音声の音響的普遍構造に着眼したパラ・非言語情報推定に関する実験的検討 *

◎ 籐野 真紀（東大・情報理工）　峯松 信明（東大・新領域）　広瀬 啓吉（東大・情報理工）

1　はじめに

人間間の音声コミュニケーションを観測すると，音声の音響情報から様々なパラ・非言語情報を抽出することで円滑なコミュニケーションを実現していることが分かる．本研究では，いくつかのパラ・非言語情報に関してその自動抽出を分節的特徴に着眼して検討した．この際，音声の分節的側面に対する新しい物理表象である音響的普遍構造分析を用いた．本分析は音声に内在する乗算性及び線形変換性歪みを表現する次元を保有しない，新しい音声の物理表象として峯松により提案された [1]．音声事象を確率論的に有限個の状態として記述し，状態間距離を情報論的に算出し，最終的に音声事象を相対論的に状態群が成す構造として捉えると，その構造は乗算性及び線形変換性歪みに一切影響を受けない．

既に，この音響的普遍構造に着眼することで，感情や発話意図などのパラ・非言語情報を含む音声に対し，その構造に様々な変化が表れることが示されている [2]．即ち，パラ・非言語情報の推定において，構造に着眼することの有効性が示唆されている．一方，パラ・非言語情報の推定には，一般的に韻律情報が用いられている．そこで本研究では，韻律情報の他に構造の情報を用い，構造の情報がパラ・非言語情報推定に与える効果を実験的に検討した．

2　音声に内在する普遍構造

音声事象をケプストラムベクトルによって構成される (混合) ガウス分布であると仮定する．事象間距離をバタチャリヤ距離で算出する．

$$
\begin{aligned}
&BD(u,v)\\
&= -\ln \int_{-\infty}^{\infty} \sqrt{P_u(\boldsymbol{x})P_v(\boldsymbol{x})}d\boldsymbol{x}\\
&= \frac{1}{8}\mu_{uv}^T \left(\frac{\sum_u + \sum_v}{2}\right)^{-1} \mu_{uv} + \frac{1}{2}\ln \frac{|(\sum_u + \sum_v)/2|}{|\sum_u|^{\frac{1}{2}}|\sum_v|^{\frac{1}{2}}}
\end{aligned}
$$

μ_u は u の平均ベクトル，μ_{uv} は $\mu_u - \mu_v$ を，$\sum_u$ は u の分散共分散行列を意味する．空間内の n 点に対して ${}_nC_2$ 個だけ存在する対角線の長さのみを抽出することは，n 点で張られる構造を規定することに等しい．さて，バタチャリヤ距離は，2 つの分布に対して共通の如何なる一次変換 $Ax + b$ を施しても距離は変わらない．つまり，A を掛ける演算は構造の回転を意味し，b を足す演算は構造の移動を意味し，構造は不変である．ケプストラム空間において，b を足す演算は乗算性歪みを意味し，収録環境・音声機器の差異，更には話者性の一部を表現する．A を掛ける演算は，例えば周波数ウォーピングを意味し，これは声道長の差異による音響的差異や，聴取者間の聴覚特徴の差異を表現する．

以上の議論より構造化された音声事象は，音声の生成 · 収録 · 伝送 · 再生 · 聴取の過程において不可避に混入する静的歪みに一切影響を受けないことが分かる (音声の音響的普遍構造)．

音響的普遍構造は，構造のサイズに関しても実験的にその意味付けが行われている．構造のサイズは，強弱勢や長短など，調音努力を表すものとして解釈できる [3]．また，非言語情報を含む音声に対し，調音努力の大小を意図表出の大小と考えれば，その構造に大きな違いが表れ，パラ・非言語情報推定において構造を見ることの有効性が示唆されている [2]．

そこで，本研究では，この構造の情報と韻律情報を用いて，パラ・非言語情報の推定を実際に試み，実験的に検証したので，以下にそれを報告する．

3　音響的普遍構造に基づくパラ・非言語情報分析

3.1　使用したデータ

使用したデータは，劇団経験者による 16 種類のパラ・非言語情報を込めた，連続発声音声/aiueo/(各 3 回) である．劇団経験者は，計 6 名であり，そのうち 1 名 (見本話者) の音声を見本音声，残り 5 名 (真似話者) が真似音声とした．即ち，真似話者は見本話者の音声を収録前に毎回聞き，可能な限り真似て頂いた．

パラ・非言語情報：

推定を試みたパラ・非言語情報は以下の A〜P まで 16 種類である．A）震えるほどの「悲しみ」，B）押し殺した「怒り」(cold anger)，C）仕方なく，D）恥じらい，E）ささやき，F）恐怖，G）感謝，H）楽しく・うきうき気分，I）迷い，J）ごまかし，K）ため息，L）自慢，M）明確，N）激しい「怒り」(hot anger)，O）思い切り，P）感謝 (より大袈裟に)

ここで，この A〜P の定義には，以下のステップを踏んでいる．[**step0**] 予め，見本話者 1 名に感情・意図を自由に表現してもらい，29 種類のパラ・非言語情報を収録．以下，29 種類→ 16 種類に分類する作業である．[**step1**]step0 で得た音声試料を筆者が聞き，任意の 2 種類の違いを 5 段階評価し，29 × 29 の距離行列を作成．[**step2**] この距離行列を Ward 法によって樹形図化 (Ward 法は累積歪みが最小となるようにマージ対象の 2 要素を選択するボトムアップクラスタリング手法である)．[**step3**] 木の高さに対してある一定の値で区切り，カテゴリー数を減らし

∗ Experimental study on estimation of the para- and non-linguistic information in speech based on the acoustic universal structure.

By Maki Fujino, Nobuaki Minematsu and Keikichi Hirose (University of Tokyo)

最終的に，上記の A〜P まで 16 種類のパラ・非言語情報カテゴリーを得た．

3.2 分析条件

分析条件を表 1 に示す．説明変数に関しては後述する．また，データ解析には，データマイニングツールである CART5.0[5] を用いた．

表 1. 分析条件

データ	各種パラ・非言語情報付き aiueo 連続発声 3 回，6 話者（見本 1 名，真似 5 名)
サンプリング	16bit/16kHz
窓	シフト長 10 ms，窓長 25 ms
説明変数	F0,power,duration, 音響的普遍構造に関する情報

説明変数：分析に用いた説明変数は，以下のように大きく 3 パターンに分けられる．[パターン X：単語単位の韻律情報] 単語 (/aiueo/) 単位で以下に記す韻律情報を取った．F_0・パワーに対して，平均，最大値，最小値，レンジ．ΔF_0・Δ パワーに対しては，平均，最大値，最小値，レンジ，絶対値の最小値．そして，継続長の計 19 次元である．ただし，F_0 に関しては，男女差を考慮して，「平静」の値を各々の非言語情報から引く処理をした．[パターン Y：母音単位の韻律情報] パターン X の 19 種類の韻律情報を，母音単位で取った．よって，95 次元である．[パターン Z：音響的普遍構造の情報] 各母音間の距離 (10 次元)+構造サイズ (1 次元) の計 11 次元である．なお，分布間距離の計算は各母音を分布として表現することを要求するが，分布化は MAP 推定で行った [4]．

3.3 評価実験 1

第 3.2 節で述べた説明変数の「パターン X(単語単位の韻律情報 19 次元), Y(母音単位の韻律情報 95 次元),Z(構造の情報 11 次元)」の組合せを変えながら，第 3.1 節で述べた，A〜P16 種類のパラ・非言語情報の識別実験を行った．話者 open であり，学習データに真似話者のデータを，テストデータには見本話者のデータを用いた．説明変数 X，Y，Z の組合せ毎に，識別結果の正解率の平均を，表 2 に示す．

表 2. パラ・非言語情報識別の正解率 (%)

X + Y	Y + Z	X + Z
66.7	**83.3**	56.3

X〜Z の中では，Y は 95 次元と他の 2 つに比べて圧倒的に大きく，CART には説明変数の数の増加に伴い識別率が向上する傾向がある為，Y による効果が大きくなると予想される．事実，Y+Z(83.3 %)>X+Y(66.7 %)>X+Z(56.3 %) となり，Y を使う時ほど，識別率は高くなっている．

そこで，Y を中心に表 2 を見てみると，Y+Z(83.3 %)>X+Y(66.7 %) となっており，Y と共に X を使うよりも Z を使う方が識別率が 16.6 %高くなっていることが分かる．Z(構造の情報) を使うことの有効性が明らかになった．

以上より，構造の情報は，いくつかの韻律情報と共に使うことで，パラ・非言語情報推定において効果的な役割を果たしていると解釈できる．

3.4 評価実験 2

本節では，第 3.2 節の Y(母音単位の韻律情報)+Z(構造の情報) に焦点をしぼる．

構造のサイズは継続時間や強弱勢などを表現する [3]．即ち，音響的普遍構造は，F_0 情報以外の，スペクトル，パワー，継続長などの情報を統合的に含んでいる．そこで，「パターン Y」を更に，F_0 に関するの情報 (以後，「パターン Y_1」) と，F_0 以外のパワーや継続長に関する情報 (以後，「パターン Y_2」) に分割し，説明変数を「Y_1(母音単位の F_0 情報 45 次元) + Z(構造の情報 11 次元)」と「Y_2(母音単位のパワー・継続長の情報 50 次元)+Z(構造の情報 11 次元)」にして，第 3.2 節と同じように評価実験を行った．識別結果の正解率の平均を，表 3 に示す．

表 3. パラ・非言語情報識別の正解率 (%)

Y_1 + Z	Y_2 + Z
75.0	47.9

表 3 を見ると，Y_1+Z では識別率が 75.0 %と Y_2+Z のそれよりも 27.1 %高くなっている．これにより，構造の情報と共に用いる韻律情報の中では，特に F_0 に関する情報が効果的な役割を果たしていると解釈できる．

4 まとめ

音響的普遍構造の情報は，感情や発話意図などのパラ・非言語情報の推定において，韻律的情報と用いることで，より効果的な役割を果たすことが確認できた．また，この韻律的情報の中でも，特に (母音単位の)F_0 の情報を用いると効果的であることが示された．

今後の課題としては，まず，安定した構造化の手法が上げられる．今回は母音の連続発声を用い，強制アラインメントを取ったが，母音境界の検出に関し，一部誤りがあった．極端な検出誤りについては，手修正を加えている．また，今回は母音のみであったが，子音を含む音声を用いた構造化手法についても検討する必要がある．

参 考 文 献

[1] 峯松他，”構造不変の定理に基づく音声の構造的表象とその距離尺度”，日本音響学会春季講演論文集，1-5-13，pp.25-26 (2005)．

[2] 浜野他，”音声の分節的特徴に着眼したパラ・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.25-30 (2004)．

[3] 朝川他，”音響的普遍構造のサイズと単語境界における音響的分離度に着眼した米語音声の音響分析”，電子情報通信学会音声研究会，SP2004-28，pp.53-58 (2004)．

[4] 丸山他，”音声の構造的表象に基づく音響的照合に関する実験的検討”，日本音響学会春季講演論文集，1-5-14，pp.27-28 (2005)

[5] http://www.hulinks.co.jp/software/cart