Vol.2018-MUS-119 No.48
Vol.2018-SLP-122 No.48
2018/6/17

符号化モデルを用いた音楽ジャンルの脳内情報表現の可視化

中井智也[†1,2]　小出（間島）真子[†3]　西本伸志[†1,2,4]

概要：音楽ジャンル認識は，ヒトの音楽に対する嗜好性を理解する上で中心的な課題である．先行研究では，音楽ジャンル認識は様々な音響特徴量に基づいて行われてきたが，個々の音楽ジャンルがどのように脳内で表現されているかは知られていなかった．本研究では，被験者に 540 曲の音楽刺激を聞かせ（各 15 秒），その際の脳活動を 3T の MRI 装置で計測した．脳活動（R）と特徴量（F）の関係を表す符号化モデル（R＝FW）を L2 正則化付き線形回帰でフィッティングし，重み行列（W）から各音楽ジャンルの脳内表現を可視化したところ，両半球の上側頭回で音楽ジャンルが表現されていることがわかった．中でもクラシックとヒップホップは正反対の脳活動パターンを示していたが，その脳活動パターンは，聴覚野情報表現のモデルである Spectro-temporal receptive field（STRF）モデルで抽出した時間スペクトル特性により説明することができた．また今回検証した 4 種類の音響特徴量モデルのうち，STRF モデルが最も高い精度を示した．これらの結果は，音楽ジャンル認識の生物学的基盤に関して新しい知見を提供するものである．

Visualization of cortical representation of music genres using encoding model

TOMOYA NAKAI[†1,2]　NAOKO KOIDE-MAJIMA[†3]
SHINJI NISHIMOTO[†1,2,4]

1. はじめに *

音楽ジャンル認識は，ヒトの音楽に対する嗜好性を理解する上で中心的な課題である．先行研究では，音楽ジャンル認識は様々な音響特徴量に基づいて行われてきたが[1]，個々の音楽ジャンルがどのように脳内で表現されているかはこれまで知られていなかった．

音楽理解の脳神経メカニズムの先行研究において，音圧，音色，調性，リズムなどの音楽情報において利用される特徴量と，両半球の上側頭回を中心とした脳活動との相関が報告されている[2], [3]．他方，上側頭回を対象としたより低次聴覚情報処理研究においては，一次聴覚野の細胞応答特性のモデルである spectro-temporal receptive field（STRF）がよく用いられてきた[4]–[6]．

本研究では，被験者に 10 種類のジャンルの音楽刺激を聴かせ，その際の脳活動を磁気共鳴画像法（MRI）で計測した．また，音楽刺激から 4 種類の特徴量（Cochlear, STRF, MIR，MFCC）を抽出し，得られた脳活動に対し，符号化モデルにより，どのような特徴量によって音楽ジャンルの脳内表現が説明できるのかを検証した（図 1）．

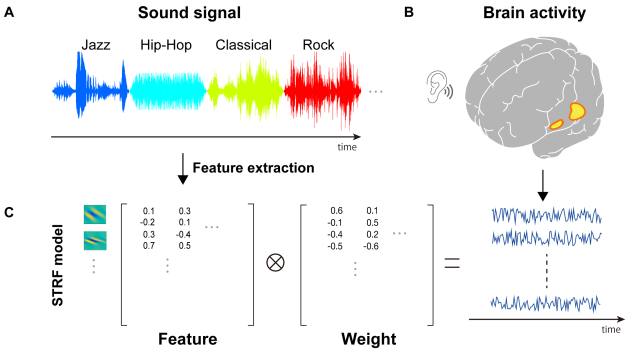

図 1　研究パラダイムの概略図

Figure 1　Schematic image of the research paradigm

2. 方法

2.1 被験者

5 名の正常な聴力を持つ男女が実験に参加した（23-33 歳，ID01-05 と対応）．実験に先立ち，被験者には実験内容に関して十分に説明をした上で，同意書に記入をしてもらった．この実験は情報通信研究機構の倫理審査委員会で承認されたものである．

2.2 刺激および課題

本研究では，GTZAN データベースから 10 ジャンルに渡る 540 曲（各ジャンル 54 曲）の楽曲を使用した（30 秒，

* †1 情報通信研究機構 脳情報通信融合研究センター
Center for Information and Neural Networks (CiNet), NICT
†2 大阪大学大学院生命機能研究科
Osaka University, Graduate School of Frontier Biosciences
†3 ブラザー工業株式会社
Brother Industries LTD
†4 大阪大学大学院医学系研究科
Osaka University, Graduate School of Medicine

Vol.2018-MUS-119 No.48
Vol.2018-SLP-122 No.48
2018/6/17

22,050 Hz）[7]. 各楽曲から 15 秒間のクリップを抜き出し，平均音圧を統制した．被験者に 3 日間にわたり 18 回の撮像（各 10 分）を実施し，合計 540 クリップの音楽刺激を聴かせた．その際の脳活動を 3T の MRI 装置（Siemens Trio Tim）で計測した．

撮像には 32 チャンネル・ヘッドコイルを用い，機能的 MRI として以下のパラメータを使用した：（EPI，TR = 1500 ms，TE = 30 ms，FA = 62°，FOV = 192 × 192 mm^2，voxel size = 2 × 2 × 2 mm^3，multi-band factor = 4）．また解剖画像として以下のパラメータを使用した：（MPRAGE，TR = 2530 ms，TE = 3.26 ms，FA = 9°，FOV = 256 × 256 mm^2，voxel size = 1 × 1 × 1 mm^3）．

2.3 聴覚および音楽情報モデル

10 個の音楽ジャンルごとに異なる活動パターンを抽出するため，one-hot ベクトルを集めた 10 次元の音楽ジャンルモデルを作成した．

各音楽クリップに対し，20-10,000 Hz に渡る 128 個のガンマトーンフィルタ列によりコクレオグラムを作成した[8]．Cochlear モデルにおいては，各ガンマトーンフィルタの出力を 128 次元の特徴量とした．STRF モデルとして，各コクレオグラムに対し，100 種類の時間スペクトルフィルタ（10 段階スケール：Ω = 0.35-8.0 cyc/oct，10 段階時間変調 ω = 2.8-64.0 Hz）を用意した[9]．コクレオグラムに対するフィルタの畳み込み演算結果を，対数周波数軸上の 20 区間および 1.5 秒ごとにそれぞれ平均した．得られた 2000 次元の特徴量を 99%の成分を保持するように特異値分解し，302 次元に削減した．

さらに，先行研究で用いられていた音楽情報処理（Music information Retrieval，MIR）の特徴量として，MIRtoolbox を用いて音圧，音色，調性，リズムに関する 24 次元の特徴量を抽出した [2], [10]．また，同じく MIRtoolbox を用いて 12 次元のメル周波数ケプストラム係数（mel-frequency cepstrum coefficients，MFCC）を抽出した．

2.4 符号化モデル解析

Statistical Parametric mapping（SPM8）を用いて体動の補正を行い，さらに各被験者の全 EPI 画像を参照用画像に合わせ込んだ．時間窓 240 秒のメディアンフィルタによりドリフトを除去し，さらに各ボクセルにおける信号を正規化した．解剖画像から皮質表面を同定し，EPI データのボクセルに合わせ込むために FreeSurfer を用いた．

脳活動（R）と特徴量（F）の関係を表す符号化モデル（R = FW）を L2 正則化付き線形回帰でフィッティングすることで，重み行列（W）を推定した（図 1）．脳活動の血流動体反応を反映させるため，1.5，3，4.5，6 秒の時間遅れ成分を特徴量行列に加えた．訓練用データサイズは 4800，テストデータサイズは 600 であった．訓練用データを 80%と 20%のサンプルに分けてモデルフィッティングを 10 回繰り返すことで，最適な正則化パラメータを推定した．

テストデータは 4 回繰り返しに関し平均を取ることにより信号ノイズ比を向上した．予測された脳活動と計測データのピアソン積率相関係数を計算することにより，各モデルによる脳活動の予測精度を算出した．False-discovery-rate（FDR）による多重比較の補正をおこなった[11]．

訓練用データを 80%と 20%に分けて音楽ジャンルモデルによるモデルフィッティングを 50 回行い，8 割以上の試行において有意であったボクセルを Region-of-interest（ROI）として定義した．

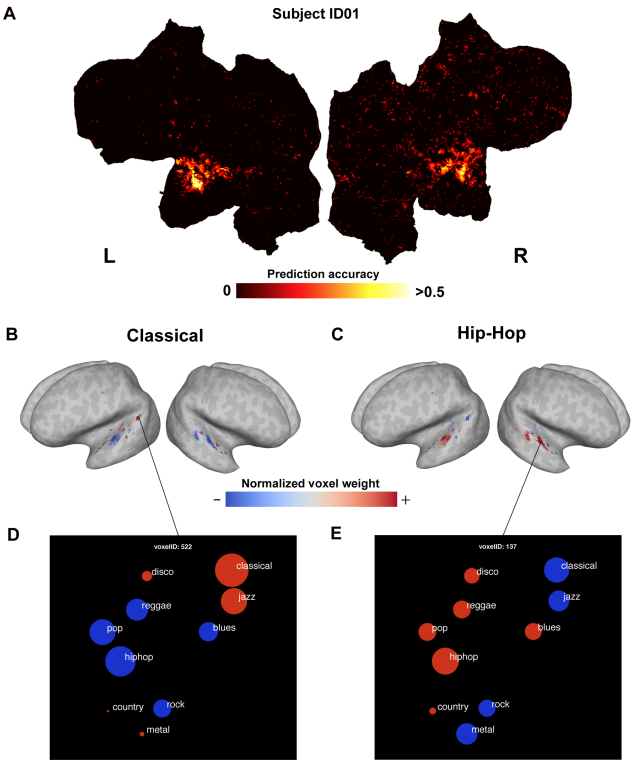

図 2 音楽ジャンルの脳情報表現

Figure 2 Cortical representation of music genres

3. 結果

3.1 音楽ジャンル特異的な脳活動パターン

音楽ジャンルモデルにより皮質上で予測精度を計算すると，両半球の上側頭回で音楽ジャンルが表現されていることがわかった（$p < 0.05$，FDR 補正）（図 2 の A，被験者 ID01 のみを例示）．音楽ジャンルモデルで作成した ROI を以後の解析でマスクとして用いた．推定した重み行列を用いて各音楽ジャンルによる脳活動パターンのマップを得た．各ジャンルを比較すると，クラシックとヒップホップは正反対の脳活動パターンを示していた（図 2 の B，C）．このような活動パターンは 5 名の被験者に共通してみられるものであった．さらに，音楽ジャンルモデルの重み行列に対し

Vol.2018-MUS-119 No.48
Vol.2018-SLP-122 No.48
2018/6/17

t-distributed stochastic neighborhood embedding（*t*-SNE）を適用し，10 ジャンルを視覚的に 2 次元上にプロットすることにより，音楽ジャンルの脳内表現を可視化した（図 2 の D, E）．また他の 4 種類の音響/音楽モデルに関しても音楽刺激に伴う脳活動の予測精度を検証したところ，全ての被験者において，STRF モデルによる脳活動の予測精度が最も高かった（表 1）．

表 1　各音響/音楽モデルによる脳活動の予測精度
Table 1　Prediction accuracy of brain activity using sound/music-feature models

被験者	ID01	ID02	ID03	ID04	ID05
Cochlear	0.227	0.223	0.146	0.233	0.159
STRF	**0.336**	**0350**	**0.354**	**0.277**	**0.336**
MIR	0.264	0.281	0.182	0.239	0.221
MFCC	0.147	0.169	0.157	0.181	0.152

4. 考察

本研究では，10 種類の音楽ジャンルを含む音楽刺激を 5 名の被験者に聴かせ，脳活動を MRI 装置で計測した．音楽刺激から特徴量を抽出し，符号化モデルにより音楽ジャンルの脳内情報表現を可視化した．音楽ジャンル特異的に脳活動が観測されたのは両半球の上側頭回であるが，この領域は先行研究において低次聴覚特徴量および音楽特徴量に基づく脳活動が報告されていた領域であった[2]–[6].

音楽ジャンルごとの脳活動パターンとして、クラシックとヒップホップの正反対の活動パターンが特徴的であったが，*t*-SNE により得られた 10 ジャンルの脳内表現により，さらにクラシックとジャズ，ブルースの脳内表現が類似しており，またロックとメタルの脳内表現が類似しているなどの結果が得られた．クラシックは音声がほとんど含まれていないのに対し，ブルースは音声が含まれていることから，今回得られた脳内表現が単に言語的要素の有無によって生じているわけではないことが示唆される．

本研究で用意した 4 種類の音響/音楽モデルのうち，音楽刺激に伴う脳活動パターンに対する予測精度が最も高かったのは STRF モデルであった．MIR と MFCC モデルは音楽情報処理で用いられてきたが，STRF モデルのほうがより脳活動を予測できていたことは，ヒト聴覚野の細胞応答特性に基づく STRF モデルが，より生物学的に妥当性のモデルであることを示している．また Cochlear モデルと比べ，STRF モデルはスペクトルの時間的ダイナミクスを捉えることができるという点で，大脳皮質における聴覚応答を説明する上でより優れているのだと考えられる．以上の結果は，ヒト脳内における音楽ジャンル認識の時間スペクトル特性に関して，新しい知見を提供するものである．

謝辞　本研究は MEXT/JSPS 科研費 JP15H05311，JP1805091（新学術「共創言語進化」#4903）の助成を受けたものです．

参考文献

[1] Sturm, B. L. and Noorzad P.. A survey of evaluation in music genre recognition. Adapt. Multimed. Retr. Semant. Context. Adapt.. 2012, p. 29–66.

[2] Alluri V., Toiviainen P., Jääskeläinen I. P., Glerean E., Sams M., and Brattico E.. Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. Neuroimage. 2012, vol. 59, no. 4, p. 3677–3689.

[3] Toiviainen P., Alluri V., Brattico E., Wallentin M., and Vuust P.. Capturing the musical brain with Lasso: Dynamic decoding of musical features from fMRI data. Neuroimage. 2014, vol. 88, p. 170–180.

[4] Patil K., Pressnitzer D., Shamma S., and Elhilali M.. Music in Our Ears: The Biological Bases of Musical Timbre Perception. PLoS Comput. Biol.. 2012, vol. 8, no. 11, e1002759.

[5] Norman-Haignere S., Kanwisher N. G., and McDermott J. H.. Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. Neuron. 2015, vol. 88, no. 6, p. 1281–1296.

[6] Santoro R. et al.. Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. Proc. Natl. Acad. Sci.. 2017, vol. 114, no. 18, p. 4799–4804.

[7] Tzanetakis G. and Cook P.. Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. 2002, vol. 10, no. 5, p. 293–302.

[8] Ellis D. P. W.. “Gammatone-like spectrograms,” web resource. 2009, http://www.ee.columbia.edu/~dpwe/resources/matlab/.

[9] Chi T., Ru P., and Shamma S. A.. Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am.. 2005, vol. 118, no. 2, p. 887–906.

[10] Lartillot O., Toiviainen P., and Eerola T.. A Matlab Toolbox for Music Information Retrieval. Data analysis, machine learning and applications, Springer, Berlin, Heidelberg. 2008, p. 261–268.

[11] Benjamini Y. and Hochberg Y.. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B. 1995, vol. 57, no. 1. p. 289–300.