音声刺激下の脳活動情報による音声特徴量の推定への取り組み

漆原理乃 † 山口裕人 ‡¶ 中井智也 ‡¶ 西本伸志 ‡¶ 小林一郎 †
† お茶の水女子大学 ‡ 情報通信研究機構 ¶ 大阪大学
†{g1420509, koba}@is.ocha.ac.jp,
‡{hyamaguchi, nakai.tomoya, nishimoto}@nict.go.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．また，脳の特定の領域のみを入力した実験により，深層学習モデル中間層における領域ごとの音声特徴量（以降、深層学習の中間層に表現される音声特徴量を「音声中間表現」と呼ぶ）の推定精度の比較を行う．

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，ヒト脳内言語情報処理機構の定量理解や解読を目指す研究が盛んになっている [1,2]．ヒトの脳に電磁波を当てることにより血中酸素濃度依存性信号 (BOLD 信号) を計測する非侵襲的手法の fMRI(functional magnetic resonance imaging) を用いることで，より具体的な脳領域と脳内意味表象との対応関係を解明することができ，音声刺激下の脳活動データを記録し，皮質全体と脳内意味表象の対応関係を捉えた研究 [3] なども報告されている．しかし，音声刺激下の fMRI 脳神経活動から直接的にテキストを生成する研究は未だ報告はない．そのため，本研究では fMRI から取得した音声刺激下の脳神経活動からの言語解読を行う．それにより，ヒト脳内における言語情報処理機構の探究し，将来，マシン・ブレインインタラクションを実現するための一歩となることを目指す．また，ヒトの脳機能を模倣した深層学習分野におけるモデル構築研究への知見も得ることも目指す．解読手法構築にあたって，Encoder-Decoder Network を用いた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言語解読が実現できることが示されている [4]．本研究では，Matsuo ら [4] の手法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することで，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．図 1 に提案手法のモデルと処理の流れを示す．

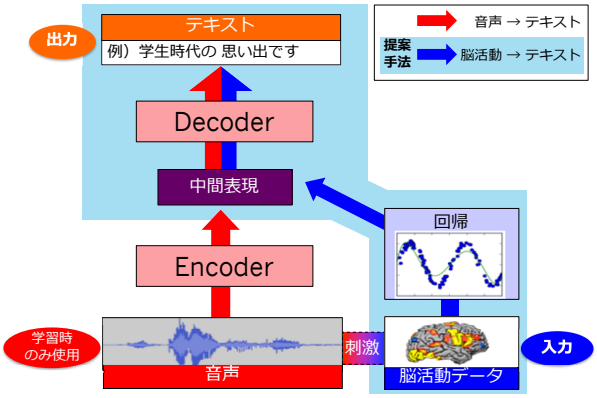

図 1: 本研究の概要図

step 1. 自動音声認識
Encoder-Decoder Network を用いた自動音声認識モデルである，Hybrid CTC/Attention Architecture [5] を使用して音声から対応するテキストを生成する．

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する．

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，テキストを生成する．

step 2. 脳活動データからの刺激音声中間表現推定
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回帰モデルにより，新規の脳活動データから対応する音声中間表現を推定する．回帰モデルとして，本研究では Ridge 回帰とニューラルネットワーク (NN) を用いる．

step 3. 推定した音声中間表現からテキスト生成
step1-2. で学習済みの自動音声認識の Decoder を用いて，新規の脳活動データから step2. で計算された音声中間表現を入力として，テキストを生成する．

3.1 自動音声認識

本手法の基盤として，Encoder-Decoder Network を用いた自動音声認識モデルである，Hybrid CTC/Attention Architecture [5] を用いる．このモデルは，Encoder として双方向 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 付きの系列変換モデル (sequence-to-sequence) に Connectionist Temporal Classification(CTC) を組み合わせたモデルを使用し，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．

3.2 脳活動データによる音声中間表現推定

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中間表現を予測するために，Ridge 回帰とニューラルネットワーク (NN) 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 4，5，6 秒と仮定し，それらを説明変数とする回帰のモデルを構築する．

4 実験

3 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

深層学習を用いた自動音声認識ツールキットである ESPnet[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) 中の講演データを使用し，CSJ で設定されている評価セット 1 から 3，および，脳活動データの刺激として使用した音声以外を訓練データとした．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 Mel-scale filterbank 特徴量を計算し，それにピッチを加えた 83 次元の音声特徴量を入力とした．モデルの詳細パラメータ設定を表 1 に示す．

[1] https://github.com/espnet/espnet

4.1.2 実験結果

生成，正解テキストを比較し，文字間違い率 (CER; Character Error Rate) の講演データごとのマクロ平均を計算したものを表 2 に示す．ESPnet を用いた実験により生成したテキストの一部を表 3 に示す．

4.1.3 考察

表 2 の文字間違い率の比較においては，評価セット 1，2，3 ともに先行研究 [5] と同様の数値となった．脳活動データの刺激として使用した音声は評価セットと比較して文字間違い率が低く良い結果となっている．また，表 3 の実際の生成テキストを見ると，「程」と「頃」の間違いはあるが，正解テキストの「御御」という 2 文字同じ音の連続している言いよどみの箇所が，生成テキストでは「御」のみの 1 文字になっており，言語モデルの有効性が確認できる．また，長い系列長においてもうまく予測できていることが確認できる．

4.2 実験 2: 脳活動からの音声中間表現推定

4.2.1 実験設定

脳活動と音声中間表現の対応関係を学習するためのデータセットとして，CSJ の 16 本を 1 人の被験者に聴かせた時の BOLD 信号を 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 ボクセルのうち大脳皮質に相当するデータ列を用いた．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．test 用データの 2 本は，1ヶ月の期間をあけて 2 回繰り返し収録された脳活動データを用い，それぞれの平均値を使用する．推定の手順としては，まず刺激となっていた音声を 1 秒単位で分割し，1 秒間の音声データを ESPnet の Encoder に入力し，25,600 次元 (1024 次元

表 1: パラメータ設定

	自動音声認識 (音声→テキスト)	脳活動による音声中間表現推定 (脳活動→音声中間表現)	
		Ridge	NN
train データ	日本語話し言葉コーパス (CSJ)	音声刺激下の脳活動データ	
学習量	919,118 sample × 8 epochs	9,841sample	9,841sample × 20,000 epochs
アルゴリズム	AdaDelta	Ridge 回帰	SGD
入力次元	83 次元	脳領域ごとに異なる 表 4 参照	
隠れ層次元	Encoder 4 層：全て 1024 次元 Decoder 1 層：全て 1024 次元		10,000 次元
出力次元	3260 次元	25,600 次元	
誤差関数	CTC		平均二乗誤差
その他	CTC:Attention = 1:1	正則化項[2] $\lambda = 1.0$	

表 2: 音声認識実験結果 (CER; 文字間違い率)

	評価セット 1	評価セット 2	評価セット 3	脳活動刺激音声
CER	6.6	5.1	6.0	4.7

表 3: 音声認識実験結果 (生成テキスト例)

正解テキスト	生成テキスト
どうぞ御御指導の程 よろしくお願いいたします	どうぞ御指導の頃 よろしくお願いいたします
続いて実験方法について 説明いたします	続いて実験方法について 説明いたします
評価実験はこのような 建物空間で行いました	評価実験はこのような 建物空間で行いました

×25 次元の系列) の音声中間表現を取得した．その後，ある音声を聴いた時の脳活動データから，上述の音声中間表現を予測するような Ridge 回帰と NN の 2 種類の回帰モデルを構築した．fMRI は脳活動を記録する際にタイムラグがあるため，今回はそれを 4，5，6 秒の複数時点としたものをそれぞれ説明変数とする，回帰のモデルを構築し，比較した． (大脳皮質を用いた NN 回帰のみは高次元のため 4 秒のみを説明変数とした．) また，脳領域における推定精度の比較のため，2 つの方法を用いて fMRI で取得したボクセルの絞り込みを行った．1 つ目は，特定の関心領域 (Region of interest:ROI) を抽出する方法である．具体的には，音声処理を行なっていると言われる，前頭葉，頭頂葉，側頭葉の 3 つの部分を使用した．2 つ目は，2 回繰り返し取得した test 用データ 1 本分を用い、2 回の試行で計測された脳活動の相関をとり、相関係数の高い上位 N 個のボクセルを抽出し使用した．脳活動データは，平均 0，分散 1 に正規化を行ったものを入力として使用した．学習の詳細設定は表 1 に示す． 大脳皮質全体を用いた実験では学習時に使用した訓練データ，使用していない評価データのそれぞれを用い，それ以外の脳領域の実験においては評価データのみを用いた．

4.2.2 実験結果

脳活動データから音声中間表現を推定し，その推定結果とその正解となる刺激音声の中間表現との相関係数を計算した結果を表 4 に示す．また，大脳皮質全体の脳活動データによって学習した Ridge 回帰モデルから予測した音声中間表現を用いて，2 種類の類似音声検索を行った．1 つ目の方法は，脳活動データから予測した音声中間表現の中からコサイン類似度の高いもの同士を検索する方法であり，その結果を表 5 に示す．2 つ目の方法は，実際の音声を Encoder に入力して計算した音声中間表現の中から，脳活動データから予測した音声中間表現に近いものをコサイン類似度によって検索する方法であり，その結果を表 6 に示す．表 5，表 6 では，類似音声検索において訓練，評価データのそれぞれコサイン類似度の高かった上位 2 件のサンプルを示す．

4.2.3 考察

表 4 から，評価データを比較すると，Ridge 回帰モデルにおいては，特定の脳領域より大脳皮質全体を使用する方が相関係数が高いことがわかる．NN 回帰モデルにおいては，特定の脳領域のみに絞り込む方が相関係数が高いことがわかる．後者は，低次元にする方が学習が容易になるためと考えられる．また，Ridge 回帰の大脳皮質の訓練および評価データの比較においては，後者においては高い相関係数を得ることができ

[2] 1.0 から 10^8 を試行した結果 1.0 を採用した

表 4: 脳領域ごとの音声中間表現推定結果

脳領域		次元	Ridge	NN
大脳皮質 (訓練)		187,656	0.99	0.23
大脳皮質 (評価)		187,656	0.36	0.21
ROI	前頭葉	33,279	0.26	0.24
	頭頂葉	26,886	0.22	0.25
	側頭葉	30,735	0.25	0.24
相関の高い上位 N 個のボクセル	N=1,000	3,000	0.26	0.29
	N=5,000	15,000	0.14	0.26
	N=10,000	30,000	0.23	0.23
	N=15,000	45,000	0.27	0.22
	N=20,000	60,000	0.29	0.21

表 5: 類似音声検索 (脳から推定した音声中間表現の中から検索)

	検索元	検索結果	cos 類似度
訓練	おー	おー	0.83
訓練	ーなんですけれども	量なんですけれども	0.82
評価	いう	思い出です	0.87
評価	まー	本当にあのー	0.87

ておらず，これはデータ数が少ないことに起因していると考えられる．類似音声検索実験においては，訓練データは表 5 では類似音声，6 では同じ音声を検索することができている．しかし，評価データにおいては訓練データのように適切に検索することができていないが，表 6 においては，テキストは異なるが，刺激音声のはじめの 1 文字の母音や伸ばし方や無音の箇所が類似している音声は検索することができた．

4.3 実験 3: 脳活動からのテキスト生成

4.3.1 実験設定

step2 における Ridge 回帰実験結果の大脳皮質の訓練，評価データから生成した音声中間表現を ESPnet の Decoder に入力することでテキストを生成した．

4.3.2 実験結果

テキスト生成結果を表 7 に示す．

4.3.3 考察

訓練データからは正解と同じテキストを生成することができたが，評価データで生成できていないのは，データ数が少なく，step2 のモデルの精度が低いためと考えられる．

表 6: 類似音声検索 (実際の音声中間表現の中から検索)

	検索元	検索結果	cos 類似度
訓練	検出はこう	検出はこう	1
訓練	さするとしょ	さするとしょ	1
評価	でー	えーま本当にでも	0.70
評価	ん	まー実際に	0.69

表 7: 大脳皮質を用いたテキスト生成結果

データセットの種類	正解テキスト	生成テキスト
訓練データ	だんだんと興味が	だんだんと興味が
訓練データ	できました	できました
評価データ	あの楽しい	映画
評価データ	ですがも	です

5 おわりに

本稿では，自動音声認識モデル Hybrid CTC/Attention アーキテクチュアを援用し，音声刺激を受けた被験者の脳活動データから刺激となっている音声をテキストとして出力する手法を提案した．実験において，評価データにおいては十分な精度となる結果は得られなかったが，訓練データにおいては言語解読が可能であることを確認した．今後の課題として，訓練データにおいてある程度の精度が観測されていることからデータ数を増やすなど推定精度向上のための工夫をするつもりである．

参考文献

[1] Huth, A. G., Nishimoto, S., Vu, A. T., Gallant, J. L.: 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] Stansbury, D. E., Naselaris, T., Gallant, J. L.: Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex, Neuron 79, pp.1025-1034, September 4, 2013, Elsevier Inc (2013).

[3] Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453. (2016)

[4] Matsuo, E., Kobayashi, I., Nishimoto, S., Nishida, S., and Asoh, H. Describing Semantic Representations of Brain Activity Evoked by Visual Stimuli. In arXiv preprint arXiv:1802.02210 (2018).

[5] S. Watanabe, T. Hori, S. Kim, J. Hershey, T. Hayashi. " Hybrid CTC/attention architecture for end-to-end speech recognition." IEEE Journal of Selected Topics in Signal Processing 11.8 (2017): 1240-1253.