

ケプストラムの回転性とそれに基づく話者認識に関する実験的検討*

◎齋藤大輔[†], 鎌田敏明^{†‡}, 峯松信明[†], 広瀬啓吉[†], 長内隆[‡], 蒔苗久則[‡] ([†]東大, [‡]科警研)

1 はじめに

音声は音韻のみならず, 年齢や性別, 声道長や音響機器などの諸要因により不可避的に変化し, これらの情報を伝達することになる. 音声の持つ情報は大きく分けて, 言語的情報, 非言語的情報, パラ言語的情報に分類する事ができる. 通常, 音声を用いたアプリケーションではこれらのうち一つに着眼して情報を抽出する. 例えば音声認識では言語的情報を抽出し, 話者認識や話者同定では, 非言語的情報に着目する. このような着目する情報を取り出す際, その他の情報が多様に变化した学習データを用いて, 着目する情報以外を平均化することでモデルを構築する事が広く行われている. 話者認識においては, 学習データを用いて混合正規分布 (Gaussian Mixture Model; GMM) によるモデル化 [1] やベクトル量子化 (Vector Quantization; VQ) [2] などによって非言語的情報を抽出する.

平均化を通して目的以外の情報を抑制するという事は, 特徴量空間において各種の情報の違いがどのように表出するかは十分に議論されていないことを意味する. そのため統計モデルを構築する為に, 多様な特徴をカバーするように大量の学習データを必要とする事が多い. 一方で学習データに含まれないような特徴を有する特異な音声に対しては, これらのモデルが有効に機能しないことがある [3]. これらはミスマッチ問題と呼ばれ, 統計的手法における問題の一つと考えられる. 特徴量空間の特性を考慮する事でこれらの問題により柔軟に対応しうる可能性がある.

筆者らはこれまでに, 声道長正規化 [4] において広く用いられているケプストラム空間の変換行列に着目し, その幾何学的性質を理論的および実験的に検討してきた [5, 6]. 声道長変化を表す変換行列は強い回転性を有し, ケプストラムの方向成分は声道長が変化する事でおおよそ同様に回転する. さらにこの回転性を詳細に分析し, 2次元部分空間における回転の組み合わせとして記述できることを示している [7, 8].

本稿ではこれまで分析再合成音を通して行ってきた回転性の分析を実音声に対して適用する. 先行研究において示してきたベクトルの内積に基づく回転角, および 2次元部分空間における回転角を分析する事で, 話者情報がケプストラム空間においてどのよう

な幾何学的特性として現れるかを検討する. さらに身長異なる話者の音声を分析する事でケプストラムの回転性に基づく話者情報表現の初期検討を行う.

2 声道長変化の定式化

話者の声道長の変化は, 音声のスペクトル表現における周波数ウォーピングとして考えることができる. 今, 周波数ウォーピングにおける変換前後の正規化角周波数を $\omega, \hat{\omega}$ ($0 \leq \omega, \hat{\omega} \leq \pi$) とする. このとき $z = e^{j\omega}$, $\hat{z} = e^{j\hat{\omega}}$ として, 周波数ウォーピングとして以下の 1 次全域通過関数を考える.

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \quad (1)$$

このとき α は $|\alpha| < 1$ の実数であり, $\alpha < 0$ の場合, 周波数軸が低域に変換され声道長は長くなる. 一方 $\alpha > 0$ の場合, 周波数軸は高域に変換され声道長が短くなる. 以後 α をウォーピングパラメータと呼ぶ.

江森らは上記の周波数ウォーピングをもとに, そのケプストラム空間における記述を導出している [9]. パワーを表現するケプストラムの 0 次項 ($c_0, \hat{c}_0$) を考慮しない場合, 周波数ウォーピングは以下の式で表現される.

$$\hat{c} = A c \quad (2)$$

$$A = \begin{pmatrix} 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots & \cdots \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad (3)$$

さらに Pitz らによれば, 式 (3) における行列 A の要素 a_{ij} はウォーピングパラメータ α を用いて以下のよう表せる [10].

$$a_{ij} = \frac{1}{(j-1)!} \sum_{m=\max(0, j-i)}^j \binom{j}{m} \times \frac{(m+i-1)!}{(m+i-j)!} (-1)^{(m+i-j)} \alpha^{(2m+i-j)} \quad (4)$$

ただし

$$\binom{j}{m} = \begin{cases} jC_m & (j \geq m) \\ 0 & (j < m) \end{cases} \quad (5)$$

とする.

* Experimental study of speaker recognition based on the rotational properties of cepstrum coefficients.
by SAITO Daisuke, KAMADA Toshiaki (Univ. of Tokyo / NRIPS), MINEMATSU Nobuaki, HIROSE Keikichi (Univ. of Tokyo), OSANAI Takashi, and MAKINAE Hisanori (NRIPS)

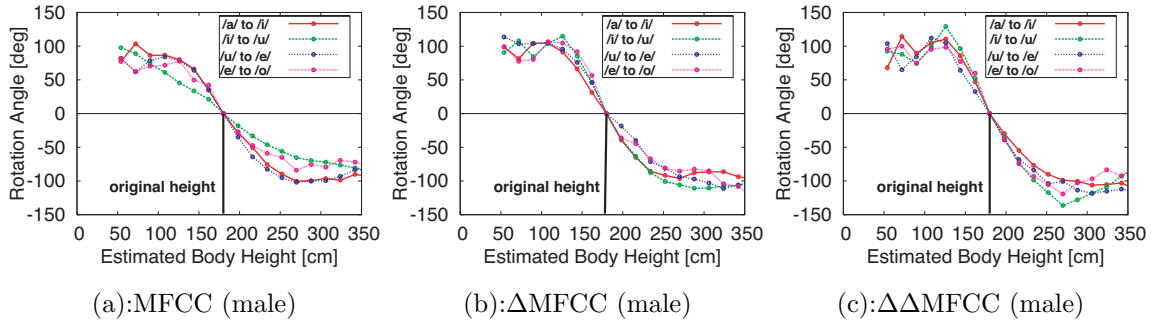


Fig. 1 男性話者（身長 180 cm）における分析再合成音の推定身長と回転角との対応関係

3 ケプストラム空間における回転性

筆者らは多次元空間での幾何学を通して、式 (3) の性質を考察してきた [5, 6]. n 次元空間において行列 R が以下の性質を満たすとき、 R を回転行列と呼ぶ.

$$R^t R = R R^t = I \quad (6)$$

$$\det R = +1 \quad (7)$$

先行研究では式 (3) が近似的に式 (6)(7) を満たすことを理論的に示している. 加えて日本語 5 母音発声 /aiueo/ について STRAIGHT[11] を用いて人工的に声道長を変化させた再合成音を用いて、その音声の MFCC パラメータの回転性を実験的に示している. Fig. 1 は二つのベクトルの角度を式 (8) で定義し、母音遷移の中心部分における声道長変換前後の角度と声道長（身長）との対応関係を示したものである.

$$\theta = \arccos \frac{a \cdot b}{|a||b|} \quad (8)$$

Fig. 1 より音声のパラメータベクトルの方向が声道長変化に依存して回転している事が分かる.

4 固有値による回転性の記述

4.1 回転行列の対角化

行列 A の回転性を詳細に記述するため、行列の固有値に基づく解析を考える. 以下 n 次元の回転行列 R_n を用いて議論を進める. R_n は n 次のユニタリ行列 U_n を用いて複素数を要素に持つ対角行列 D_n の形に対角化できる.

$$R_n = U_n D_n U_n^\dagger \quad (9)$$

一方 2 次元の回転行列 R_2 は一つの回転パラメータ θ を用いて以下のように表される.

$$R_2(\theta) = U_2 D_2(\theta) U_2^\dagger \quad (10)$$

ここで

$$R_2(\theta) = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \quad (11)$$

$$D_2(\theta) = \begin{pmatrix} e^{j\theta} & 0 \\ 0 & e^{-j\theta} \end{pmatrix} \quad (12)$$

R_n の固有方程式は $n=2m$ ($m \in \mathcal{N}$) の時、絶対値 1 の m 組の複素共役解を持ち、 $n=2m+1$ の時、1 と m 組の複素共役解を持つ. 故に式 (9) の D_n は m 次の回転パラメータベクトル $\Theta = (\theta_i | i=1, 2, \dots, m)$ を用いて以下のように表現できる.

$$D_n(\Theta) = \begin{cases} \begin{pmatrix} D_2(\theta_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & D_2(\theta_m) \end{pmatrix} & (n : \text{even}) \\ \begin{pmatrix} 1 & \cdots & 0 \\ & D_2(\theta_1) & \\ \vdots & & \ddots \\ 0 & \cdots & D_2(\theta_m) \end{pmatrix} & (n : \text{odd}) \end{cases} \quad (13)$$

さらに式 (10) を用いることで R_n は実数の回転行列のみを用いて以下のように表現することができる [12].

$$R_n(\Theta) = R_n'' R_n'(\Theta) R_n''^t \quad (14)$$

ここで、 $n=2m+1$ のとき、 $R_n'(\Theta)$ は以下のようになり、 $n=2m$ では式 (15) の 1 行目、1 列目を削った形となる.

$$R_n'(\Theta) = \begin{pmatrix} 1 & \cdots & 0 \\ & R_2(\theta_1) & \\ \vdots & & \ddots \\ 0 & \cdots & R_2(\theta_m) \end{pmatrix} \quad (15)$$

このとき R_n'' は式 (9) におけるユニタリ行列と対応する実回転行列であり、回転行列 R_n の固有ベクトルと

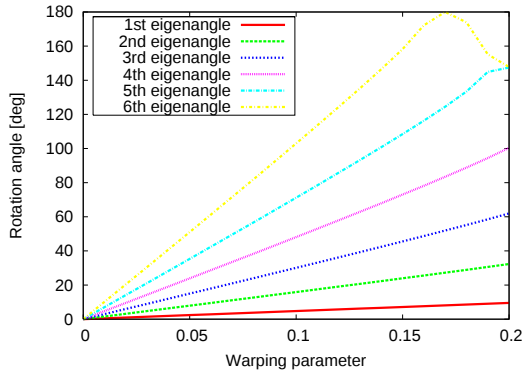


Fig. 2 固有回転角と α の対応関係

等価の情報を持つ。これは回転行列 $R_n(\Theta)$ が持つある固有姿勢へ補正する回転行列と考える事ができる。 R_n^{tt} はその逆操作である。

4.2 部分空間上での回転角

式 (8) による回転角の定義は、多次元空間においては曖昧性を有しているといえる。すなわち、特徴量空間上で全く異なる変化が内積に基づく角度では全く同じ値に観測されることがある。また逆に、同一の回転行列で表される変化も、Fig. 1 のように内積角で観測した場合、特徴量空間における座標の違い（ケプストラム空間では音韻、性別の違い）によって異なる角度で表現される。

今 $n = 2m$ において、ベクトル $X = (x_i | i = 1, 2, \dots, n)$ を回転パラメータ $\Theta = (\theta_i | i = 1, 2, \dots, m)$ を持つ回転行列でベクトル Y へと変換した場合、変換前後の内積角 θ' は以下ようになる [7, 8]。

$$\begin{aligned} \cos \theta' &= \frac{X \cdot Y}{|X||Y|} \\ &= \frac{1}{|X|^2} \sum_{i=1}^m v_i^t R_2(\theta_i) v_i \\ &= \frac{1}{|X|^2} \sum_{i=1}^m |v_i|^2 \cos \theta_i \end{aligned} \quad (16)$$

ただし、 v_i はベクトル X の $2i-1, 2i$ 次元目の成分で構成される 2 次元ベクトルである。

固有値分析との対応から θ_i を回転角の小さいものから順に第 1 固有角、第 2 固有角と呼ぶ事にする。さらにこれらの固有角に対応する 2 次元の部分空間を第 1 回転平面、第 2 回転平面と呼ぶ。固有姿勢の α に対する依存性が低いことから、回転平面は α によらず一定であると仮定する。各回転平面における固有角について、 α との対応を示したものが Fig. 2 である。Fig. 2 によれば、 α の変化に応じて各回転平面上でケプストラムベクトルが連続的に回転していることがわかる。

Table 1 音響分析条件

サンプリング条件	44.1 kHz / 16 bit
ダウンサンプリング	44.1 kHz $\rightarrow$ 16 kHz
フレーム窓	Hamming window
フレーム	25 ms length / 5 ms shift
ケプストラム	FFT Cepstrum (1-12)

今、実際の音声、これらのパラメータの回転平面に着眼して分析することを考える。実際の音声は様々な要因で変化するため、全ての固有回転角が Fig. 2 に示すように規則的に変化する訳ではないと考えられる。しかし声道長変換行列 A では語彙情報を保存したまま音声情報を変換しているため、声道長に関する情報を多く含んでいる。そのためこれらの要因が 2 次元部分空間上で異なる回転性として表出している。すなわち 2 次元部分空間への射影成分に着目する事で、二つのベクトル間の回転的な差異を調べる事ができる。この情報は音声に含まれる話者情報の表現として有用となりうる。

5 実験

5.1 実験方法

実際に話者の情報がどのように回転性として表出するかを調べるために、実験を行った。音声サンプルとして科警研の協力のもと、身長が 160 cm, 170 cm および 180 cm の 3 群についてそれぞれ 5 名ずつ、計 15 名の成人男性話者の音声を用いた。発話内容は日本語 5 母音の孤立発声であり、各母音につき 2 発声存在する。これらについて Table 1 の分析条件のもとケプストラムベクトルに変換した。これらのベクトル時系列について、発話区間のみを切り出し、発話区間での平均ベクトルを 1 発声の特徴量とした。

以下の実験では全て、対象とする 2 発声のベクトル間の回転角に着眼して分析を行った。このときベクトル間の回転角として、1) 式 (8) に基づく 2 ベクトル間の内積角、2) 声道長変換行列 A の回転平面に基づく各部分空間上での回転角の二つの指標について分析した。2) の回転角（以下固有角と呼ぶ）は、声道長変換行列 A の固有ベクトルによって張られる基底空間に着目した。これは式 (14) における姿勢行列 R_n'' に相当する。この基底空間において着目する回転平面への射影ベクトルを求める事で、各回転平面上で対象とする 2 発声のベクトル間の固有角を算出した。なお A の固有ベクトルの基底空間は α にあまり依存しないことから [7]、今回は $\alpha = 0.05$ の A について基底空間を求めた。

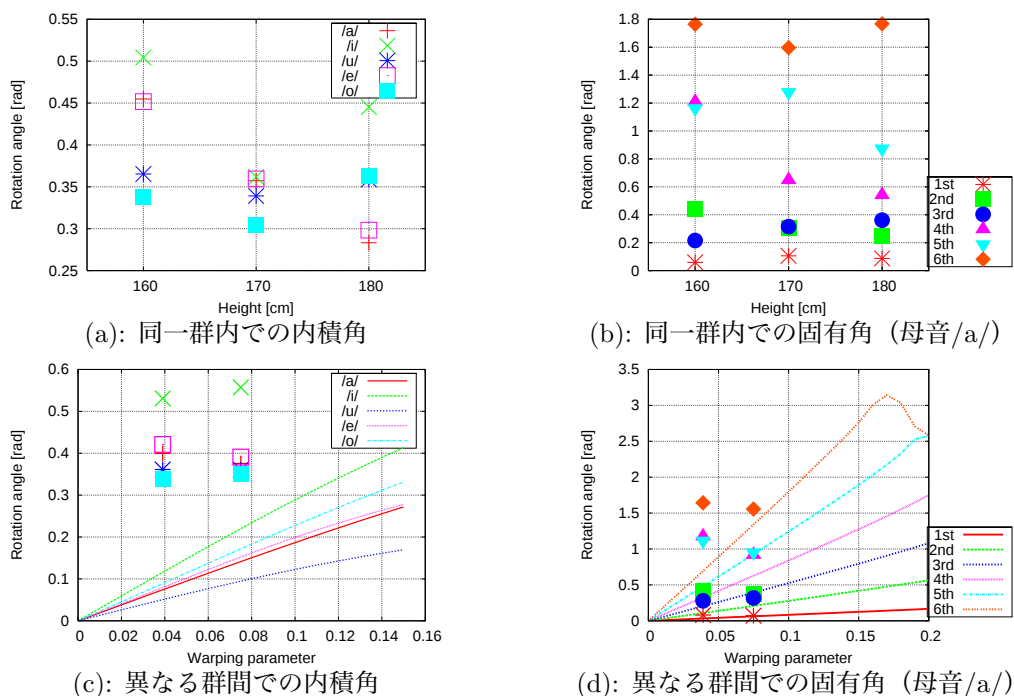


Fig. 3 実験結果

5.2 実験結果

同一群内および異なる群間の音声について、内積角および固有角を算出した結果を Fig. 3 に示す。Fig. 3 において (a) は同一身長群内での 2 音声間の内積角の平均を母音毎にプロットしたものである。一方 (b) は、母音/a/ について、同一群内での固有角の平均をプロットしたものである。また (c) および (d) は異なる群間での内積角および固有角とその身長比の関係を示している。身長比とウォーピングパラメータの対応は先行研究 [6] を参照した。(c) および (d) におけるグラフはそれぞれ、(c): 実験データ各母音 30 発声の平均声を行列 A で変換した時の内積角、(d): 行列 A の固有角 (Fig. 2) を同時に示している。

(a) より、同一身長の話者の同一母音であっても、内積角は大きな変動がある。これは内積角が曖昧性を有していることを示している。一方 (b) のように固有角に分解した場合、低次の固有角と高次の固有角との間に差があることがわかる。内積値は式 (16) のように固有角の線形和となるため、発話内の変動が特に高次の固有角の変動によって表れたと考えられる。

(c) のように、異なる群間での比較においても、内積角は行列 A による値よりも大きくなる。これについても話者性の多様な違いが影響していることが考えられる。一方 (d) のように固有角に分解して捉えた場合、第 1 固有角および第 3 固有角は、行列 A による値との対応がよく、身長情報をうまく表現しうる可能性がある。一方その他の固有角は大きなばらつきがあり、これらの回転角は声道長以外の話者性や発話の揺らぎの影響を受けやすい事が示唆される。

6 おわりに

本稿では、ケプストラムの回転性に着眼して、話者情報を抽出することに関して検討を行った。筆者らはこれまでに声道長変換を表す行列がケプストラム空間において強い回転性を有する事に言及し、固有値分析を通して部分空間上の回転性について分析してきた。今回内積角および部分空間上の固有角を実際の音声と比較し、適切な部分空間を選択すれば、実際の音声の声道長の情報をうまく抽出しうることを示唆した。今後は高次の固有角の変動と話者性および発話の揺らぎの関係を検討し、ケプストラムの回転性に基づく話者認識を検討していく。

参考文献

- [1] J. Samuelsson, ICASSP, vol.1, I-165-8, 2004.
- [2] 長内他, 音講論 (春), 1-11-18, pp.57-58, 2006.
- [3] M. Russel, S. D'Arcy, SLaTE2007, 2007.
- [4] E. Eide, H. Gish, ICASSP, vol. 1, pp. 346-348, 1996.
- [5] 齋藤他, 信学技報, SP2007-128, pp.189-194, 2007.
- [6] D. Saito et al., ICASSP, pp.4485-4488, 2008.
- [7] 齋藤他, 音講論 (春), 1-Q-19, 2008.
- [8] D. Saito et al., INTERSPEECH2008, 2008 (to appear).
- [9] 江森, 篠田, 電子情報通信学会論文誌 D-II, Vol.J83-D-II, No.11, pp.2108-2117, 2000.
- [10] M. Pitz, H. Ney, IEEE Trans. Speech and Audio Processing, Vol.13, pp.930-944, 2005.
- [11] H. Kawahara et al., Speech Communication, 27, 3-4, pp.187-207, 1999.
- [12] 高橋他, Visual Computing/グラフィックスと CAD 合同シンポジウム, 2007.