様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ５ 月 ２４ 日現在

機関番号：１４４０１
研究種目：挑戦的研究（萌芽）
研究期間：2018～2021
課題番号：１８Ｋ１９８１９

研究課題名（和文）深層学習でカクテルパーティ問題を解く

研究課題名（英文）Solving the cocktail party problem using deep learning

研究代表者
北澤　茂（Kitazawa, Shigeru）

大阪大学・生命機能研究科・教授

研究者番号：００２５１２３１

交付決定額（研究期間全体）：（直接経費）　　4,800,000円

研究成果の概要（和文）：私たちは、大勢の人が話しているパーティーでも相手の言うことを聞き取れる。本研究では脳と比較可能な神経回路モデルを作ることを目標に「重なった音声信号の１つに注意を向ける人工神経回路を自律学習で作り出す」という問題に挑戦した。注意の機構を備える人工神経回路モデル（transformer）に環境音データベースの音声情報を入力して、ラベルなしで情報量を最大化する自律学習を行わせたところ、環境音を複数の対象として「認識」してそのいずれかに「注意」を向ける機構が獲得できることが示された。このaudio-transformerは「カクテルパーティ問題」の謎を解くための有力な神経モデルとなるだろう。

研究成果の学術的意義や社会的意義
カクテルパーティー効果は私たちが日常で体験できる現象だが、その神経基盤は未知である。本研究では「特定の音に対して注意を向ける」人工神経回路を自律的な学習で作りだすことに成功した。最近Googleのグループなどが二人の音声を聞き分けることだけに特化した人工神経回路を発表しているが、それらは正解を与えて学習させる「教師付学習」を用いている。我々はそのような強制的な学習を行わずとも、環境音を聞いているうちに「自然に」音の特徴を使って聞き分けるように人工神経回路を育てることが可能であることを示した。ヒトは教師付学習を行っていないので、我々の得たモデルはよりヒトの脳に近いことが期待される。

研究成果の概要（英文）：We can hear what others are saying even at a party where many people are talking. In this study, we attempted to create an artificial neural network that directs attention to one of the overlapping speech signals by autonomous learning, with the goal of creating a neural circuit model that can be compared to the human brain. We showed that an artificial neural circuit model (transformer) equipped with an attention mechanism could "recognize" environmental sounds as multiple objects and acquire a mechanism to "pay attention" to any one of them. This audio-transformer may be a promising neural model for solving the mystery of the "cocktail party problem".

研究分野：認知神経科学

キーワード：カクテルパーティー効果　深層学習　transformer

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

私たちは、大勢の人が話しているパーティーでも相手の言うことを聞き取れる。脳はどのようにして声を聴き分けているのか。「カクテルパーティ問題」として 50 年以上研究されてきたにもかかわらず、手法上の制約から「私たちの脳は聞きたい音声にトップダウンの注意を向けているらしい」ということしかわかっていない。本研究の開始当初には、音声認識の能力を獲得した多層人工神経回路とヒトの脳の活動を詳細に比較して、カクテルパーティ問題の謎を解くことを構想していた。

２．研究の目的

本研究の第一段階では、多層人工神経回路を教師付学習によって声の弁別を行うように訓練し、第二段階では、同時に複数の話者が発する音声刺激をヒト被験者に提示した際の脳活動を計測して、人工神経回路の応答と比較する計画であった。しかし、コロナ禍が障害となって第二段階のヒトを対象とするデータ取得が実施できなかった。また、本研究を開始した後に Google Research のグループが徹底した教師付学習とビデオ画像を併用する方法で混合音声を分離する人工神経回路を開発することに成功した[1]。そこで、本研究では第一段階に立ち戻り、教師付学習に頼らずに、「重なった音声信号から 1 つを選んで注意を向ける機構を自律的に人工神経回路に獲得させる」という一段階高度な問題に挑戦することとした。

３．研究の方法

画像への注意が定義できる人工神経回路(vision transformer, Dosovitskiy ら 2020)[2] に、音信号をスペクトログラムとして入力し、さらに自然な注意を自律的に獲得することが知られる学習法(ラベルなし自己蒸留法, Caron ら 2021)[3] を適用して、得られる情報量を最大化する自律学習を行わせた（図 1a)。YouTube 動画を使用した視聴覚大規模データセット VGG-Sound の音声データを入力音声として使用した。各音声を 8kHz にサンプリングした後、修正離散コサイン変換を行い、周波数成分の時間変化のスペクトログラム(帯域 0-4kHz, 10 秒間)を作成した。さらにスペクトログラムの一部をランダムに切り取り、transformer への入力とした。学習後の transformer と学習には用いていない 50 カテゴリーの環境音データベース ESC-50 からサンプルした音データを使って 2 種類のテストを行った。

(1) 環境音データベースの各カテゴリーから 8 個ずつサンプルした計 400 個のデータをそれぞれ transformer に入力して、最上層（第 4 層）の classification token の 384 個の人工神経の活動を記録した。384 次元のデータを tSNE 法により 2 次元に次元圧縮して表示した。
(2) 2 つの信号を混合して学習済みモデルに与えて、2 つの信号のいずれか一方に注意を向ける head があるかどうかを S/N 比の変化を用いて検討した（図 1b)。

４．研究成果

(1) Audio-transformer は聴覚世界を自律的に分節化した（図 2）

ラベルのない音信号を「聴いて」自律学習しただけの audio-transformer に 50 カテゴリー400 個の環境音データを入力したところ、最上層の classification token の活動は、おおよそカテゴリーに相当するようなクラスターを形成した（図２)。例えば、1 のボックスには“ hand saw ”、2 のボックスには ”helicopter ”、3 のボックスには ”crying baby ”とラベルされた音信号が集中していた。自律学習の際にはこれらの信号は用いていないし、学習信号のラベルも学習には一切用いていなかったことに注意されたい。つまり、自律的に学習した audio-transformer は音信号の分類に関する一切の事前の知識なしに、聴覚世界をヒトが付けたラベルと似た構造に分節化したのだ。

(2)Audio-transformer のヘッドは重なった音の 1 つだけに注意を向けた(図 3)

この audio-transformer は 2 つの音が重なっているときに、どちらか 1 つに「注意」を向けるだろうか。50 カテゴリーの中から 2 カテゴリーを選び、それぞれから 40 個のデータをサンプルして 1600 通りの 2 カテゴリーを重ねた信号を作成して、audio-transformer に入力してみた。研究に用いた Audio-transformer には音に対して注意を向ける ”ヘッド (head) ”と呼ばれる構造が 6 個ある。最上層における classification token の 6 個の head の中に、重なった音を分離して片方だけに注意を向けるような head はあるだろうか。図 2 でクラスターを示した ”helicopter ”の信号と ”crying baby ”を例に 6 個の head の挙動を見てみよう（図 3a)。crying baby の信号は、間歇的に「エーン、エーン」と泣いている。一方 helicopter は規則的に「ブルブルブル」と音を立てる。重ねた複雑な信号(mixed sound)の入力に対して、6 個の head(0-5)の注意の map を観察すると、head #1 と 2 は赤ん坊の泣き声に対して「注意」を向けているように見える。実際、1600 通りの重ね合わせに関して、各 head が注意を向けた音の成分だけを取り出して復元した信号が mixed sound に比べてどれだけ S/N 比を改善したかを計算すると（図 3b)、head 1 と 2 は mixed sound の中から赤ん坊の泣き声の S/N 比だけを 2-3dB 向上

させていることが確かめられた。特に head#1 は helicopter の音の S/N 比を 2db 程度下げているので、混合音の中から crying baby の泣き声を「聴き取ろうと」しているかのようである。ここで再度強調したいのは、audio-transformer は無理やり混合音を分離するように強制されたわけではない、という点である。ここが教師付学習を用いて音声分離に成功した幾多の研究グループとの決定的な相違点である。ただ、世界にあふれる音を聞き続けるだけで、重なった音を分離して「聴きとる」ような注意の head が transformer に獲得されたのだ。ヒトはカクテルパーティ効果を発揮するような訓練を受けたわけではない。生後の自律的な学習の過程を経て、いつの間にか音を分離する能力を獲得したのだ。その意味で、本研究の audio-transformer はヒトのカクテルパーティ効果の神経基盤のモデルとして、教師付学習で無理に作った人工神経回路よりも優れているものと期待される。この audio-transformer は「カクテルパーティ問題」の謎を解くための有力な神経モデルとなるだろう。

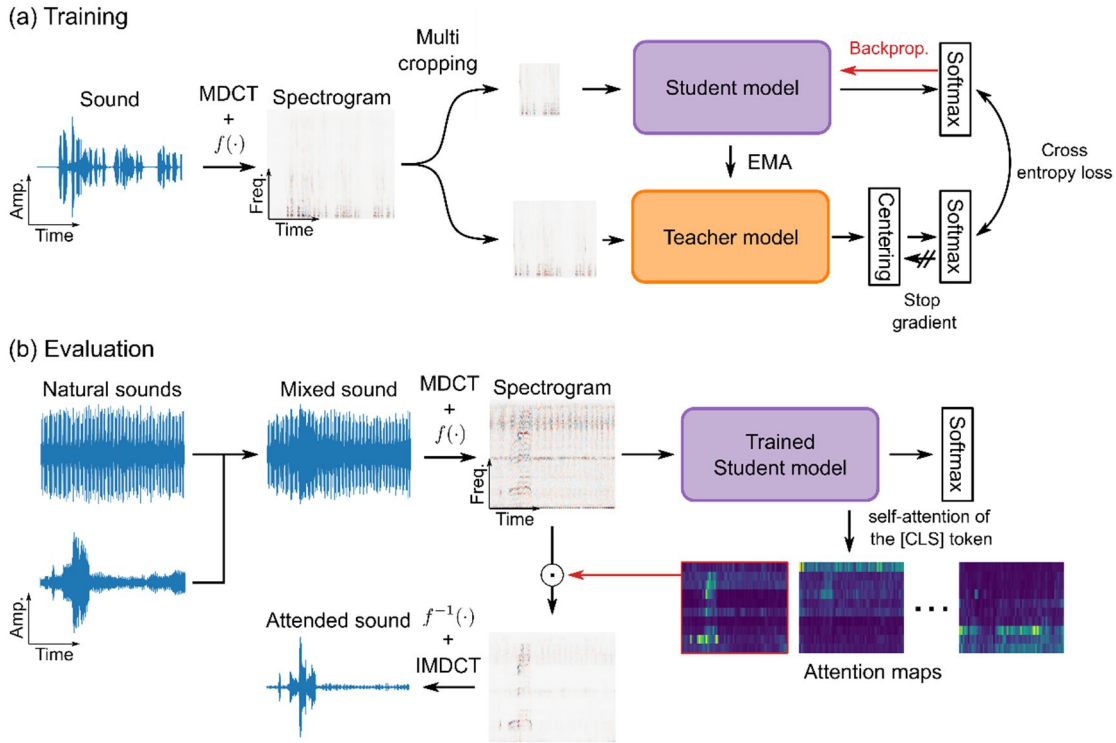

図 1　Transformer を使った音声自律学習法（a）と分離のテスト手法(b)

(a) Transformer は 4 層、6 ヘッド（6 個の異なる注意を獲得する）として、Caron ら(2021)の自己蒸留法を用いて情報量最大化に相当するラベルなし自律学習を行った。学習時には音声信号の混合は一切行っていない。

(b) 学習には用いていない 50 カテゴリーの環境音データベースからサンプルした 2 つの信号を混合して学習済みモデルに与えて、2 つの信号のいずれか一方に注意を向ける head があるかどうかを S/N 比の変化を用いて検討した。

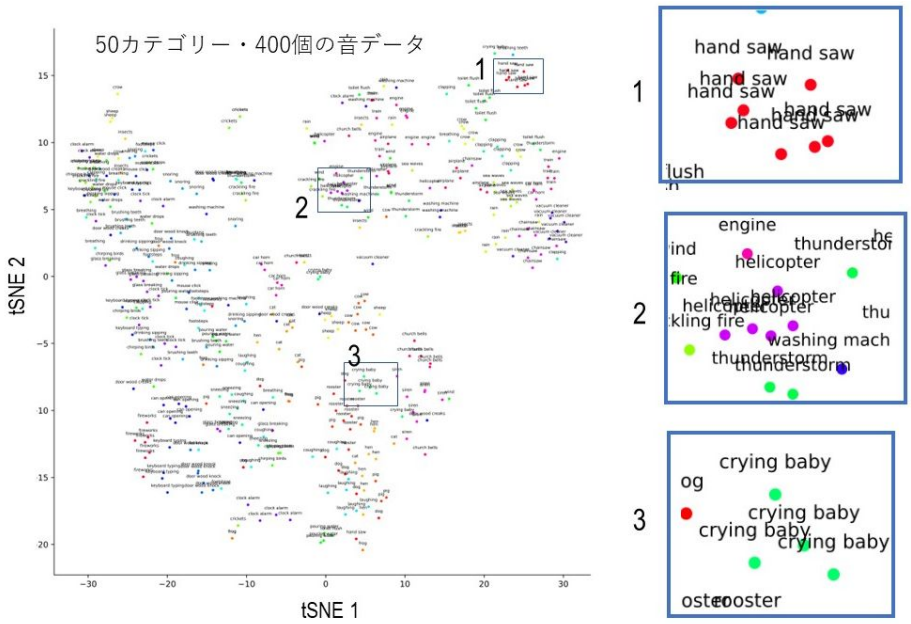

図 2 Audio-transformer は聴覚世界を自律的に分節化した

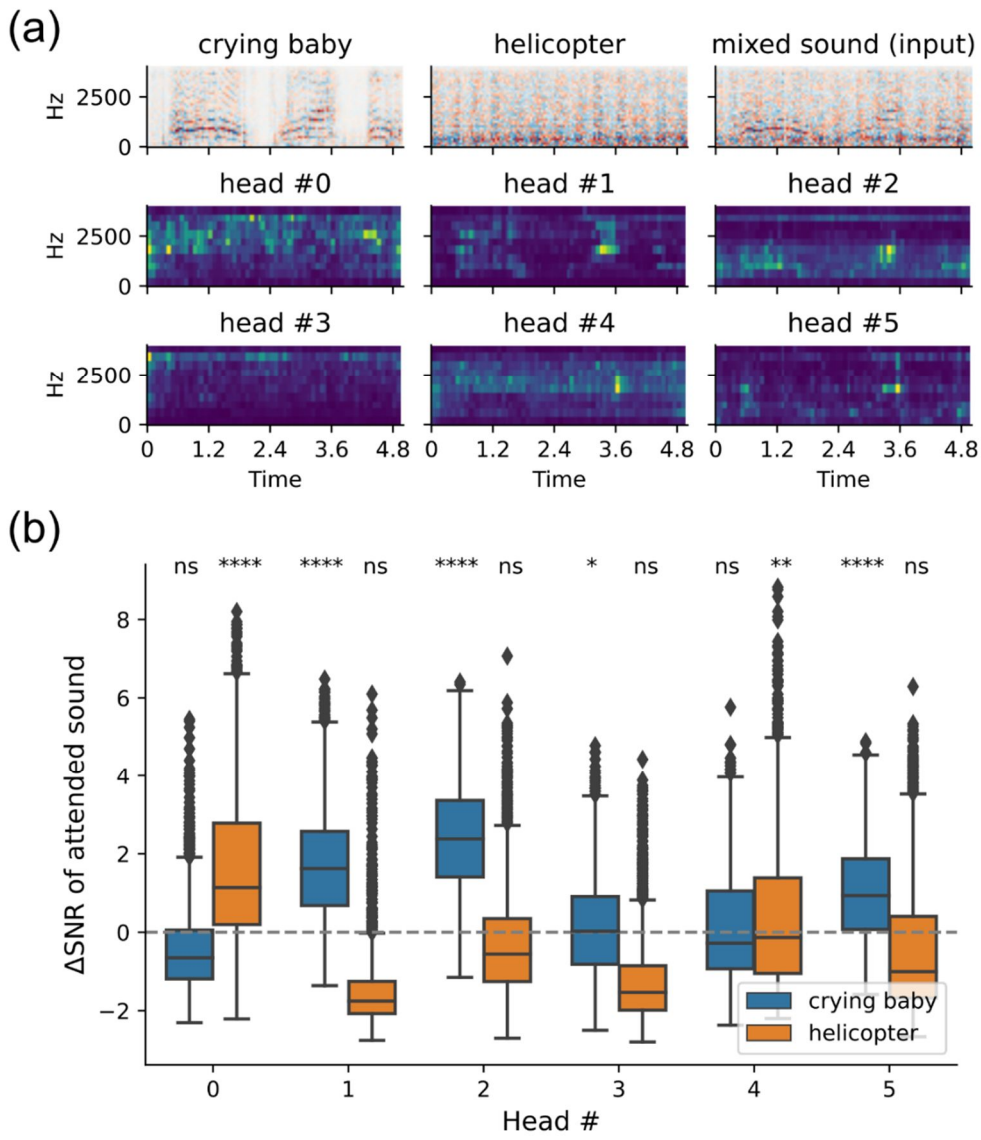

図 3　Audio-transformer は重なった音の一方だけに「注意」を向けた。

1. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., and Rubinstein, M. (2018). Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. Acm T Graphic *37*. 10.1145/3197517.3201357.
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Geigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929v2
3. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. arXiv:2104.14294v2 [cs.CV]

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 Takuto Yamamoto *; Shigeru Kitazawa
２．発表標題 Emergence of color constancy in an autoencoder with biologically plausible “batch normalization”
３．学会等名 第29回日本神経回路学会全国大会
４．発表年 2019年

１．発表者名 Akiyama O, Kitazawa S
２．発表標題 Emergence of visual receptive field remapping in a convolutional neural network for sensory prediction
３．学会等名 第41回日本神経科学学会大会
４．発表年 2018年

〔図書〕　計1件

１．著者名 北澤　茂	４．発行年 2020年
２．出版社 中外医学社	５．総ページ数 192
３．書名 医師・医学生のための人工知能入門	

〔産業財産権〕

〔その他〕

https://kitazawa-lab.jp/index.html

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	山本　拓人 (Yamamoto Takuto)	大阪大学・医学部 (14401)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関