The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

聴神経の計算論再考：教師なし学習か課題最適化か？

Reconsidering the computational model for the auditory nerve: unsupervised learning or task-based optimization?

寺島 裕貴　　古川 茂人
Hiroki Terashima　Shigeto Furukawa

NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

Efficient coding has been a leading computational principle to understand the sensory systems. The auditory nerves have been explained by unsupervised learning of natural sounds: their filter shapes resemble the basis optimized to code human voice. However, the previous study did not clearly distinguish sounds recorded in studios and those in the environment. We found that the model fails to explain the auditory nerves when applied to environmental recordings because of reverberations. How can we model the auditory nerves including the environmental modulations? We hypothesized that the auditory nerves are optimized to perform auditory tasks we face in the environment. To test this, we trained a deep convolutional neural network to classify phonemes based on their reverberated waveforms. The filters learned in the first layer showed characteristics similar to the auditory nerves. The results suggest that the auditory nerves efficiently encode task-related information rather than the entire incoming signal.

1. はじめに

哺乳類の耳の物理的なサイズは種によってさまざまだが，聴神経が持つ周波数チューニングのフィルタ特性は種を越えて普遍的な分布に従うため [4]，何らかの機能的意義が示唆される．この特徴をどのように理解すればよいだろうか．

脳神経系の情報処理様式を理解する方策として，脳の各部位を計算アルゴリズムと対応付けていくという考え方がある．その中で，低次感覚系は教師なし学習と対応付くと考えられてきた．初めに，大脳皮質一次視覚野が自然画像のスパース符号を用いて説明され，神経系と自然刺激統計性が深い関係にあることが示された．これを受けて聴覚系では，自然音，その中でも特にヒト音声のスパース符号が聴神経のフィルタ特性と一致することが報告された [3]．

この研究 [3] の結果を素直に捉えると，聴覚末梢は教師なし学習とよく対応するように思われる．しかし前報 [7] では，この定説に疑問を投げかける現象を報告した．すなわち，先行研究で用いられた「自然音」よりもさらに耳への入力に近い，屋内外で話者から離れた場所で録音した音声を用いた場合には，先行研究のようには聴神経特性を説明できなかった．

本報告では，自然残響下でもなお聴神経特性を説明できるモデルとして，残響下の音素識別タスクを課した深層ニューラルネットワークを提案する．この結果は，単純な教師なし学習よりも，自然な音響的タスクに対する最適化のほうが聴覚末梢を理解するための枠組みとしてより有望であると示唆している．

2. 手法

2.1 自然音の教師なしスパース符号化

Lewicki [3] の結果と直接比較できるよう，同様の基底学習法を用いた．入力 $\boldsymbol{x}$ を $\boldsymbol{x} = \sum_i a_i \boldsymbol{\phi_i}$ のように基底 $\boldsymbol{\phi_i}$ の線形和で表現し，係数 a_i がよりスパースになる基底 $\Phi = \{\boldsymbol{\phi_i}\}$ を下記のように更新して学習した．詳細は原論文 [3] を参照．

$$\Delta\Phi \propto \Phi\Phi^T \frac{\partial}{\partial\Phi} \log p(\boldsymbol{x}|\Phi) \tag{1}$$

$$= \Phi(I - \mathrm{sign}(\boldsymbol{a})\boldsymbol{a}^T) \tag{2}$$

また，得られた基底ベクトルそれぞれのフィルタ特性の解析手法も原論文に準拠し，周波数チューニングとそのシャープネス（Q 値）を求めた．

2.2 自然音・自然残響データベース

自然な屋内環境におけるヒト音声の録音として，NTT 乳幼児音声データベース [1] のうち大人同士の会話としてタグ付けされたもののみを抽出して用いた．また，音素識別タスクには TIMIT データベースを用い，[2] に倣って 61 の音素ラベルを 39 クラスに再編した．自然な空間残響のシミュレーションに際しては，屋内外で録音された自然インパルス応答データベース [5] を用いた．

2.3 深層ニューラルネットワークによる音素識別

自然な音響的タスクへ最適化されたフィルタ形状を調査するため，音声波形を直接入力する深層ニューラルネットワークを構築した．入力は 16 kHz 長さ 2,000 の音声波形で，ネットワークの構造は 5 層の畳み込みニューラルネットワークである（詳細は図 2）．訓練後 1 層目に獲得されたフィルタの特性を [3] と同様に分析した．

聴神経フィルタの時間窓と類似の時間スケールにおける自然な音響的タスクとして，入力波形の中心部に位置する音素の識別をニューラルネットワークに課して訓練した．その際，自然残響下での課題を模擬するため，各入力には自然なインパルス応答をデータベース [5] からランダムにひとつ選んで畳み込んだ．

3. 結果

3.1 残響下自然音スパース符号と聴神経の不一致

聴覚末梢の教師なし学習モデル研究によれば，純粋なヒト音声録音のスパース符号と聴神経フィルタの特性は類似する [3]．

連絡先: 寺島裕貴，NTT コミュニケーション科学基礎研究所，〒243–0198 神奈川県厚木市森の里若宮 3–1，teratti@teratti.jp

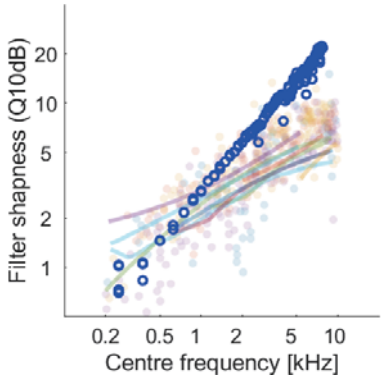

図 1　ヒト音声の自然環境下録音について最適なスパース符号の周波数チューニングとそのシャープネスの分布．聴神経の特性（半透明）よりもシャープなチューニング特性を示す．

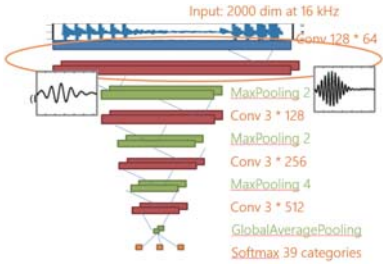

図 2　自然な残響下のヒト音声を入力とし，音素識別タスクで訓練したニューラルネットワークのアーキテクチャ．第 1 層の畳み込みフィルタと聴神経を比較する．

前報 [7] では，より自然な屋内環境で録音したヒトの音声でこの結果を再現できるかどうか調べるため，NTT 乳幼児音声データベースの中から大人同士の発話のみを抽出し，同アルゴリズムを適用した．

図 1 に各フィルタの周波数チューニングとそのシャープネスの分布を示す．背景に半透明で重畳しているのは，多種の哺乳類で記録されたデータの再プロットである [4]．前報 [7] で報告したように，自然界における残響の影響を主に受け，シャープネスが聴神経よりも高い方向にシフトした分布が得られた．

3.2　音素識別課題最適化による聴神経特性の再現

聴覚末梢を理解するために，教師なし学習以外の枠組みとして有望なものは何だろうか．近年，自然な視覚課題に対する最適化が視覚系階層性の理解に有用だと考えられつつある [6]．聴覚系の特性もまた，自然な音響的課題に対する最適化を用いて説明できるとの仮説のもと，深層ニューラルネットワークを構築して解析を行った．図 2 に示す畳み込みニューラルネットワークを，自然な残響を模擬した波形を入力して TIMIT 音素識別で訓練した．

訓練後の第 1 層に獲得されたフィルタ特性の分布を図 3 に示す．残響下の音声波形が入力されているにもかかわらず，哺乳類で見られる聴覚フィルタ特性分布を再現できていることがわかる．これは単純な教師なし学習を用いた場合（図 1）と対照的である．

4.　結論

聴覚末梢の特性は，自然音声の教師なし学習として理解できるとされてきた．しかしわれわれは前報で，音声の自然な録音であっても，聴覚系への入力により近い残響下での録音を用いると聴覚末梢の特性を説明できなくなることを報告した．本報告では，残響下の音声を入力として用いつつ，深層ニューラルネットワークを音素識別課題で訓練することで聴覚末梢の特性を説明した．聴覚末梢を理解するための枠組みとして，単純な教師なし学習よりも自然な音響的課題に対する最適化がより適切だと考えられる．

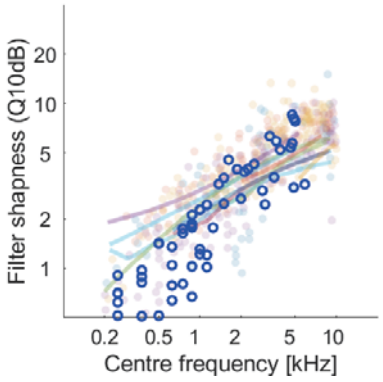

図 3　自然な残響下のヒト音声を入力として訓練したニューラルネットワーク第 1 層の周波数チューニングとそのシャープネスの分布．聴神経の特性（半透明）と似たチューニング特性分布を示す．

謝辞

本研究は JSPS 科研費 JP15H05915（新学術領域研究：多元質感知）の助成を受けたものです．

参考文献

[1] Amano S, Kato K, and Kondo T. Development of japanese infant speech database and speaking rate analysis. In *ICSLP - INTERSPEECH*, 2002.

[2] Lee KF and Hon HW. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**(11):1641–1648, 1989.

[3] Lewicki MS. Efficient coding of natural sounds. *Nature Neuroscience*, **5**(4):356–363, 2002.

[4] Ruggero MA and Temchin AN. Unexceptional sharpness of frequency tuning in the human cochlea. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(51):18614–18619, 2005.

[5] Traer J and McDermott JH. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences of the United States of America*, **113**(48):E7856–E7865, 2016.

[6] Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, and DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, **111**(23):8619–8624, 2014.

[7] 寺島裕貴, 古川茂人. 自然残響の考慮による聴覚音声スパース符号化再考. In 第 *31* 回人工知能学会全国大会論文集, 2K5–OS–33c–2. 人工知能学会, 2017.