

1 はじめに

音響的普遍構造は、音声に対して不可避に混入する線形変換性・乗算性歪みを一切保有しない音声表象として提案されている [1]。これは、確率論・情報論・相対論による構造音韻論の物理実装に相当する。[2] では本表象を、語学学習者の表象として利用しているが、60 文発声を用いて音素群の構造化を検討している。構造化をかけるデータ量を減らした場合、例えば一発声を構造化した場合に、どの程度非言語情報が消失されるのか、は興味深い対象である。本稿では、非言語情報として話者性に着眼し、一発声の構造化による消失の度合いを定量的に分析する。

2 音響的普遍構造 [2]

2.1 音声の構造化

音声に含まれる各音素を、ケプストラムベクトルによって構成される多次元ガウス分布と仮定する。今、 n 個の音素に対する ${}_nC_2$ 個の音素間距離を求めることで、音声を構造化することを考える。音素間距離としてはバタチャリヤ距離の平方根を用いる。分布 u と分布 v のバタチャリヤ距離は以下の式で与えられる。

$$BD(u, v) = -\ln \int_{-\infty}^{\infty} \sqrt{P_u(x)P_v(x)} dx \quad (1)$$

P_u, P_v は、それぞれ u, v の確率密度関数である。このとき、全てのケプストラムベクトル c に対して共通の一次変換 $Ac + b$ をかけた場合、バタチャリヤ距離はその前後で不変である。音声認識・合成の話者・環境適応などで用いられているように、行列 A の乗算は声道長特性、聴覚特性など話者・聴取者の生理学的な差異を表現し、ベクトル b の加算は音響機器の差異を表現する。そして、これらの不可避的な歪みに対して、上記で定義される構造は一切影響を受けない。

2.2 構造間差異のスカラー化

M 個の分布からなる二つの構造 $\{P_i\}, \{Q_i\}$ ($1 \leq i \leq M$) を考える。各構造の重心を P_G, Q_G とすると、距離尺度としてバタチャリヤ距離の平方根を用いた場合、以下の近似式が成立する。

$$\sqrt{\frac{1}{M^2} \sum_{i < j} (\overline{P_i P_j} - \overline{Q_i Q_j})^2} \quad (2)$$

$$\approx \sqrt{\frac{1}{M} \sum_i (\overline{P_i P_G} - \overline{Q_i Q_G})^2} \quad (3)$$

今、二つの構造の重心が G に重なるように両者をシフトさせ、いずれかの構造を重心周りに回転させることで $\sum |\theta_i|$ ($|\theta_i| = \angle P_i G Q_i$) を最小にすることを

考える。 $|\theta_i|$ が十分に小さい時、 $|\overline{P_i G} - \overline{Q_i G}| \approx \overline{P_i Q_i}$ が成立するので、 $\sum |\theta_i|$ が十分に小さい時には、

$$\sqrt{\frac{1}{M^2} \sum_{i < j} (\overline{P_i P_j} - \overline{Q_i Q_j})^2} \approx \sqrt{\frac{1}{M} \sum_i \overline{P_i Q_i}^2} \quad (4)$$

が成立する。即ち、式 (2) は距離行列のみを用いることで、 $\sum |\theta_i|$ が十分に小さいという条件の下、シフト & 回転後の二つの構造間差異の近似式となる。

3 話者性の消失に関する定量的実験

一発声の構造化でどの程度非言語情報が消失されるのかを分析するため、本稿では複数話者が複数回発声した孤立 5 母音をそれぞれ構造化し、話者間構造差異と話者内構造差異を比較することで、話者性の消失度合いを定量的に分析した。また、分析条件による消失度合いの変動についても調べた。

3.1 孤立 5 母音の収録

成人男性 8 人、女性 7 人に対して孤立 5 母音を 5 回発声させた。この際、構造サイズが調音努力などの影響を受けるので [3, 4]、以下に示す方法で発話スタイルの統一を試みた。まず孤立 5 母音のサンプル音声をスピーカー提示し、その直後に、発話スタイル、継続長などを真似る形で発声させた。なお、 F_0 については話者にとって自然な F_0 で、かつ一定となるよう指示した。最終的に、各話者の 5 つの 5 母音構造を分析し、そのサイズが極端に大きい・小さい話者を除き、男女 4 名ずつを分析対象とした。

3.2 話者性の消失に関する分析

8 人の話者の 5 つの構造に対して、話者間構造差異 (${}_8C_2 \times 5^2 = 700$ 個)、話者内構造差異 ($8 \times {}_5C_2 = 80$ 個) を求め、両者の比較を通して話者性の消失の度合いを検討する。分析条件を表 1 に示す。

表 1. 分析条件

サンプリング	16bit / 16kHz
窓	窓長 25msec、シフト長 10msec
分析区間	中心部の 1 モーラ相当区間 (140msec)
パラメータ	FFTcep.(1~12 次元)

構造差異の定義としては、(I) 構造のシフト & 回転を行わずに式 (4) 右辺を用いる方法と、(II) シフト & 回転後の構造差異近似式 (2) を用いる方法を試みた。また、音声を構造化した場合、構造サイズの正規化を行なった上で二つの構造を比較することも可能となる。これは構造を音声認識に利用しようとする場合などにおいて、特に有効と考えられる。そこで、(II') 各構造が平均構造サイズと等しくなるよう正規化した後の式 (2) による構造差異も検討した。

* Quantitative analysis of removal of non-linguistic information in speech by structuralizing speech events

By Takao Murakami, Nobuaki Minematsu, and Keikichi Hirose (University of Tokyo)

表 2. 分析結果

構造差異の導出方法	(I)	(II)	(II')
話者間差異の平均	13.8	3.92	2.40
話者内差異の平均	4.99	3.68	2.22
危険率 (p)	0.0%	13.9%	1.97%

それぞれ話者間差異と話者内差異の平均を求め、話者間・話者内差異に対する分散分析を行ない、その有意差を検定した。結果を表 2 に示す。

(I) と (II) を比較すると、(II) の方が話者間差異が小さく、5%水準で有意差がないことが分かる。これは、音声の構造化によって話者性の消失が行なわれていることを示す。また、正規化を行なわない方が、より話者性が消失されることが (II) と (II') を比較することで分かる。これは、正規化を行なわない場合、構造サイズがばらつくので、話者間・話者内差異のばらつきが共に大きくなるのが原因と見られる。

次に話者内差異に注目してみる。同一話者内では $A = E$ と $b = \vec{0}$ であると仮定すると、話者内差異は (I) と (II) で等しくなると考えられる。しかし表 2 を見ると、(II) の方が (I) より小さい。これは、式 (2) が $\sum |\theta_i|$ が十分に小さいことを前提とした近似式であることが原因の一つだと考えている。 $|\theta_i|$ が大きい場合、 $|\overline{P_i G} - \overline{Q_i G}| < \overline{P_i Q_i}$ が成立し、その結果、式 (2) は構造差異を小さく見積もる傾向がある。

話者性の消失度合いを示す異なる指標として、「全話者間差異のうち、話者内差異の平均を下回るものの割合」を (II)、(II') に対して求めてみると 48.1%、39.4%であった。両分布が同一である場合 50%であることを考慮すると、(II) は話者性がよく消失していると考えられる。また、(II') は有意差があると判定されたものの、(I) に比べれば構造差異は十分小さく、話者間差異の約 4 割は話者内差異の平均を下回る。

ここで、話者間差異 < 話者内差異を満たす男性話者と女性話者の構造、及び同男性話者の二つの構造を、多次元尺度法により二次元空間にマッピングした時の図を図 1、図 2 に示す。話者性の消失に対する構造化の有効性を示す一例であると言える。

3.3 分析条件による変動

3.2 では、シフト長 10msec、分析対象区間 140msec という条件であったが、同一音声区間に対して、シフト長を短くすることでフレーム数を上げる、即ちより詳細に音響事象を捉えようとする、話者性の消失度合いはどのように変動するのかを調べてみた。

(II) について、シフト長を 10msec、5msec、2msec

表 3. 分析条件による有意差検定結果の変動

シフト長	フレーム数	有意差検定の結果
10msec	14	13.9%
5msec	28	$2.78 \times 10^{-6}\%$
2msec	70	$1.06 \times 10^{-9}\%$

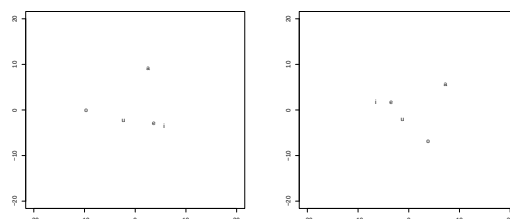


図 1. 異なる話者（男性と女性）の二つの構造

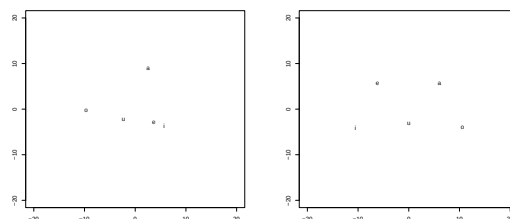


図 2. 同一話者（男性）の二つの構造

とした場合の検定結果を表 3 に示す。シフト長を短くしてフレーム数を上げると、有意差が大きくなることが分かる。音声の構造化は、静的な音響機器特性及び、話者の生理学的かつ静的な身体特性を消失するものであり、各話者特有の発声・発音の癖や方言差などは以前として構造に反映される。音響モデルに代表される音のモデル化に対して、音構造のモデル化は、これら要因がより明確化されと考えられる。分析分解能を下げることでこれらを非明瞭化することは可能であるが、逆に、話者間差異と同時に音韻差異も不明瞭になる可能性がある。ベイズ推定など少数データからのパラメータ推定なども必要であろう。

4 まとめ

孤立 5 母音を構造化することで話者性の消失がどの程度行なわれるかを分析した。結果、構造のシフト & 回転を近似する式を用いた場合、話者間差異と話者内差異の有意差が小さくなった。それでも尚、消失することのできない話者性は、話者特有の発声・発音の癖、方言差によるものと考えている。現在、スペクトルの高域（約 2.2kHz 以上）情報を明示的かつ事前に消去することによる効果も検討している。

参考文献

- [1] 峯松他, “構造音韻論の物理実装に基づく音声の新しい音声表象”, 電子情報通信学会音声研究会, SP2004-27, pp.47-52 (2004-6)
- [2] 峯松, “音声の音響的普遍構造の歪みに着目した外国語発音の自動評定”, 電子情報通信学会音声研究会, SP2003-180, pp.31-36 (2004-1)
- [3] 浜野他, “音声の分節的特徴に着目したパラ・非言語情報推定に関する実験的検討”, 電子情報通信学会音声研究会, SP2003-197, pp.143-150 (2004-3)
- [4] 朝川他, “音声の音響的普遍構造に着目した英語強勢・弱勢母音の分析”, 日本音響学会春季講演論文集, 3-7-17, pp.315-316 (2004-3)
- [5] 北村他, “スペクトル高域成分の変形と話者識別”, 日本音響学会春季講演論文集, 3-9-20, pp.397-398 (????-?)