2K5-OS-33c-2

自然残響の考慮による聴覚音声スパース符号化再考

Reconsidering sparse codes of natural sounds with natural reverberations

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

The auditory periphery has been understood as a result of unsupervised learning because its filter characteristics match sparse codes of natural sounds. In this study, we report that they turn out to be unmatched when considering more natural sounds for our ears. Hypothesizing that the discrepancy results from reverberations in the natural environments, we replicated a similar shift by simulating natural reverberations. The results suggest that there might be a more appropriate framework for the auditory periphery than unsupervised learning, e.g., supervised learning.

1. はじめに

脳神経系の情報処理様式を理解する方策として，脳の部位をアルゴリズムと対応付けていくという考え方がある [2]．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 [7]．初めに，大脳皮質一次視覚野が自然画像のスパース符号として説明され，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 [6]．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 [5]．

この研究 [5] 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．しかし本研究では，この定説に疑問を投げかける現象を報告する．まず，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環境で録音した音声を用いた場合に，先行研究のような聴神経特性との一致が見られないことを報告する．次に，この不一致の原因は自然な空間が持つ残響特性にあるという仮説を立て，インパルス応答を畳み込むことで残響のシミュレーションを行い，同様の不一致傾向を再現できることを示す．

われわれの耳には残響込みの音が届いているのにも関わらず，残響の考慮によってスパース符号は聴神経の特性と一致しなくなってしまう．この結果は，聴覚末梢が単純な教師なし学習と対応するという定説への再考を促し，例えば教師あり学習といった他の枠組みによるモデル化がより適切だという可能性を示唆する．

2. 手法

2.1 聴覚スパース符号化モデル

Lewicki [5] 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 $\boldsymbol{x}$ を以下のように基底 ϕ_i の線形和で表現することを考える．

$$\boldsymbol{x} = \sum_i a_i \phi_i \tag{1}$$

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，terashima.hiroki@lab.ntt.co.jp

係数 a_i がよりスパースになる基底を学習するため，以下の自然勾配を用いたミニバッチ（100 標本）で基底を十分な回数（> 10000）更新した．詳細は原論文 [5] を参照．

$$\Delta\Phi \propto \Phi\Phi^T \frac{\partial}{\partial\Phi} \log p(\boldsymbol{x}|\Phi) \tag{2}$$

$$= \Phi(I - \mathrm{sign}(\boldsymbol{a})\boldsymbol{a}^T) \tag{3}$$

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 [5] に準拠した．各基底ベクトルについて，まずピリオドグラムを求め（Hamming 窓），最大パワーを示す周波数をその基底ベクトルの中心周波数とする．ピークの両側でパワーが 10 dB 低下する周波数幅を使い，シャープネス（Q 値）を求めた．

2.2 自然音データベース

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．先行研究と条件を揃えるため 16 kHz にリサンプリングし，8 ms（128 点）のセグメント 100 万標本を入力として用いた．また，先行研究の結果を再現するための対照実験には，直接音録音に近い IPA 音声データベース [4] を用いた．

自然な空間残響のシミュレーションに際しては，データベース SMILE2004 [12] に収録されたインパルス応答を IPA 音声データベースに畳み込んだ．用いたインパルス応答は 22 種（各種室内 18 種・屋外伝搬 4 種）である．

3. 結果

3.1 自然環境における録音のスパース符号

聴覚末梢の教師なし学習モデル研究によれば，ヒト音声のスパース符号と聴神経フィルタの特性は類似する [5]．この研究で用いられたヒト音声は，純粋な音声録音を目的としたものだった．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 上に学習された基底を，図 1 下に周波数選択性のシャープネスがどのように分布しているかを示す．シャープネス分

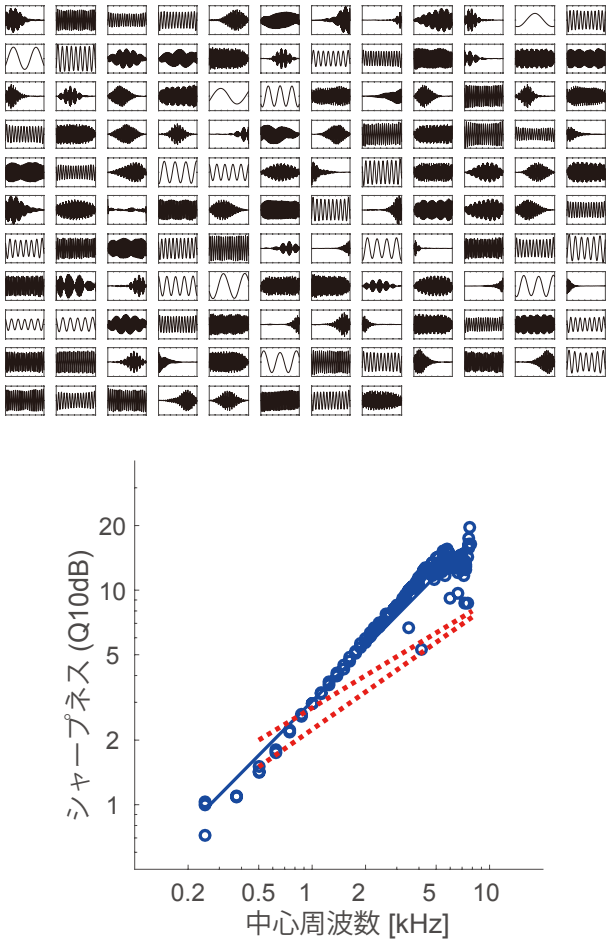

図 1: **自然な音声録音のスパース符号.** （上）学習された基底. （下）周波数選択性とシャープネスの分布.

布に二本の赤点線で示すのは，先行研究 [5] が参照しているネコ聴神経のフィルタ特性についての二研究 [3, 8] の結果の線形フィットである．先行研究によれば，ヒト音声のスパース符号は赤点線似た分布を示すはずである．しかし図に示すように，実際にはシャープネスが高い方向にシフトした分布が得られ，神経生理学的な分布と一致しない（線形フィットの結果を青線で示す）.

分布の不一致が実装上の不具合でないことを確かめるため，先行研究で用いられた音声録音により近いと考えられる IPA 音声データベースを用いて同様の解析を行った．その結果、図 2 の分布が得られた。これは神経生理学的な分布と近いもので，先行研究の結果と類似した．したがって，先に NTT 乳幼児音声データベースから得られた分布の不一致は実装上の不具合ではないと考えられる．では，なぜシャープネスが高い方向にシフトした分布が得られたのだろうか.

3.2 自然な反響を考慮した際のスパース符号

前節では，同じヒト音声の録音でも，教師なし学習を適用した際に聴神経の特性に一致する場合と一致しない場合の両方があることを示した．入力音を比較した結果われわれは，この差異の源が空間の残響にあるのではないかという仮説をたてた．話者が発した音声が聴き手の耳に届く際には，屋外であれ屋内であれ，木々や壁などによる反響が多数加わった残響が，直接音と重ね合わさってくる．先行研究で用いられた音や IPA 音声データベースが直接音に近い特性を持っているのに対し，

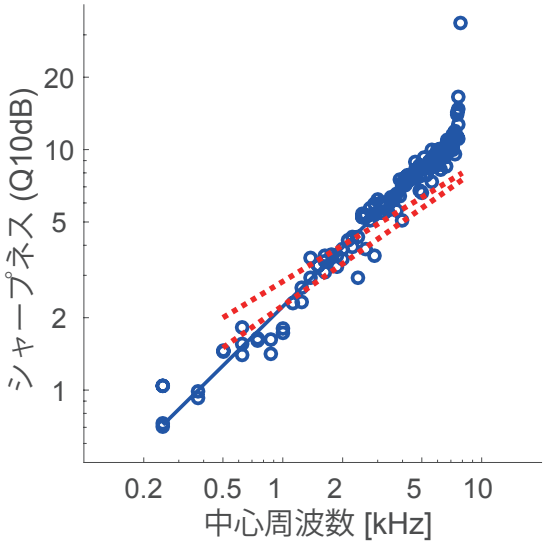

図 2: **残響が無い場合のシャープネス分布.**

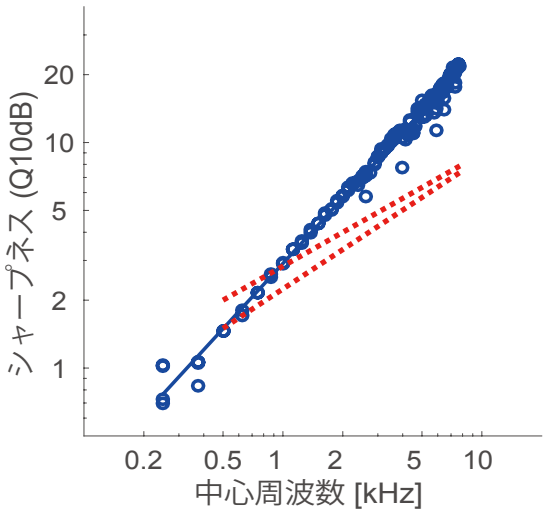

図 3: **残響をシミュレートした場合のシャープネス分布.**

NTT 乳幼児音声データベースは屋内での録音であり，多くの残響が含まれている．こうした変調の有無が聴神経との一致・不一致と対応しているのではないだろうか.

空間残響の有無が聴神経との一致・不一致を説明するとの仮説を検証するため，直接音に近い IPA 音声データベースに人工的に残響を加える実験を行った．さまざまな屋内外環境のインパルス応答データベース（SMILE2004）を原音声に畳み込んで残響込みの音声録音をシミュレートし，その結果を入力として同様の解析を行った．図 3 に，残響をシミュレートした音声を入力として用いた場合のシャープネス分布を示す．仮説の通り聴神経の特性分布よりも高い方向に分布がシフトしており，図 1 下と同様の傾向が再現された.

4. 考察

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．しかし本研究では，音声の自然な録音であっても，そのスパース符号が必ずしも聴覚末梢の特性と一致しないことを報告した．また，その違いの源が自然な空間残響にあ

るという仮説を設定し，残響をシミュレートして同様のずれ方を再現した．残響シミュレーションを行ったほうが聴覚系への入力により近いことから，聴覚末梢の単純な教師なし学習としてのモデル化には再考が必要だと考えられる．

単純な教師なし学習としてのモデル化以外には，他にどのような解釈が適当だろうか．近年，視覚野の階層性が教師あり学習としてよくモデル化できることが報告されている [11]．聴覚系も，反響を除いた音源の処理に特化していると考えれば，教師あり学習の枠組みでよく理解できる可能性がある．あるいは，聴覚末梢の特性は蝸牛の物理特性で決まっており，最適化の枠組みが不適当なのかもしれない．

では，聴覚系の中で教師なし学習は起きていないのだろうか．別の候補としては，比較的高次の処理段階にあたる大脳皮質の低次聴覚野が考えられる [9]．低次聴覚野は，最初に教師なし学習モデルが提案された一次視覚野 [6] と解剖学的特徴を共有しており，アルゴリズムも類似している可能性が高い [10]．今後は聴覚系全体を俯瞰した上で，教師なし学習と教師あり学習双方との対応を処理段階ごとに注意深く行う必要があるだろう．

参考文献

[1] Amano S, Kato K, and Kondo T. Development of japanese infant speech database and speaking rate analysis. In *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002*, 2002.

[2] Doya K. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, **10**(6):732–739, 2000.

[3] Evans EF. Cochlear nerve and cochlear nucleus. In *Auditory System*, pages 1–108. Springer, 1975.

[4] International Phonetic Association. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press, 1999.

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

[6] Olshausen BA and Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**(6583):607–609, 1996.

[7] Olshausen BA and Field DJ. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, **14**(4):481–487, 2004.

[8] Rhode WS and Smith PH. Characteristics of tone-pip response patterns in relationship to spontaneous rate in cat auditory nerve fibers. *Hearing Research*, **18**(2):159–168, 1985.

[9] Terashima H and Okada M. The topographic unsupervised learning of natural sounds in the auditory cortex. In *Advances in Neural Information Processing Systems 25 (NIPS2012)*, pages 2321–2329, 2012.

[10] Terashima H. *Computational Model for Auditory Cortex: An Analogy to Visual Cortex*. PhD thesis, The University of Tokyo, 2014.

[11] 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.

[12] 日本建築学会．DVD 版 建築と環境のサウンドライブラリ, 2004.