm.cl.nec.co.jp

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、声道長パラメータを用いた話者正規化の手法が提案されている。本稿では、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表現し、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予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、などの利点がある。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

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

Tadashi EMORI Koichi SHINODA
NEC C&C Media Research Laboratory

4-1-1 Miyazaki Miyamae-ku Kawasaki
044-856-2304
t-emori@ccm.cl.nec.co.jp, shinoda@ccm.cl.nec.co.jp

Abstract

In recent works, vocal tract length normalization methods which achieve a remapping of the frequency axis using warping functions have been proposed for a large vocabulary speech recognition system. In this work, we introduce an estimation method of the parameter characterizing individual speakers, using the remapping of the frequency axis in cepstrum domain derived from all-pass transforms. In Japanese 5000-word task speech recognition experiments, we report reductions in word error rate of 7.1% absolute. When the normalization method is combined with CMN, word error rate reduction is 14.6%.

key words speech recognition, Hidden Markov Model, speaker normalization, vocal tract length, maximum-likelihood estimation

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。隠れマルコフモデルは、様々な要因による発声のゆ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。このような特徴をもつため、話者の違いを発声のゆらぎの一つと捉え、多数話者の発声データを用いて確率分布を学習することにより、誰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、話者性から生じる発声のゆ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。

話者正規化は、特徴抽出の段階で、話者性から生じる発声のゆ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)[1]、声道長正規化 (Vocal Tract Length Normalization; VTLN)[2] 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、話者性のみならず、雑音、反響、回線などの違いにより生じる発声の音響的特徴の変化に比べ十分長時間のゆ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ゆらぎを取り除く方法である。話者の声道長の違いにより、声道 (Vocal Tract) の共鳴周波数 (フォルマント周波数) が異なるため、スペクトルの形状が話者により異なることが知られている [2]。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」の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。したがって、予め声道長パラメータを複数用意し、話者ごとにもっとも適当なパラメータを選択するという方式 (ML-VTLN) が提案され、多くの認識システムで用いられている [3, 7, 8, 9]。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、ゆう度計算が必要となり、計算量が増大すること、あるいは、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。

本稿では、声道長正規化を 1 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、パターンマッチングの段階で、HMM の最ゆう推定で求める。したがって、ML-VTLN に比べ、予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必要な計算量は少ない。また、アフィン変換を用いた SAT に比べると、推定の対象となるパラメータが少なく、少量の発声でのパラメータ推定が可能であるという利点をもつ。

次章で、提案手法のアルゴリズムを述べ、第 3 章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、通常スペクトル上のワーピング関数として表される。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。ここでは、周波数軸の変換をケプストラムの z 変換を用い、z 空間上で式 (1) で示される 1 次全域通過フィルタの位相特性を用いる [4, 5]。

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

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、$\alpha > 0$の場合高域に変換する。以後、この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をc_nを用い、変換後のケプストラムを$\hat{c}_n$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 $S(z)$ とした場合、

$$S(z) = \sum_{n=-\infty}^{\infty} c_n z^{-(n+1)}, \tag{2}$$

と表すことができる。式 (2) の逆変換は、コーシー積分

$$\frac{1}{2\pi j}\oint z^{-(n+1)}dz = \begin{cases} 1 & n=0 \\ 0 & n\neq 0 \end{cases}, \tag{3}$$

を用いることで求めることができる。

$$c_n = \frac{1}{2\pi j}\oint S(z)z^{-(n+1)}dz \tag{4}$$

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 (4) と同様の式を得ることができる。

$$\hat{c}_n = \frac{1}{2\pi j}\oint \hat{S}(z)z^{-(n+1)}dz \tag{5}$$

ここで、$\hat{S}(z) \equiv S(\hat{z})$ と定義する。すなわち、$\hat{S}(z)$ は、c_nを$\hat{z}$によって変換したものと仮定する。

$$\hat{S}(z) \equiv S(\hat{z}) = \sum_{n=-\infty}^{\infty} c_n \hat{z}^{-(n+1)} \tag{6}$$

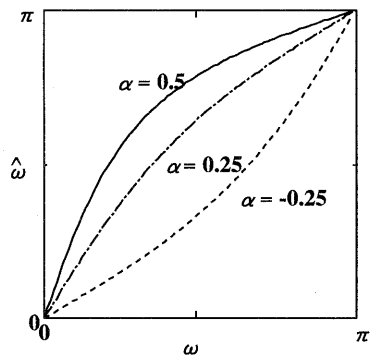

図 1: ワーピング関数

式 (6) を式 (5) に代入すると、c_nと$\hat{c}_n$の関係を得ることができる。

$$\hat{c}_n = \sum_{m=-\infty}^{\infty} c_m \frac{1}{2\pi j} \oint \hat{z}^{-(m+1)} z^{-(n+1)} dz \qquad (7)$$

一方、式 (1) を z について展開する。

$$\begin{aligned} \hat{z} &= \frac{z-\alpha}{1-\alpha z}, \\ &= (z-\alpha)\sum_{k=0}^{\infty}(\alpha z)^k, \\ &= \sum_{k=0}^{\infty}(\alpha^k z^{k+1} - \alpha^{k+1} z^k). \end{aligned} \qquad (8)$$

式 (7) に式 (8) を代入し、各 c_nを抽出すると、c_n,$\hat{c}_n$,αだけの次式を得ることができる。

$$\hat{\mathbf{c}} = \mathbf{A}\mathbf{c}. \qquad (9)$$

但し、$\hat{\mathbf{c}}, \mathbf{A}, \mathbf{c}$ は、次のとおりである。

$$\begin{aligned} \hat{\mathbf{c}} &= \begin{pmatrix} \hat{c}_0 & \hat{c}_1 & \hat{c}_2 & \hat{c}_3 & \cdots \end{pmatrix}^T, \\ \mathbf{A} &= \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ 0 & 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots & \cdots \\ 0 & -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}, \\ \mathbf{c} &= \begin{pmatrix} c_0 & c_1 & c_2 & c_3 & \cdots \end{pmatrix}^T. \end{aligned} \qquad (10)$$

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、話者性によるフォ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。図 2と図 3は、15 次元のケプストラムでリフタリングを行ったスペクトルである。前述の通り、$\alpha > 0$ 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。

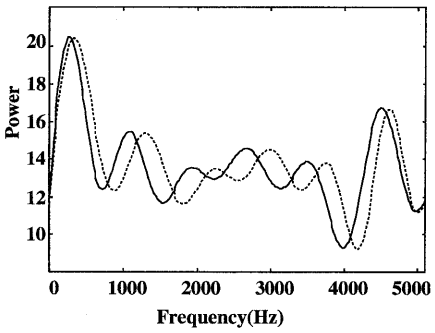

図 2: $\alpha = 0.1$ の場合:実線が変換前のスペクトル、破線が変換後のスペクトル

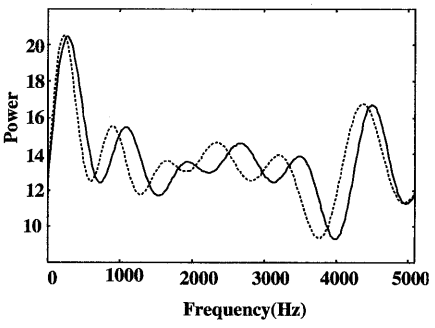

図 3: $\alpha = -0.1$ の場合:実線が変換前のスペクトル、破線が変換後のスペクトル

2.2 ワーピングパラメータの最ゆう推定

本節では、データから話者毎の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、話者毎に、モデルθに対し、観測系列 $\mathbf{o}_t$、状態系列 $\mathbf{q}$ のときの同時確率 $P(\mathbf{o}_t, \mathbf{q}|\Theta)$ を最大にすることである。なお、$\Theta \equiv (\theta, \alpha)$ とする。

$$\alpha = \arg\max_{\alpha} P(\mathbf{o}_t, \mathbf{q}|\Theta). \qquad (11)$$

従来、予め用意された複数のαの値を用いて $P(\mathbf{o}_t, \mathbf{q}|\Theta)$ を求め、最大になる値をαとして選ぶ、ML-VTLN と呼ばれる方法が知られている [3, 7, 8, 9]。このような方法では、αの値毎に、$P(\mathbf{o}_t, \mathbf{q}|\Theta)$ 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、演算量をおさえるために、予め用意するαの値を少なくした場合、フォ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、最ゆう推定法により、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、最尤推定の最大化すべきQ関数(目的関数)を、

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

とする。$P(\mathbf{o}_t, q_t = j|\Theta)$は、モデル$\Theta$が与えられたとき、時刻$t(t = 1 \sim T)$における状態が$j(q_t = j)$で、観測系列$\mathbf{o}_t$を生成する場合の同時確率を表す。$J$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。

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

ここで、共分散行列$\mathbf{\Sigma}_j$は、$\sigma_{mj}^2(m$は、次元を表す)を対角成分にもち非対角成分は0とした対角行列とし、μ_jは、状態jの平均分布ベクトルとする。また、ケプストラムの次元数をMとする。式(12)をαについて微分し、0とおく。

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

式(14)の解が、最適なαの候補である。McDonough[4]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の多項式になため、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。そこで、複数の話者で学習された音響モデルは、未知の話者に対し、フォ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。このとき、式(10)の$\mathbf{A}$は、2次以降の項を無視して、

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

とすることができる。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。

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

式(16)を解くと、次のようになる。

$$\alpha = \frac{\displaystyle\sum_{j=1}^{J}\sum_{t=1}^{T} \gamma_t(j) \left[\sum_{m=1}^{M} \frac{1}{\sigma_{mj}^2}\Delta c_{mjt}\bar{c}_{mt}\right]}{\displaystyle\sum_{j=1}^{J}\sum_{t=1}^{T} \gamma_t(j) \left[\sum_{m=1}^{M} \frac{1}{\sigma_{mj}^2}\bar{c}_{mt}^2\right]}. \tag{17}$$

このとき、次の記号を用いた。

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

ここで、$\gamma_t(j)$は、Forward-Backwardアルゴリズムより求めた、ある時刻tにおける状態jに存在する事後確率である。

式(17)をHMMの学習の過程に組み入れることで、周波数軸上におけるフォルマント位置のゆらぎを補正した声道長正規化学習を行うことができる。HMMのパラメータの推定時に計算される事後確率$\gamma_t(j)$を用いるため、わずかな演算量の増加でαの計算を行うことができる。

2.3 声道長正規化アルゴリズム

声道長正規化学習のアルゴリズムについて説明する。声道長正規化学習のアルゴリズムは、図4に示すように次の5つのステップからなる。Step1では、Baum-Welchアルゴリズムで、入力された学習音声のケプストラム系列$\mathbf{C}_s$からHMMのパラメータを推定する。ここでsは、話者を表すパラメータである。Step2では、Step1で求められた話者毎の事後確率を用い、式(17)から各話者毎に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系列$\hat{\mathbf{C}}_s$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、HMMのパラメータの更新($\theta \rightarrow \theta'$)と、声道長正規化されたケプストラムの更新($\hat{\mathbf{C}}_s \rightarrow \mathbf{C}_s$)を行いStep1へ飛ぶ。この手法を、VTLN-R(Vocal Tract Length Normalization using Rapid Maximum-Likelihood Estimation)と呼ぶ。

次に、認識アルゴリズムを説明する。認識アルゴリズムは、α_sを推定する過程と、α_sを用いて声道長正規化したケプストラム系列$\hat{\mathbf{C}}_s$を求める過程と、$\hat{\mathbf{C}}_s$を用いて認識を行う過程がある。まず、認識におけるα_sは、図4のStep2と同様に、話者毎に推定用の音声から式(17)を用いて推定される。次に、Step3と同様に、認識音声のケプストラム系列$\mathbf{C}_s$から、先の過程で計算されたα_sを用い、話者毎に声道長正規化されたケプストラム系列$\hat{\mathbf{C}}_s$を求める。次に、$\hat{\mathbf{C}}_s$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11.025kHz、帯域300~5000Hz、フレーム間隔16ms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[6]。HMMの共分散行列は、対角成分のみを使用している。状態毎の混合分布数は、2である。学習には、音素バランス単

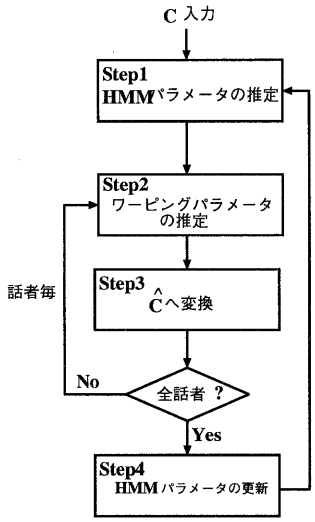

図 4: 学習アルゴリズム

語約 2000 単語、成人男性 18 名、成人女性 17 名、老人 13 名、子供 8 名の計 56 名の音声を用いた。評価用音声は、成人男性 45 名、成人女性 45 名、電子協 100 地名を 3 回づつの発声したものを用いた。認識におけるワーピングパラメータの推定用に、成人男性 45 名、成人女性 45 名、電子協 100 地名の音声を、評価用とは別に用意した。認識辞書には、電子協 100 地名単語を含む 5000 単語の辞書を用いた。

3.2 予備実験

声道長正規化の実験を行う前に、α_sの推定に使用するケプストラムの次元数を決めるための予備実験を行った。本節では、学習音声に声道長正規化は行わず、認識時の音声にのみ声道長正規化を行った (通常の適応化に対応)。HMM として前節の条件で学習を行ったものを、評価音声として男女 6 名の電子協 100 地名の音声を、辞書として電子協 100 地名をそれぞれ用いた。発声数は、1 名につき 100 発声で、α_sを推定する単語数は 1 単語とした。認識方法は、予め推定したα_sによって声道長正規化したケプストラムに対し、認識を行った。話者 12 名の認識率の平均を表 1に示す。表 1の NA は、認識時に声道長正規化を行わない場合、10D、6D、5D、4D は、認識時に、それぞれ 1 ～ 10 次元、1 ～ 6 次元、1 ～ 5 次元、1 ～ 4 次元のケプストラムでα_sを推定し、声道長正規化を行い、認識した場合の結果を表す。結果、4 次元での認識性能が最も性能が高くなっている。これは、ワーピングパラメータの推定においては、スペクトルほう絡の細かな変化成分を含めて推定するよりも、スペクトルほう絡の大

表 1: α_sの推定に使用するケプストラム次元数と認識率 (%)

次元数	NA	10D	6D	5D	4D
認識率 (%)	82.0	83.1	82.2	84.8	85.5

表 2: 認識時のみ声道長正規化による誤認識率 (%)

話者	SI	SA
男性	21.3	20.6
女性	21.2	20.9
平均	21.2	20.7

局的な変化成分のみで推定する方が良いためと考えられる。以降の実験では、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。更に、単語中の母音だけをα_sの推定に使用した。

話者男女 45 名の計 90 名について認識時にだけ声道長正規化を行う実験を行った。結果を表 2に示す。表 2の数値は、単語誤認識率 (%) を表す。表 2の SI は、前節の実験条件で学習を行った HMM の実験結果である。男性女性 45 名づつの平均と、男女 90 名の平均の誤認識率でる。1 名あたりの評価音声は、電子協 100 地名の 3 回発声で、合計 300 発声である。SA は、HMM に SI の実験で用いたものを使い、認識時に声道長の適応を行った場合の実験である。α_sの推定は、推定用の電子協 100 地名の発声を用い、式 (17) で推定を行った。表 2から、声道長の適応により、認識誤りが全体で、5.0% 減少している。

3.3 実験結果

以下、提案方式を用いた声道長正規化学習の実験結果を示す。実験結果を表 3に示す。表 3は、各実験における単語誤認識率を示す。

表 3の SI は、前節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。値は、それぞれ男性女性 45 名づつの平均と、男女 90 名の平均の誤認識率でる。

ML-VTLN 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 等による手法を用いて、声道長正規化および、認識を行ったものである [9]。ただし、今回は比較のため、ワーピング関数として、Zhan 等の用いた関数を用いずに式 (1) を用いた。ワーピングパラメータα_sは、予め SI で使用した HMM を用いて推定した値を中心とし、前後 0.3 づつ 0.05 刻みで値を振り、推定を行った。認識時におけるα_sは、1 名につき 100 発声の推定用の音声で認識実験を行い、最も認識性能が良くなるようなα_sを選び、実験用の音声に適用した。認識時のα_s推定の範囲は、$|\alpha_s| < 0.5$ で、0.05 刻み

表 3: 声道長正規化による誤認識率 (%)

話者	SI	ML-VTLN	VTLN-R	CMN	VTLN-R +CMN
男性	21.3	20.6	19.9	19.9	18.6
女性	21.2	20.9	19.6	18.6	17.7
平均	21.2	20.7	19.7	19.3	18.1

とした。表 3から、SI から男性 3.1%、女性 1.6%、全体で、2.4% の認識誤りの改善が得られた。

VTLN-R 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。声道長正規化学習の初期 HMM として、SI の HMM を用いた。表 3から、男性 6.6%、女性 7.7%、全体で 7.1%の認識誤りの改善が得られた。これは、声道長正規化を認識時だけ行う場合よりも、改善の割合が大きい。また、ML-VTLN よりも、改善の割合が大きい。

表 3の CMN は、比較のため CMN による平均正規化学習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 を用いた。学習の入力として、話者毎に全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、予め実験音声とは別の音声データの CM を初期値とし、1 発声毎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 (19) の様に、K番目の発声までのケプストラムの総和の平均である。

$$\begin{aligned} \mathbf{c}^{(K+1)} &= \bar{\mathbf{c}}^{(0)} + \sum_{k=1}^{K} \sum_{n=1}^{N^{(k)}} \mathbf{c}^{(k)}[n] \\ \bar{\mathbf{c}}^{(K+1)} &= \frac{\mathbf{c}^{(K+1)}}{\sum_{k=1}^{K} N^{(k)}} \end{aligned} \tag{19}$$

Kは発声回数、$N^{(k)}$は、k発声目のフレーム数である。$\bar{\mathbf{c}}^{(K)}$は、K番目の発声のケプストラムから差し引く CM である。実験の結果、全体で認識誤りが、9.3%減少した。

VTLN-R+CMN は、提案方式による声道長正規化学習と、CMN による平均正規化学習を組合せた場合の実験結果である。本実験では、声道長正規化学習の入力 $\mathbf{c}$ のかわりに、予め話者毎に求めておいた 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。男性 12.5%、女性 16.7%、全体で 14.6%の認識誤りの改善が得られた。これは、単独に声道長正規化や CMN を行うよりも効果がある。

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、話者正規化学習を行う手法 VTLN-R を提案し、単独で 7.1%、CMN と組み合わせた場合 14.6%誤りが減少した。VTLN-R は、従来の ML-VTRN など、予め声道長パラメータを複数用意し、そこから話者毎に最適なパラメータを選択する手法に比べ、計算量が少いという利点がある。

本手法は、話者正規化の観点からは、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、話者正規化学習と見なすことも可能である。アフィン変換などのより一般的な線形変換を用いた手法に比べると、極めて少量の発声でも頑健なパラメータ推定が可能であるという特徴をもち、一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、特に効果的と考えられる。

参考文献

[1] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, Vol. 55, pp. 1304-1312, 1974.

[2] Ellen Eide, Herbert Gish, "A Parametric Approach to Vocal Tract Length Normalization, "Proc. ICASSP96 , Vol.1, pp346-3483, 1996

[3] Li.Lee and Richard C.Rose, "Speaker Normalization Using Efficient Frequency Warping Procedures, "Proc. ICASSP96 , Vol.1, pp.353-356, 1996

[4] Jhon McDonough, Willam Byrne, "Speaker Adaptation With All-Path Transforms, "Proc. ICASSP99 , paper number 2093, 1999

[5] A.V.Oppenheim and D.H.Johnson:"Discrete representation of signals, "Proc. IEEE, **60**, pp. 681-691, June, 1972

[6] 渡辺, 磯谷, 塚田 ," 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

[7] Steven Wegmann, Don Macllaster, Jeremy Orloff and Barbra Peskin, "Speaker Normalization On Conversational Telephone Speech, "Proc. ICASSP96 , Vol.1, pp.339-341, 1996

[8] L.Welling, S.Kanthak and H.Ney, "Improved Methods For Vocal Tract Normalization, "Proc. ICASSP99 , Paper number 1436, 1999

[9] Puming Zhan, Martin Westohal, "Speaker Normalization Based On Frequency Warping, "Proc. ICASSP97 , pp.1039-1042, 1997