音声言語情報処理 19－5
(1997. 12. 11)

正準相関分析法を音素代表ベクトルに適用した教師なし話者正規化

桜木美春　　有木康雄
龍谷大学 理工学部
〒 520-21 大津市瀬田大江町横谷 1-5

あらまし 従来の不特定話者 HMM は、観測空間上で個々の話者空間の存在を無視して作成されているため、出力確率の分布が広がり、認識率が低下するという問題が生じていた。この問題に対して本研究では、各話者ごとに話者空間を設定し、観測正準相関分析法を用いて基準話者と入力話者を対応づけることにより、HMM の出力確率分布の広がりを押さえている。また、教師あり話者正規化においては、入力話者は基準話者と同じ発話を話さなければならないという問題がある。この問題に対して本研究では、Viterbe アルゴリズムにより各話者の発話を音素ごとにセグメンテーションし、各話者の音素毎に代表ベクトルを求め、この代表ベクトルの対応付けを話者間で行う教師なし話者正規化法を提案している。

キーワード 不特定話者 HMM, 教師あり話者正規化, 教師なし話者正規化, Viterbe アルゴリズム

Unsupervised Speaker Normalization by Canonical Correlation Analysis to Phoneme Representation Vectors

Miharu Sakuragi and Yasuo Ariki
Faculty of Science and Technology, Ryukoku University
1-5, Yokotani Ooe-Cho Seta Otsu-Shi 520-21, Japan

Abstract Conventional speaker-independent HMMs ignore the speaker differences and collect speech data in an observation space. This causes a problem that the output probability distribution of the HMMs becomes vague so that it deteriorates the recognition accuracy. To solve this problem, we construct the speaker subspace for an individual speaker and correlate them by o-space canonical correlation analysis between the standard speaker and input speaker. In order to remove the constraint that input speakers have to speak the same sentences as the standard speaker in the supervised normalization, we propose in this paper an unsupervised speaker normalization method which automatically segments the speech data into phoneme data by Viterbi decoding algorithm and then associates the mean feature vectors of phoneme data between speakers by o-space canonical correlation analysis.

key words speaker-independent HMMs, supervised speaker normalization, unsupervised speaker normalization, Viterbe decoding algorithm

1 はじめに

大語彙音声認識では一般的に、不特定話者HMMが広く使われている。しかし、不特定話者HMMは複数話者の音声データを大量に用いて作成されるため、HMMの確率分布の広がりが大きくなり、認識率が低下するという問題が生じる。これは、図1に示すように、話者Aと話者Bには、それぞれの音素構造を表現した話者固有の空間があるにも関わらず、従来の不特定話者HMMではこの話者空間を無視し、観測空間上で音声データを処理しているためである。

この問題に対して、各話者の音声データを用いて話者空間を構築し、音声データを話者空間へ射影して得られる、話者正規化データを用いて、不特定話者HMMを構築する方法が考えられる[1][2][3][4]。

話者正規化において、我々はこれまで正準相関分析法[5]を用いて、2人の話者の話者空間を対応づける方法について提案してきた[1][2]。この方法は、元来DPマッチングによる単語認識において、話者適応化法としてK.Choukriによって提案されたものである[6]。しかし、この手法は、2人の話者のうちモデルとなる基準話者の話者空間が入力話者の音声データに依存して形成されるため、HMMを用いて認識する場合には、入力話者ごとに基準話者のHMMを作り直さなければならないと言った問題が生じる。

この問題に対して、我々はCLAFIC正準相関分析法を用いた話者正規化法を提案している[2]。これは、始めに基準話者Aの話者空間をCLAFIC(Class featureing information compression)法を用いて作り、それから正準相関分析法によって入力話者Bの話者空間を作るものである。この方法を用いることにより、多数の話者の音声データを基準話者Aの話者空間に射影して正規化することができる。しかし、正準相関分析法の根本的な問題として、基準話者Aと入力話者Bは、各々の話者空間を作成するために、同じ発話内容を発声する必要がある。これは教師あり話者正規化法と呼ばれる手法では避けることのできない問題である。もし、この問題を取り除くことができれば、入力話者Bは基準話者Aの発話内容にとらわれず、自由に発話しても、基準話者Aに正規化することができる。これは教師なし話者正規化法と呼ばれる。

これまで、教師なし話者適応化法は報告されている[7][8]が、本稿では、話者正規化法として、どんな話者が何を発話してもよい、教師なし話者正規化法について提案し、従来法の教師あり話者正規化法との比較実験を音素認識実験により行なった。実験の結果、提案手法である教師なし話者正規化により、従来法である教師あり話者正規化とほぼ同程度の認識率が得られた。

2 教師あり話者正規化法

2.1 正準相関分析法

図1に示すように、観測空間で観測される話者Aの音声データをX_A、話者Bの音声データをX_Bとする。音声データX_Aは、話者Aの音声波形を時刻tで短時間スペクトル分析した特徴ベクトルx_{At}の時系列である。従って、音声データX_Aは、特徴ベクトルx_{At}^T、$(1 \leq t \leq M)$を行ベクトルとする行列として表され、列は周波数$i, (1 \leq i \leq N)$に対応している。

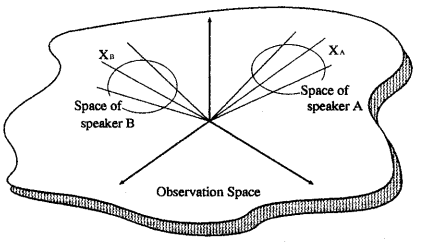

図1: 観測空間と話者空間

正準相関分析法は、話者正規化および適応化においてよく知られた方法である。正準相関分析法の手順を以下に示す。

STEP(1) 話者Aの音声データ1文と話者Bの音声データ1文をDPマッチングして、特徴ベクトル間の対応付を行う。これを複数の文に適用して音声データの行列X_A, X_Bを求める。

STEP(2) X_AとX_BをQR分解して、$X_A = QR$、$X_B = PS$を求める。

STEP(3) $\Omega = Q^T P$を求め、$\Omega\Omega^T$を固有値分解して固有値の大きい順にv'_{Ai}を求める。同様に$\Omega^T\Omega$を固有値分解して固有値の大きい順にv'_{Bi}を求める。$v_{Ai} = R^{-1}v'_{Ai}$、$v_{Bi} = S^{-1}v'_{Bi}$として相関の高い斜交軸v_{Ai}、v_{Bi}を順次求める。話者Aと話者Bの音声データを、こうして得られた各々の話者空間に射影して正規化データを得る[1]。

2.2 観測正準相関分析法

正準相関分析では入力話者Aの話者空間が、基準話者Bの音声データに依存して形成される。このため、HMMを用いて認識する場合には、入力話者毎に基準話者のHMMを作り直さなければいけないといった問題が生じる。この問題に対し、我々は観測正準相

関分析法による話者正規化を提案した。これは基準話者 A の話者空間を観測空間に固定し、基準話者 A と入力話者 B の相関が最大になるような入力話者 B の話者空間の軸を求めるものである。観測正準相関分析法の手順を以下に示す。

STEP(1) 基準話者 A の音声データ1文と入力話者 B の音声データ1文をDPマッチングして特徴ベクトル間の対応付を行う。これを複数の文に適用して音声データの行列 X_A, X_B を求める。

STEP(2) 観測空間を基準話者 A の正規直交基底として固定する。

STEP(3) 入力話者 B の音声データ X_B を用いて、基準話者 A の軸に相関の高い軸 v_B を次式により求める [1]。

$$v_B = \frac{\sqrt{C}\Sigma_{22}^{-1}\Sigma_{21}v_A}{\sqrt{v_A^T\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}v_A}} \qquad (1)$$

ここで、Σ_{12}、Σ_{21} はそれぞれ基準話者 A、入力話者 B の相互相関行列である。また Σ_{22} は入力話者 B の自己相関行列 であり、C は軸 v_A の分散を表している。基準話者 A と入力話者 B の音声データを、こうして得られた各々の話者空間に射影して正規化データを得る。

3 教師なし話者正規化法

3.1 教師あり話者正規化法の問題点

上で述べた観測正準相関分析を用いた教師あり話者正規化には、次のような問題点があると考えられる。

(1) 発話した文の長さが長くなると、DPマッチングによる2話者間の文の対応付けに失敗し、観測正準相関分析法による教師あり話者正規化の精度が低下してしまう。この問題を解決するには、2人の話者が発話した文中の音素を、確実に対応付けることが必要である。最も簡単な方法としては、文単位での対応付けを行なうのではなく、単語単位で対応付けを行うことが考えらる。しかし、これでは連続音声を用いていないため、話者正規化の精度が悪くなってしまう。したがって、2人の話者が発話した文を対応づけるのに、DPマッチングを用いない方法が必要になる。

(2) 観測正準相関分析法では、2話者間で文の対応付けをおこなうのにDPマッチングを用いている。このため入力話者は、正規化用データとして基準話者 A と同じ文を発話しなくてはならない。これは入力話者 B にとっては負担となる。この問題を解決するためには、DPマッチングを用いない正準相関分析法が望まれる。これにより、入力話者 B は基準話者 A の発話内容にとらわれず、自由に発話をすることができるようになる。このような話者正規化法を、教師なし話者正規化法と呼ぶ。

3.2 教師なし話者正規化

DPマッチングを用いない教師なし話者正規化の一つとして、入力話者 B の音声データを、不特定話者HMMを用いて、Viterbeアルゴリズムにより、音素ごとにセグメンテーションする方法を提案する。なお、基準話者 A の音声データのセグメンテーションについては目視で行う。セグメンテーションを行った後、分割された音素データを用いて、各音素ごとの平均ベクトルをもとめる。基準話者 A と入力話者 B 間の対応付けとして、この音素毎の平均ベクトルを用いて正準相関分析を実行する。これにより、入力話者 B は基準話者 A の発話内容にとらわれず、自由に発話することができ、DPマッチングを用いることなく話者空間の対応付けをおこなうことができる。

以下に教師なし話者正規化の手順を示す。

step(1) 入力話者 B の正規化用音声データを用意する。入力話者 B の音声データの発話内容は基準話者 A と全く異っていてもよい。

step(2) 入力話者 B の正規化用音声データを、不特定話者HMMを用いてViterbeアルゴリズムにより音素毎に分ける。一方、基準話者 A の方は目視により音素毎に分けておく。その後、各話者の音素毎の代表ベクトル（平均ベクトル）を求め、その代表ベクトルを基準話者 A、入力話者 B の対応付けられた特徴ベクトルとして、分散共分散行列 Σ_{12}、Σ_{21}、Σ_{22} を求める。

step(3) 観測空間を基準話者 A の話者空間として固定しておく。STEP(2)で求めた Σ_{12}、Σ_{21}、Σ_{22} を用いて、入力話者 B の話者空間を形成する軸 v_B を、観測正準相関分析の式(1)を用いて求める。その後、入力話者 B の話者空間に、入力話者 B の音声データを射影し、正規化データを作る。

step(4) 基準話者 A の音声データを用いて HMM_A を作り、入力話者 B の正規化データを認識する。

本稿では、観測正準相関分析を用いた教師なし話者正規化の有効性を、教師あり話者正規化法との比較実験により示す。

4 話者正規化による音素認識実験

4.1 実験条件

本稿では、観測正準相関分析を用いた教師なし正規化と教師あり正規化の効果を調べるために、複数話者に対する音素認識実験を行った。表1に実験条件を示す。

表 1: 実験条件
(AA: 音響分析)

	サンリング周波数	12kHz
	高域強調フィルタ	$1-0.97z^{-1}$
A	特徴量	LPC ケプストラム (16 次)
A	フレーム長	20ms
	フレーム周期	5ms
	窓関数	ハミング窓
H	状態数	5 状態 3 ループ
M	分散共分散行列	対角行列
M	タイプ	混合分布型 HMM
	混合数	4

表 2: 話者正規化で用いたデータベース

基準話者	MTK
入力話者	MHO, MMY, MHT, MSH, MYI (男性話者)
	FYM, FTK, FKS, FKN (女性話者)
正規化用データ	ATR 音韻バランス文 B セット a,h,i (150 文) のうちの偶数番目 の 75 文
HMM の学習データ 初期モデル:	ATR 音韻バランス文 B セット a ～ j の 500 文
連結学習 :	MTK の正規化用データと同じ
認識用データ	入力話者の ATR 音韻バランス文 B セット a,h,i のうち奇数番目の 75 文

話者正規化で用いたデータを表2に示す。音声データとしては ATR の音韻バランス文 B セットを使用した。基準話者を MTK として固定し、入力話者として男性 5 名、女性 4 名の計 9 名を用いた。話者正規化用データとして 75 文を使用し、音素 HMM は MTK の 75 文 (a,h,i の偶数番目) を用いて作成した。認識は話者 9 名の 75 文 (a,h,i の奇数番目) に対して行なった。

4.2 従来法との比較実験

教師なし話者正規化法の有効性を調べるために、表3に示す 4 つの実験、*No-norm Sv-norm Usv1-same Usv2-same* を、男性 5 名女性 4 名の計 9 名に対して行った。以下に 4 つの実験内容について述べる。

No-norm は正規化を行なわない実験で、基準話者 A で作った HMM を用いて音素認識したものである。*Sv-norm* は教師あり話者正規化を、*Usv1-same* と *Usv2-same* は教師なし話者正規化を入力話者 B に対して行ない、入力話者 B の正規化された音声データを、基準話者 A で作った HMM を用いて音素認識したものである。また *Usv1-same* は、入力話者 B の音声データを目視により音素毎に分けているのに対して、*Usv2-same* は入力話者 B の音声データを、Viterbe アルゴリズムにより音素毎に分けている。なお、*Usv1-same* と *Usv2-same* の実験では、入力話者 B の正規化用データは基準話者と同じものを使用している。

実験結果を表 (4) に示す。

表 4: 平均音素認識率 (%)
(正規化用データとして同じデータを使用)

	男性平均	女性平均	全体平均
No-norm	50.2	36.9	44.3
Sv-norm	59.0	60.0	59.6
Usv1-same	**57.7**	**60.0**	**58.7**
Usv2-same	**56.2**	**58.1**	**57.0**

表 4より、*Usv2-same* は正規化をかけない場合に比べて、全体平均で 12.7% 認識率が向上し、また教師あり話者正規化法 (*Sv-norm*) と比べて、同程度の認識率を得ることができた。これは、普通、音素間のわたりと言ったような部分は音響的に不安定であるため、正準相関分析において、全ての特徴ベクトルを対応付ける必要はないためと考えられる。教師なし話者正規法においては、対応付けとして安定した音素のみの対応付けを行なっているため、表 4に示すように音素認識率の低下が少なかったものと考えられる。また、男性話者で作った HMM を用いて女性話者の音声データを認識すると、正規化を行なわない場合 (*No-norm*) は 36.9% と低い認識率であるが、教師なし話者正規化 (*Usv2-same*) では 58.1% となり、21.2% もの認識率の向上が得られた。

4.3 異なる発話内容についての検討

4.2では、従来法である教師あり話者正規化と比較するために、入力話者の正規化用データを基準話者と同じものに設定して音素認識実験を行なった。ここでは、入力話者の発話内容が基準話者と全く異った場合と、同じ場合との比較実験を行なった。つまり、表3の *Usv1-same*、*Usv2-same*、*Usv1-diff*、*Usv2-diff* の 4 つの実験を行なった。

ここで、*Usv1-same* と *Usv2-same* は、入力話者の

表 3: 実験内容

	正規化なし	正規化あり				
		教師あり	教師なし			
			ラベル付け		発話内容	
			Manual	Viterbi	Same	Different
No-norm	○					
Sv-norm		○				
Usv1-same			○		○	
Usv2-same				○	○	
Usv1-diff			○			○
Usv2-diff				○		○

正規化用データの発話内容を、基準話者と同じにした場合の実験であり、*Usv1-diff*、*Usv2-diff*は、入力話者の正規化用データの発話内容を、基準話者と異るものにした場合の実験である。また、*Usv1-same*と*Usv1-diff*は入力話者の音素毎のセグメンテーションを目視で行なった場合であり、*Usv2-same*と*Usv2-diff*は入力話者の音素毎のセグメンテーションを、Viterbiアルゴリズムで行なった場合である。実験結果を表5に示す。

表 5: 平均音素認識率 (%)
(正規化用データとして異なった発話内容を使用)

	男性平均	女性平均	全体平均
Usv1-same	57.7	60.0	58.7
Usv2-same	56.2	58.1	57.0
Usv1-diff	**54.7**	**57.5**	**56.0**
Usv2-diff	**54.9**	**57.2**	**55.9**

表5より、入力話者の正規化用データを基準話者と異るものにしても、認識率の低下は少ないと言える。また、*Usv2-diff*は正規化を行なわない場合(表4の*No-norm*)に比べて、11.6% 認識率が向上した。

4.4 話者正規化の文章発話数の検討

話者正規化では、充分な正規化を行なうのに、どのくらいの文章数が必要であるかが問題となる。なぜなら、特定話者HMMで作成するのに必要な文章数と同等の文章数が、話者正規化で必要となるのであれば、話者正規化の意味がなくなるからである。

そこで、教師なし話者正規化(*Usv2-diff*)において、正規化に必要な文章数を75文から10文ずつ減らして音素認識実験を行なった。また、教師あり話者正規化との比較も行なった。実験条件は表1に示した通りである。基準話者はMTKを使用し、認識は男性として

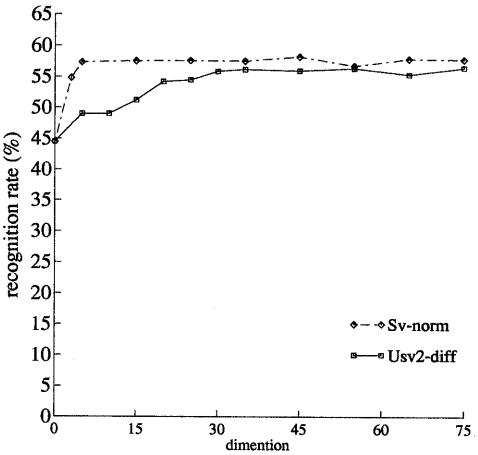

図 2: 正規化における文章発話数と認識率の関係(入力話者は男性)

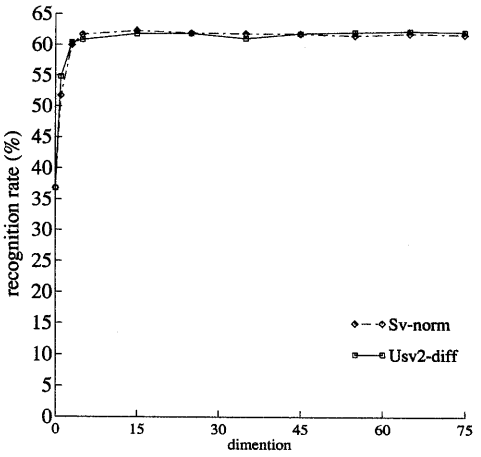

図 3: 正規化における文章発話数と認識率の関係(入力話者は女性)

MYIを、女性としてFTKを使用した。結果を図2、3に示す。図2、3の横軸は正規化を行なうために用いた文章数であり、縦軸は音素認識率を表している。

図2、3より、男性話者では25文程度で、女性話者では5文程度で正規化が収束することがわかる。また女性話者の場合、教師あり話者正規化とほぼ同程度の結果を得ることができた。

5 まとめ

入力話者は基準話者と同じ文を発話しなければならないという、教師あり話者正規化法の問題を解決する教師なし話者正規化の方法を示した。まず不特定話者HMMを用いて、Viterbeアルゴリズムにより、音素毎にセグメンテーションを行なう。次に音素の代表ベクトルを求め、正準相関分析法により話者空間の対応付けを行なう。また、従来法である教師あり話者正規化法との比較実験により、音素認識率がさほど低下しないことを確認した。

今後の課題としては各音素データの代表ベクトルを増やしたり、他の特徴量の検討、話者適応との比較を行なう予定である。

参考文献

[1] 田頭茂明、有木康雄： 部分空間射影による話者正規化を用いた不特定話者HMM, 信学技法,SP95-98,(1995-12).

[2] 田頭, 西島, 有木: 話者部分空間への写像によ話者認識と話者正規化, 信学技報,SP95-28,(1995).

[3] 羅, 尾関: アフィン変換を用いた音声特徴量の正規化, 信学技報, SP96-10,(1996).

[4] 石井, 外村: 重回帰写像モデルを用いた話者正規化と話者適応化方式, 信学技報, SP96-91,(1996).

[5] 柳井晴夫、高根芳雄： “多変量解析法”、朝倉書店、pp.99-113, (1977).

[6] K.Choukri G.Chollet Y.Grenier： ”Spectral transformations through Canonical Correlation Analysis for speaker adaptation in ASR”, ICASSP86, pp.2659-2662, (1986).

[7] 古井: スペクトル空間の階層的クラスタ化による話者適応化, 信学技報, SP88-21,(1988).

[8] 中村: ファジィクラスタリングによる教師なし話者適応化, 日本音響学会春季講演論文集,1-5-20,pp.43-44(1991).