音声刺激下の脳活動情報からのテキスト生成への取り組み

漆原 理乃 †
† お茶の水女子大学大学院

小林 一郎 ‡
‡ お茶の水女子大学基幹研究院

{g1420509, koba}@is.ocha.ac.jp

1 はじめに

近年, 脳神経活動の意味表象を捉える研究が盛んになっている. 本研究では, Functional Magnetic Resonance Imaging (fMRI) で観測した音声刺激下の脳活動データから, 人が脳内に想起した高次意味表象を言語として解読することを目指し, 深層学習を用いて, 音声刺激による脳活動データからその意味表象をテキストとして生成する手法を構築する. しかし, fMRI により観測する脳活動データは取得のためのコストが大きく, 大量の学習データを要する深層学習を十分に行うための大規模なデータ収集は困難である. そのため, 自動音声認識手法を援用することで少量データを効率的に活用する.

2 関連研究

近年, 脳神経活動の多点計測技術の発展と機械学習技術の高度化により, Huth ら [1] や Stansbury ら [2] の先行研究のように, ヒト脳内情報表現の定量理解や解読を目指す研究が盛んになっている. しかし, これらは動画像もしくは静止画像視聴下における脳神経活動を対象としており, 音声 (言語) を刺激とした脳神経活動を対象とした研究は少ない. そのため, 本研究では, 音声刺激を対象とした脳神経活動を解読することを目的とする. 解読手法構築にあたって, Matsuo ら [3] により, 深層学習を援用することで, 動画像視聴下における脳活動から認知内容の解読が実現できることが示されている. 本研究では, このような背景から深層学習を用いるが, その際, Matsuo ら [3] の少量の脳活動データの効率的な利活用方法を参考にし, 自動音声認識手法を援用することで, 音声刺激下の脳活動データから, その刺激となっていた音声のテキストを生成する手法を構築し, 脳内意味表象の解読を目指す.

3 提案手法

本提案手法は, 深層学習を用いて, 音声刺激を受けた脳活動データを入力として, その時に刺激となっていた音声のテキストを生成し, 人が頭の中で想起した言葉に対応する意味表象を言葉として解読することを目指す. しかし, fMRI により観測する脳活動データは取得のためのコストが大きく, 大量の学習データを要する深層学習を十分に行うための大規模なデータ収集は困難である. そのため, Encoder-Decoder Network に基づく自動音声認識手法を援用することで少量データを効率的に活用する. 具体的には自動音声認識の Encoder から得られる中間表現に, 脳活動データを回帰させ, その結果を自動音声認識の Decoder に入力することで, テキスト生成を行う. 図 1 に本提案手法の概要図, 3.1 節に提案手法の処理の流れを示す.

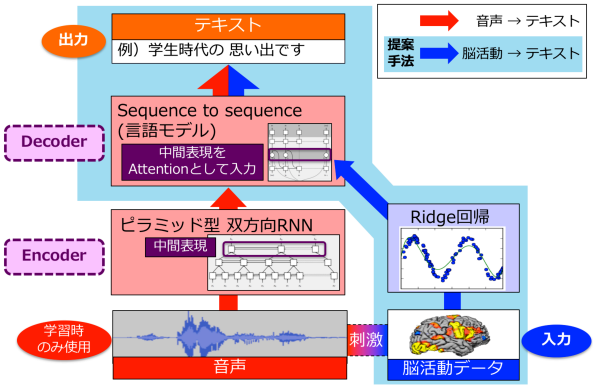

図 1: 本研究の概要図

3.1 提案手法の処理の流れ

実行時の処理は以下のようになる.

step 1. 自動音声認識

step 1-1. Encoder: 音声中間表現の抽出
自動音声認識の Encoder を用いて, 音声から音声中間表現を抽出.

step 1-2. Decoder: テキスト生成
step1-1. において抽出された音声中間表現を, 自動音声認識の Decoder に入力し, テキストを生成.

step 2. 脳活動データの特徴量推定
脳活動データとその刺激である音声の中間表現 (step1-1 の出力) との対応関係を学習した Ridge 回帰により, 脳活動データから対応する中間表現を推定.

step 3. 脳活動データの特徴量からテキスト生成
step1-2. で学習済みの自動音声認識の Decoder を用いて, step2. で計算された脳活動データの特徴量を入力として, テキストを生成.

3.2 自動音声認識

本手法の基盤として, Chan ら [4] によって提案された, Encoder-Decoder Network を用いた自動音声認識モデルである, Listen, Attend and Spell (LAS) を用いる. LAS は Encoder として, ピラミッド型の双方向 Recurrent Neural Network (RNN) を用い, 入力音声を中間表現に変換する. Decoder としては, Attention 付きの単方向 RNN を用い, Encoder で得た中間表現を Attention として入力し, 通常の言語モデル同様に次にくるであろう単語, もしくは文字を 1 つずつ生成する. Decoder で Attention を用いることで, 中間表現の系列と出力系列との対応関係も学習できる.

3.3 脳活動データの特徴量推定

音声刺激を受けた被験者の脳活動データを入力とし, その時の刺激となっている音声の LAS の Encoder によって抽出された中間表現を予測するために, Ridge 回帰を用いる. 学習には脳活動データと被験者がその時に聴いている音声を用いる. また, fMRI は脳活動を記録する際にタイムラグがあり, 今回は 4 秒と仮定して, Ridge 回帰のモデルを構築する.

4 実験

3.1 節に示す提案手法の処理の流れに沿って行った 3 つの実験を以下に示す.

4.1 実験 1：自動音声認識

4.1.1 実験設定

システムの実装は, 深層学習のフレームワーク TensorFlow を用いたコード[1]を使用した. 学習のためのデータセットとして,「日本語話し言葉コーパス」(CSJ) 中の 3,254 本の講演データを使用した. CSJ で設定されている評価セット 1 から 3 と脳活動データの刺激として使用された音声は除外した. 音声の前処理として, 転記基本単位 (IPU) で分割し, フレームサイズ 25ms, フレームシフト 10ms でフレームごとに MFCC 特徴量を取得し入力とした. 出力は発音形 (PhoneticTranscription) を用い, 発音されている文字に start-of-sentence (sos) と end-of-sentence (eos), バッチサイズ中の系列長に合わせるための padding(pad) を追加し, 84 次元とした. 学習に関する詳細設定は表 1 に示す. 評価は CSJ の評価セット 1 の 10 本の講演データを用いて行った.

4.1.2 実験結果

epoch ごとに学習用データの Loss を記録し, その減少により学習の進度を確認し, 収束するまで学習を行った. 図 2 には, 先行研究 [4] と同様に取得した Decoder における Attention の値の可視化結果を示す. また, 評価セット 1 で評価を行い, 生成したテキストの一部と, 生成テキストと正解テキストとをそれぞれ比較し編集距離のマクロ平均を計算したものを表 2 に示す.

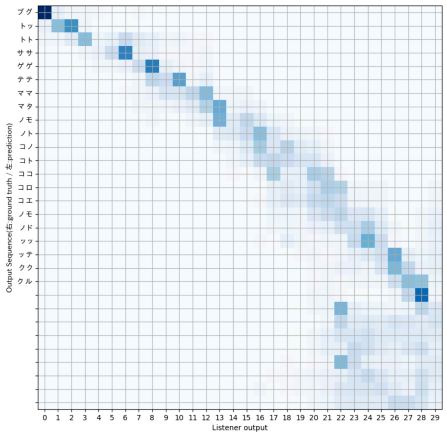

図 2: 実験 1 attention の可視化

4.1.3 考察

図 2 より, Attention がうまく機能している, つまり中間表現の各要素が, 出力系列のどの部分に対応する

[1] https://github.com/thomasschmied/Speech_Recognition_with_Tensorflow

表 1: 各パラメータ設定（詳細）

	自動音声認識	脳活動データの特徴量抽出
train データ	日本語話し言葉コーパス (CSJ)	音声刺激による脳活動データ
学習量	919,118 sample × 70 epochs	100,938 sample (9,841sample を元に線形補間し増強)
アルゴリズム	Adam	Ridge 回帰
学習に関する パラメータ	学習率: 0.00001	L2 正則化項: 1.0
隠れ層次元	入力：494 次元 Encoder 2 層：全て 450 次元 Decoder 2 層：全て 450 次元 Embedding：10 次元 出力：84 次元	Attention となる中間表現: 62,552 次元 - 900 次元 RNN の隠れ層: 62,552 次元 - 1800 次元
誤差関数	交差エントロピー	
その他	Cyclic Learning Rate Max 学習率：0.00003 Step size: 700	

表 2: 評価セット 1 における音声認識実験結果

正解テキスト	生成テキスト
コンカイノ	コンカイノ
トユーブンデハ	トユユブブンンワワ
コノヨーニ マチガウカノーセーガ アリマスノデ	コノヨーニニ チチーカカノノーー ガガアリママスデデ
編集距離のマクロ平均: 11.3 (平均文字数: 21.8)	

かを学習することができていることがわかる．しかし，表 2 より，短いテキストは生成に成功しているが，長いテキストは正確な生成が困難であるといえる．

4.2 実験 2：脳活動データの特徴量推定

4.2.1 実験設定

脳活動と音声特徴量の対応関係を学習するためのデータセットとして，日本語話し言葉コーパス (CSJ) の 16 本を 1 人の被験者に聴かせた時の血中酸素濃度依存性信号 (BOLD 信号;Blood Oxygenation Level Dependent Signal) を fMRI を用いて 1 秒ごとに記録した脳活動データ，および fMRI のデータ収集と同期させた CSJ の音声を使用する．立体撮像 96 × 96 × 72 ボクセルのうち皮質に相当する 62,552 次元分のデータ列を用い，その時聴いている音声から LAS の Encoder であるピラミッド型の双方向 RNN により得られた 900 次元の音声中間表現との対応と，1800 次元の RNN の隠れ層 (Long short-term memory(LSTM) の隠れ層，1,2 層目のセル c,h で共に 450 次元により合計 1800 次元．Decoder では，この隠れ層を初期値として予測を開始する) との対応を学習した．脳活動データは，1 秒ごとに取得されているため，系列長を音声中間表現の系列数に合わせるように線形補間を行い，最大値 1，最小値 0 に正規化を行ったものを入力として使用した．train 用データは 14 本，test 用データは 2 本の講演データを聴いた脳活動データとした．学習には Ridge 回帰を用い，その詳細設定は表 1 に示す．

4.2.2 実験結果

test 用データの脳活動データから Ridge 回帰を行い，その刺激となっていた音声の中間表現を予測した結果の一部を図 3 に示す．

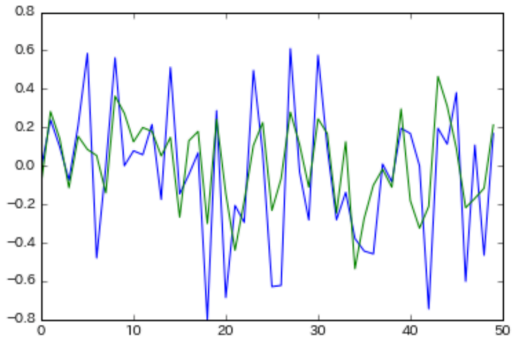

図 3: 実験 2 脳活動データの Ridge 回帰結果．900 次元の特徴量の中で，50 次元を抽出した結果を可視化．青線 - 刺激となっていた音声の中間表現 (正解データ)．緑線 - 脳活動データから Ridge 回帰を行った結果．

4.2.3 考察

脳活動データに対して Ridge 回帰を行うことで, 音声中間表現の振幅の増減の位置は予測可能であることが確認できる. しかし, 具体的な数値は合致していない. これに関しては, fMRI が持つ脳活動データの記録する際のタイムラグも関係していると考えられる. 今回はそのタイムラグを 4 秒と仮定したが, 今後は数秒ずつずらした複数時点の脳活動データを説明変数として取り入れ, Ridge 回帰を行いたい.

4.3 実験 3：脳活動からのテキスト生成

実験 1 で学習した自動音声認識モデル LAS の Decoder を使用して, 実験 2 で取得した脳活動データの特徴量を入力として, テキスト生成を行う.

4.3.1 実験設定

実験 2 の学習において未使用の test 用の 2 本の講演データを聴いた脳活動データを用いて, Ridge 回帰を行い, 900 次元の中間表現と RNN の隠れ層 1800 次元を取得し, LAS の Decoder に前者の中間表現は Attention として入力し, 後者は RNN の隠れ層の初期値として使用した.

4.3.2 実験結果

脳活動データから生成したテキストと音声から生成したテキスト, 正解テキストをそれぞれ比較し編集距離のマクロ平均を計算したものを表 3 に示す. また, 生成したテキストの一部を表 4 に示す.

表 3: テキスト生成実験結果 (編集距離のマクロ平均)

脳活動からの生成	10.1
音声からの生成	3.2
平均文字数	6.6

表 4: テキスト生成実験結果 (生成例)

正解	脳活動から生成	音声から生成
テーマワ	エーーー	テーマワ
アルト	デテ	アルト
マダ	イイ	マダ
ヤッパリ	ノノ	ヤッパリ

4.3.3 考察

表 3 より, 脳活動データから生成したテキストの編集距離のマクロ平均が test 用データの平均文字数よりも大きくなっていること, 表 4 より, 実際の生成テキスト例も正解もしくは音声から生成したテキストと一致しておらず, 生成がうまくいっていないことがわかる. 4.2.3 節で述べた, 脳活動データの特徴量の推定値の誤差が影響していると考えられる.

5 おわりに

本稿では, 自動音声認識モデル LAS を援用し, 音声刺激を受けた被験者の脳活動データから聴いている音声をテキストとして出力する手法を提案した. 今後の課題として, 実験 2 における fMRI が持つ脳活動データの記録する際のタイムラグも考慮に入れた上で, 脳活動データから特徴量を推定する手法を見直し, 実験結果のさらなる評価・分析を実施し, 生成するテキストの精度を向上させていきたい.

謝辞

本研究を進めるにあたっては, 情報通信研究機構脳情報通信融合研究センターの西本伸志氏にデータを提供して頂いた. ここに謝意を表する.

参考文献

[1] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

[2] Dustin E Stansbury, Thomas Naselaris, and Jack L Gallant. Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, 79(5):1025–1034, 2013.

[3] Eri Matsuo, Ichiro Kobayashi, Shinji Nishimoto, Satoshi Nishida, and Hideki Asoh. Describing semantic representations of brain activity evoked by visual stimuli. *arXiv preprint arXiv:1802.02210*, 2018.

[4] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 4960–4964. IEEE, 2016.