音声言語情報処理 34－16
(2000. 12. 21)

音韻空間への射影に基づく話者正規化による音素認識

西田 昌史　　有木 康雄

龍谷大学 理工学部

〒520-2194 滋賀県大津市瀬田大江町横谷1-5
Tel: 077-543-7427
E-mail: nishida@arikilab.elec.ryukoku.ac.jp, ariki@rins.ryukoku.ac.jp

あらまし 音声認識の実用化において，不特定話者の音声を高精度に認識することが必要となってくる．しかし，話者の違いによる変動と音素コンテキストの変動により，認識精度が劣化してしまう．これに対して，音声データから音韻性と話者性を分離し，音韻性のみを抽出することができれば，頑健な話者正規化による音声認識が可能となる．また，話者性を抽出することで，話者性を考慮した頑健な話者適応による音声認識も可能になると考えられる．これまでに話者内分散の大きい空間を音韻空間，話者内分散の小さい空間を話者空間とみなし，話者空間へ射影することにより話者照合を行う方法を提案してきた．本研究では，音声データから話者空間への射影成分を取り除き，音韻空間への射影成分のみを用いて話者正規化を行う方法について検討を行った．

キーワード　不特定話者音声認識，話者正規化，話者認識，部分空間法，音韻空間，話者空間

Phoneme Recognition by Speaker Normalization based on Projection to Phonetic Space

Masafumi Nishida　　Yasuo Ariki

Faculty of Science and Technology, Ryukoku University

1-5 Yokotani, Oe-cho, Seta, Otsu-shi, Shiga, 520-2194 Japan
Tel: +81-77-543-7427
E-mail: nishida@arikilab.elec.ryukoku.ac.jp, ariki@rins.ryukoku.ac.jp

Abstract　In practical speech recognition, it is required to recognize precisely the speech data spoken by many speakers. But it is a difficult problem, because of difference among speakers and phonetic contexts. To this problem, if phonetic information and speaker information included in speech data can be separated and only phonetic information is extracted, robust speaker normalization will be realized. If speaker information included in speech data can be extracted, robust speaker adaptation will be also realized. We have already proposed a speaker verification method by projection to speaker space, under the assumption that a space with large variation of within-speaker is a "phonetic space" and a space with small variation of within-speaker variance is a "speaker space". In this study, we studied a speaker normalization method by projection to phonetic space based on this insight.

key words　speaker independent speech recognition, speaker normalization, speaker recognition, subspace method, phonetic space, speaker space

1 はじめに

近年，音声認識においては，隠れマルコフモデル (Hidden Markov Model: HMM) を用いた不特定話者音声認識システムに関する研究が盛んに行われている．不特定話者音声認識システムは，多数話者の発声データを用いて学習されるため，話者の発声の多様性に対して頑健であり，特定話者音声認識システムに比べて，使用者が事前に発声する必要がないという利点がある．しかしながら，このような不特定話者音声認識システムは，使用者の音声を事前に学習した特定話者音声認識システムに比べて，一般に認識精度が低い．また，一部の話者に対し極端に認識精度が低くなる現象が見られる．これらの問題を解決するために，話者適応化，話者正規化といった手法が適用されてきた．

話者適応は，特定話者が発声した少量の適応データを用い，不特定話者の音響モデルを特定話者へ近づける方法である．この話者適応の代表的な手法として，以下の二つが挙げられる．一つは，多数話者のデータで学習された不特定話者モデルを事前知識として利用し，適応データ量に応じた効率的なパラメータ推定法として，最大事後確率推定法（Maximum A Posteriori probability estimation : MAP 推定法）[1]-[4]が用いられている．もう一つの方法は，少量の適応データしか得られない場合に，適応データから学習できない多数のパラメータが存在するので，これらの未学習パラメータを補間する方法として，MLLR(Maximum Likelihood Linear Regression)[5]-[7]，移動ベクトル場平滑化 (Vector Field Smoothing: VFS)[8]-[10]が用いられている．

話者正規化は，話者性から生じる発声のゆらぎを取り除く手法である．代表的な手法として，ケプストラム平均正規化（Cepstrum Mean Normalization: CMN），声道長正規化（Vocal Tract Length Normalization : VTLN）[12]-[14]が挙げられる．CMNは，入力データからケプストラムの長時間平均を差し引く手法であり，話者性のみならず，マイクロホンや電話回線の伝送特性等のゆらぎを取り除く方法である．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．

従来の不特定話者モデルは，多数話者の音声データを用いて学習されていることから，話者の違いによる変動と音素コンテキストの変動が混在し，広がりの大きいモデルとなっており，認識精度劣化の要因の一つになっていると考えられる．これに対して，音声データから音韻性と話者性を分離し，音韻性のみを抽出することができれば，頑健な話者正規化による音声認識が可能となる．また，話者性を抽出することで，話者性を考慮した頑健な話者適応による音声認識も可能になると考えられる．

これまでに，我々は，音声データを特異値分解により音韻性と話者性に分離し，話者正規化する方法を提案した[15]．また，最近，話者内分散の大きい空間を音韻空間，話者内分散の小さい空間を話者空間とみなし，音声データを話者空間へ射影することにより話者照合を行う方法を提案し，その有効性を示した[16]．これを踏まえて，本研究では，音声データから話者空間への射影成分を取り除き，音韻空間への射影成分のみを用いて話者正規化を行う方法について検討を行う．

2節では，話者空間への射影による話者認識法および話者照合実験と結果，3節では，音韻空間への射影による話者正規化法，4節では，話者正規化による音素認識実験と結果について述べる．

2 話者空間への射影による話者認識

2.1 部分空間分離

音声中に含まれている音韻性と話者性を分離し，話者性のみを抽出することができれば，頑健な話者認識が可能になると考えられる．発話内容による音声の特徴変動は，音声中に含まれている音韻のばらつき，つまり音韻性により生じると考えられる．また，話者性は，発話内容に依存しないと考えられる．そこで，話者内分散の大きい空間を音韻空間，話者内分散の小さい空間を話者空間とみなし，主成分分析に基づく部分空間法により音韻性と話者性を分離することを目的として，話者空間への射影による話者照合を行う[16]．

観測空間で観測される話者cの学習音声データをn次元特徴ベクトルの集合$\{x_t^{(c)}\}$ $(t=1,2,\cdots,N)$とする．この学習データから，平均ベクトル$\mu^{(c)}$および分散共分散行列$R^{(c)}$を次式で求める．

$$\mu^{(c)} = \frac{1}{N}\sum_{t=1}^{N} x_t^{(c)} \tag{1}$$

$$R^{(c)} = \frac{1}{N}\sum_{t=1}^{N}(x_t^{(c)}-\mu^{(c)})(x_t^{(c)}-\mu^{(c)})^T \tag{2}$$

この分散共分散行列$R^{(c)}$を固有値分解すると，次式のようになる．

$$R^{(c)} = \Phi^{(c)}\Sigma^{(c)}\Phi^{(c)T} \tag{3}$$

ここで，$\Sigma^{(c)}$は，分散共分散行列$R^{(c)}$の固有値$\lambda_i^{(c)}(i=1,\cdots,k,\cdots,n)$を対角成分にもつ対角行列である．また，$\Phi^{(c)}$は，分散共分散行列$R^{(c)}$の正規直交基底ベクトル$\varphi_i^{(c)}(i=1,\cdots,k,\cdots,n)$を列ベクトルとする行列である．

固有値分解で得られた固有値$\lambda_i^{(c)}$は，対応する正規直交基底ベクトル$\varphi_i^{(c)}$上のデータの分散を表しており，ここでは，固有値が大きい上位k個の正規直交基底ベクトルにより張られた空間を音韻空間，音韻空間の補空間で

ある$n-k$個の正規直交基底ベクトルにより張られた空間を話者空間と呼ぶ．したがって，分散の大きい軸で構成された音韻空間は音韻性が強い空間であり，分散の大きい軸を除くことによって得られた，音韻空間の補空間である話者空間が，音韻性を抑えた話者性を表す空間であるという仮説が成り立つ．

そこで，音韻性を抑えた話者空間に学習データを射影して，各話者の話者空間でGMMにより話者モデルを学習し，この話者モデルを用いて話者照合を行う方法が考えられる．

2.2 話者空間への射影による話者照合

話者cの学習音声データのn次元特徴ベクトル集合を$\{x_t^{(c)}\}$，話者cの学習データを主成分分析して得られた正規直交基底ベクトルを$\varphi_i{}^{(c)}(i=1,\cdots,k,\cdots,n)$とすると，$n-k$個の$\varphi_i^{(c)}$により張られる部分空間が話者空間となる．この話者空間に学習データ$\{x_t^{(c)}\}$を次式のように射影する．

$$\begin{aligned}\hat{x}_t^{(c)} &= \sum_{i=k+1}^{n}(x_t^{(c)}-\mu^{(c)},\varphi_i{}^{(c)})\varphi_i{}^{(c)}+\mu^{(c)} \\ &= P^{(c)}(x_t^{(c)}-\mu^{(c)})+\mu^{(c)} \qquad (4)\end{aligned}$$

ここで，$P^{(c)}$は，話者cの話者空間への射影行列を表す．式(4)によって学習データを話者空間へ射影し，話者空間でGMMにより話者モデルを学習する．

話者空間への射影の概念図を図1に示す．図1中のP_A，P_Bはそれぞれ話者A，話者Bの話者空間を表している．楕円は，学習データの分散を表している．図1に示すように，話者空間は分散の小さい軸で構成された空間であるので，各話者A，Bの学習データを各話者空間に射影すると，観測空間に比べて話者間分散は固定のまま，話者内分散が小さくなる．

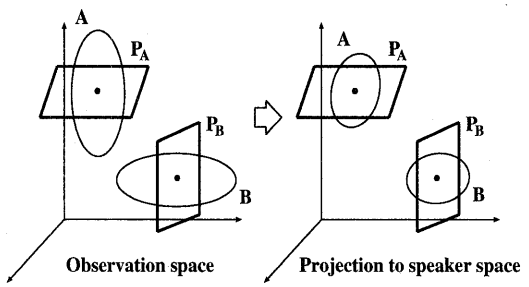

図 1: 話者空間への射影

照合は，入力特徴ベクトルの集合$\{x\}$を式(4)により申告話者cの話者空間に射影し，GMMにより対数尤度$\log P(\hat{x}|\lambda^{(c)})$を求める．この対数尤度を式(5)に示す尤度比尺度に基づき正規化する．この正規化した対数尤度が閾値より大きければ，本人の音声であると判定する．

$$\log \bar{P}(\hat{x}|\lambda^{(c)}) = \log P(\hat{x}|\lambda^{(c)}) - \max_{i\neq c}\log P(\hat{x}|\lambda^{(i)}) \qquad (5)$$

2.3 話者照合実験

2.3.1 実験条件

使用したデータは，男性20名・女性10名が約10ヵ月に渡る2時期（時期1，時期2）に発声した文章データである．分析条件を表1に示す．学習には時期1に発声した5文章を用い，評価には，時期1から約10ヵ月後の時期2に発声した各15文章を1文章づつ用いた．登録話者は男性10名・女性5名で，詐称者は登録話者と異なる男性10名・女性5名である．

表 1: 分析条件

サンプリング周波数	12kHz
高域強調	$1-0.97z^{-1}$
フレーム長	20ms
フレーム周期	5ms
窓タイプ	ハミング窓
特徴パラメータ	24次メルフィルタバンク 18次ケプストラム

2.3.2 実験結果と考察

話者空間においてGMMにより話者モデルを学習した場合について，話者照合実験を行った．実験結果を表2に示す．表2の*base*は観測空間におけるGMM，*LDA*は，判別分析により得られる分離能力の高い部分空間においてGMMで学習した場合で，最も誤り率が小さかった14次元の時の結果を示している．*speaker space*は，話者空間においてGMMで学習した場合で，最も誤り率が小さかった4-18次元の時の結果を示している．

表 2: 話者照合結果(%)

mixture number	base	LDA (14dim)	speaker space (4-18dim)
2	3.20	3.11	1.63
4	2.19	2.65	1.29
16	2.18	1.79	1.79
64	1.50	1.29	1.50

表2の結果から，判別分析による方法は，観測空間に比べると，3~18%の照合精度の改善が得られ，判別分析により構成された話者分離能力が大きい部分空間の有効性が確認できた．話者空間においては，混合分布数が4

の時に等誤り率1.29% が得られ，判別分析の方法において混合分布数が4の場合に比べると，51%の照合精度の改善が得られた．さらに，話者空間において混合分布数が4の場合は，判別分析の方法において混合分布数が64の場合と同等の照合精度が得られた．したがって，話者内分散の大きい空間を音韻空間，話者内分散の小さい空間を話者空間とみなすことで，話者空間は，音韻性を抑えた話者性を表す空間であるということがわかった．また，この結果から，話者空間への射影による話者照合法は，20秒という少量の学習データに対して，混合分布数を64から1/16に削減できる．このため，少量のパラメータ数で話者モデルを表すことができ，高速に学習および照合を行うことができる．

3 音韻空間への射影による話者正規化

話者正規化の代表的な手法として，ケプストラム平均正規化（CMN）が挙げられる．CMNは，式(6)に示すように，入力特徴ベクトルx_tの各次数毎に，短文章の音声区間全体での平均ベクトルを計算し，この平均ベクトルを各時点の入力特徴ベクトルから差し引く手法である．CMNにより，話者性のみならず，マイクロホンや電話回線の伝送特性等を正規化することができる．

$$\hat{x}_t = x_t - \frac{1}{N}\sum_{t=1}^{N} x_t \tag{6}$$

ここで，Nは入力音声の総フレーム数であり，x_tは入力音声の時刻tにおける特徴ベクトルである．

2節で述べたように，話者内分散の大きい空間は音韻性が強いため，この空間を除くことで音韻性の影響を抑えられ，話者認識精度を向上させることができる．これを踏まえて，ここでは，不特定話者の音響モデルを学習する際に用いた各話者に対して部分空間を構成し，音韻性のより強い軸を選択し，その軸で構成される音韻空間へ射影することで話者正規化を行う方法について検討する．また，本研究では，学習データおよび評価データに対してCMNを行い，CMNを行ったデータに対して音韻空間への射影による話者正規化を行う．

まず，学習における話者正規化法について述べる．観測空間で観測される話者cの学習音声データをn次元特徴ベクトルの集合$\{x_t^{(c)}\}$ $(t = 1, 2, \cdots, N)$とする．この学習データから，相関行列$S^{(c)}$を次式で求める．

$$S^{(c)} = \frac{1}{N}\sum_{t=1}^{N} x_t^{(c)} x_t^{(c)T} \tag{7}$$

この相関行列$S^{(c)}$を固有値分解すると，次式のようになる．

$$S^{(c)} = \Phi^{(c)}\Sigma^{(c)}\Phi^{(c)T} \tag{8}$$

ここで，$\Sigma^{(c)}$は，相関行列$S^{(c)}$の固有値$\lambda_i^{(c)}(i = 1, \cdots, k, \cdots, n)$を対角成分にもつ対角行列である．また，$\Phi^{(c)}$は，相関行列$S^{(c)}$の正規直交基底ベクトル$\varphi_i^{(c)}(i = 1, \cdots, k, \cdots, n)$を列ベクトルとする行列である．

ここでは，固有値分解で得られた正規直交基底ベクトル$\varphi_i^{(c)}$から音韻性の強い軸を選択し，k個の正規直交基底ベクトルにより張られた空間を音韻空間と呼ぶ．この音韻空間を話者毎に構成する．そして，式(9)により，各話者の学習データをフレーム毎に各話者の音韻空間へ射影し，射影量が最大となる話者の音韻空間を選択する．この音韻空間への射影ベクトルを用いて不特定話者の音響モデルを学習する．したがって，学習データ毎に音韻性のより近い話者の音韻空間へ射影することで，話者正規化を行っていることになる．

$$\hat{x}_t = \sum_{i=1}^{k}(x_t, {\varphi_i}^{(s)})\varphi_i^{(s)} \tag{9}$$

$$s = arg\max_j ||\sum_{i=1}^{k}(x_t, {\varphi_i}^{(j)})\varphi_i^{(j)}||^2$$

音韻空間への射影の概念図を図2に示す．

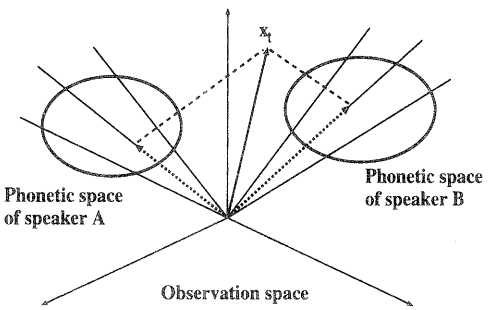

図 2: 音韻空間への射影

次に，認識における話者正規化法について述べる．入力音声データを学習時と同様に，式(9)により，フレーム毎に各話者の音韻空間へ射影し，射影量が最大となる話者の音韻空間を選択する．この音韻空間への射影ベクトルにより，音声認識を行う．したがって，入力音声データのフレーム毎に音韻性のより近い話者の音韻空間へ射影することで，話者正規化を行っていることになる．

4 話者正規化による音素認識実験

4.1 実験条件

音韻空間への射影による話者正規化の有効性を示すため，音素認識実験を行った．

音響モデルとしては，monophone HMMを用いた．音響モデルの学習には，まず，ATR連続音声データベースa～jセットから6名が発声した1名300文章のデータとその視察ラベルを用いて，初期モデルを作成した．次に，日本音響学会新聞記事読み上げコーパス(JNAS)のうち，男性話者20 名が発声した1名150文章のデータを用いて，連結学習を行なった．音響分析条件とHMMのトポロジーを表3に示す．

表 3: 音響分析とHMM

音響分析	サンプリング周波数	16kHz
	特徴パラメータ	MFCC(1-24次)
	フレーム長	20ms
	フレーム周期	10ms
	窓タイプ	ハミング窓
HMM	状態数	5状態3ループ
	タイプ	monophone HMM
	混合数	2
	学習方法	連結学習

評価用データには，日本音響学会新聞記事読み上げコーパス(JNAS)のうち，学習話者と異なる男性話者10名が発声した1名10文章を用いて，音素認識実験を行った．

4.2 実験結果と考察

音韻空間への射影による話者正規化，ならびにCMNによる話者正規化における音素認識実験を行った．実験結果を表4に示す．実験は，音素正解率および音素正解精度で評価を行った．表4のSIは，話者の正規化を行っていない場合，CMNは，CMNにより話者の正規化を行った場合，*phonetic space* $+CMN$ は，音韻空間への射影による話者正規化の結果を示している．

表 4: 話者正規化による音素認識結果

	正解率(%)	正解精度(%)
SI	60.82	52.73
CMN	63.06	54.52
Phonetic space +CMN	63.33	54.44

表4の結果から，CMNによる話者正規化の有効性が確認できた．音韻空間への射影による話者正規化では，各話者の部分空間の軸のうち，3本を1セットとして軸を取り除いて，音素認識実験を行った結果，14～16次の軸を取り除いた場合，音素正解率63.33%，音素正解精度54.44%が得られた．その結果，CMNによる話者正規化に比べると，音素正解率が若干ではあるが向上した．また，部分空間の分散の小さい軸，つまり高次の空間(21～24次)を取り除いた場合，音素正解率62.55%，音素正解精度54.09%が得られ，高次の空間にも音韻性はある程度存在しており，高次と低次の間の中間の次元には，あまり音韻性が存在していないと考えられる．

今後は，各話者の部分空間の軸に重み付けすることで，音韻性を強調した空間を構成した話者正規化を行う予定である．さらに，音韻性と話者性の分離に基づく話者正規化法ならびに話者適応法について，検討を行っていく予定である．

5 おわりに

従来の不特定話者モデルは，多数話者の音声データを用いて学習されていることから，話者の違いによる変動と音素コンテキストの変動が混在し，広がりの大きいモデルとなっており，認識精度劣化の要因の一つになっていると考えられる．これに対して，音声データから音韻性と話者性を分離し，音韻性のみを抽出することができれば，頑健な話者正規化による音声認識が可能となる．また，話者性を抽出することで，話者性を考慮した頑健な話者適応による音声認識も可能になると考えられる．

これまでに，我々は，話者内分散の大きい空間を音韻空間，話者内分散の小さい空間を話者空間とみなし，音声データを話者空間へ射影することにより話者照合を行う方法を提案し，その有効性を示した．これを踏まえて，本研究では，音声データから話者空間への射影成分を取り除き，音韻空間への射影成分のみを用いて話者正規化を行う方法について検討を行った．

今後は，各話者の部分空間の軸に重み付けすることで，音韻性を強調した空間を構成した話者正規化を行う予定である．さらに，音韻性と話者性の分離に基づく話者正規化法ならびに話者適応法について，検討を行っていく予定である．

参考文献

[1] C.-H.Lee, C.-H.Lin and B.-H.Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models", IEEE Trans. on Signal Processing, vol.39, no.4, pp.806-814, 1991.

[2] J.-L.Gauvain and C.-H.Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", IEEE Trans. on Speech and Audio Processing, vol.2, no.2, pp.291-298, 1994.

[3] E.R.Buhrke and C.Liu, "A Generalization of the Maximum a Posteriori Training Algorithm for Mixture Priors", Proc. ICASSP, vol.2, pp.993-996, 2000.

[4] 中川 聖一, 越川 忠, "最大事後確率推定法を用いた連続出力分布型HMMの適応化", 音響誌, vol.49, no.10, pp.721-728, 1993.

[5] C.J.Leggetter and P.C.Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression", Proc. ARPA Spoken Language Technology Workshop, pp.104-109, 1995.

[6] O.Siohan, C.Chesta and C.-H.Lee, "Joint Maximum a Posteriori Estimation of Transformation and Hidden Markov Model Parameters", Proc. ICASSP, vol.2, pp.965-968, 2000.

[7] 石井 純, 外村 政啓, "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, pp.29-35, 1997.

[8] L.He, J.Wu, D.Fang and W.Wu, "Speaker Adaptation based on Combination of MAP Estimation and Weighted Neighbor Regression", Proc. ICASSP, vol.2, pp.981-984, 2000.

[9] 大倉 計美, 杉山 雅英, 嵯峨山 茂樹, "混合連続分布HMM移動ベクトル場平滑化話者適応方式", 信学論(D-II), vol.J76-D-II, no.12, pp.2469-2476, 1993.

[10] 大倉 計美, 飯田 正幸, "音素識別誤りに基づく平滑化制御を用いた移動ベクトル場平滑化話者適応", 信学技報, SP96-23, pp.23-29, 1996.

[11] 外村 政啓, 小坂 哲夫, 松永 昭一, "最大事後確率推定法と適応データ量に応じた平滑化手法を用いた話者適応", 信学論(D-II), vol.J81-D-II, no.3, pp.465-471, 1998.

[12] E.Eide and H.Gish, "A Parametric Approach to Vocal Tract Length Normalization", Proc. ICASSP, vol.1, pp.346-348, 1996.

[13] L.Welling, S.Kanthak and H.Ney, "Improved Methods for Vocal Tract Normalization", Proc. ICASSP, pp.761-764, 1999.

[14] 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", 信学技報, SP99-101, pp.49-54, 1999.

[15] 田頭 茂明, 西島 政幸, 有木 康雄, "話者部分空間への写像による話者認識と話者正規化", 信学技報, SP95-28, pp.25-32, 1995.

[16] 西田 昌史, 有木 康雄, "話者補空間射影による話者認識", 信学技報, SP2000-12, pp.17-22, 2000.