想起音声識別のための変分自己符号化器に基づく表現学習*

◎矢野　肇, 高島遼一, 滝口哲也 (神戸大), 中川誠司 (千葉大)

1　はじめに

ブレイン・コンピュータ・インターフェイス（brain computer interface: BCI）は発話や身振りが困難な身体障害者のための意思伝達手段として研究開発が盛んに行われている．近年では，自由度の高いコミュニケーションが可能な，頭の中で想起した音声を脳活動から読み取る BCI の研究が注目されている [1].

我々はこれまで，音声想起時の脳磁図（magneto-encephalography: MEG）を調査し，機械学習を用いて想起音声の識別を行ってきた [2, 3, 4]．先行研究 [3, 4] では，複雑な識別関数を学習可能な畳み込みニューラルネットワークが個人ごとに学習されたが，学習パラメータ数はサポートベクターマシン等の機械学習手法と比べると圧倒的に多く，実用に耐えうる高い精度を持つ識別器を個人ごとの少ない MEG データを用いて安定して学習することは容易ではない．

近年，大規模なニューラルネットワークを用いた表現学習の研究が盛んに行われており，教師ラベルのない大量のテキスト，画像，音声のみで学習されたネットワークの一部を分類等の下位のタスクに用いることで高い性能が得られることが知られている [5, 6, 7]. 音声等と異なり，脳活動データは大量に用意するのが難しく，大規模なニューラルネットワークを用いた表現学習は困難だが，タスクや条件の異なるデータを集め，できるだけ多くの学習データを用いて表現学習をすることで，少量のデータしか得られない想起音声の分類精度を向上できる可能性がある．表現学習の手法の中でも，変分自己符号化器（variational autoencoder: VAE）[8] は，比較的小さなネットワーク構造でも利用可能で，大規模なデータセットを必ずしも必要としない．また，すでに VAE は運動想起の脳波の分類や [9]，被験者非依存の脳波表現の学習 [10] に用いられている．

本研究では，個人ごとに少量のデータしか利用できない条件下で想起音声の分類精度を向上させるため，VAE を用いて複数人の音声想起時 MEG の特徴表現を学習し，想起音声の分類モデルの学習に利用した．具体的には，学習した VAE のエンコーダーの重みを転移させて，新たな個人（VAE の学習時にデータを用いていない人）の想起時の MEG を分類する畳み込みニューラルネットワークの学習を行った．さらに，Conditional VAE (CVAE) [11] を用いて新たな被験者の想起時の MEG データを生成し，拡張したデータセットを用いて想起音声の分類モデルを学習した．

2　方法

2.1　脳磁図データ

先行研究 [2] で収録された音声想起時の脳磁図データを用いた．このデータは 8 名の被験者（男性 7 名，女性 1 名，20–40 歳）の脳磁図データからなる．被験者は，3 種類の日本語単語（“ あまぐも ”，“ いべんと ”，“ うらない ”）のうちの 1 つを 2 回聴取した後，音声を聞いた通りに想起する試行を繰り返し行う．音声の持続時間は 800 ms で，呈示間隔は 500 ms であった．この試行は単語ごとに少なくとも 100 回行われた．試行中の脳磁図は被験者の頭を覆うように配置された 122 個のセンサで計測された．

計測された MEG の前処理として，サンプリング周波数が 200 Hz となるようにダウンサンプリングし，1Hz 以下の低周波成分，1000 fT/cm を超える変動，及び眼球運動に由来するアーティファクトを除去した．頻繁に異常な信号が観測された MEG センサの信号は取り除き，他の正常なセンサの信号を元に補間した．1 回目の音声呈示及び想起のタイミングを基準に −100–900 ms の信号を音声聴取時及び想起時の MEG として切り出した．最後に，切り出された MEG 波形を全ての時間及びセンサで共通な平均値と標準偏差で標準化した．

2.2　使用したネットワーク

本研究で用いたニューラルネットワークの概要を Fig. 1 に示す．

2.2.1　EEGNet

音声想起時の MEG の分類器には EEGNet [12] を用いた．EEGNet は一般的な畳み込みの代わりに depthwise separable convolution を用いた軽量なネットワークであり，学習データ数が少量でも，比較的精度が高いモデルを学習できる．EEGNet は時間方向の畳み込み層，センサ方向の depthwise convolution 層，時間方向の separable convolution 層，及び全結合層から成る．本研究では，EEGNet の構造を少し変更し，separable convolution 層を 2 層にして用いた．

---

*Representation learning based on variational autoencoders for imagined speech classification. by YANO, Hajime, TAKASHIMA, Ryôichi, TAKIGUCHI Tetsuya (Kobe Univ.) and NAKAGAWA, Seiji (Chiba Univ.).

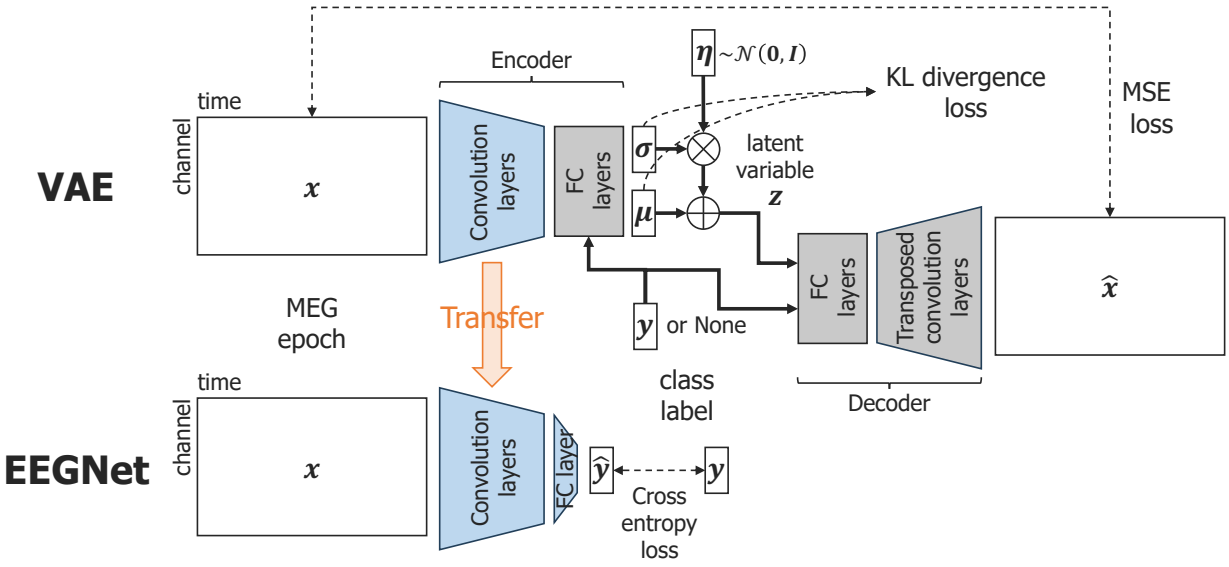

Fig. 1　Overview of the network architectures

2.2.2　VAE

VAE は，入力データ $\boldsymbol{x}$ から潜在変数 $\boldsymbol{z}$ の近似事後確率分布 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ の平均 $\boldsymbol{\mu}$ 及び標準偏差 $\boldsymbol{\sigma}$ を出力するエンコーダーと，q_ϕ からサンプリングした潜在変数を入力として元のデータを再構成するデコーダーからなる [8]．エンコーダーとデコーダーはニューラルネットワークで構成され，それぞれのパラメータを $\boldsymbol{\phi}$，$\boldsymbol{\theta}$ で表す．最終的に VAE を学習するための損失関数は次のようになる．

$$\mathcal{L}_{\mathrm{VAE}} = \mathbb{E}_{q_\phi}[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] + \mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))] \quad (1)$$

第 1 項は復元されたデータと元のデータの再構成損失を表し，本研究では平均 2 乗損失を用いた．第 2 項は潜在変数の近似事後分布 q_ϕ と事前分布 $p(\boldsymbol{z})$ の間の KL ダイバージェンスであり，$p(\boldsymbol{z})$ には多次元標準正規分布を用いた．

本研究では，VAE のエンコーダーには EEGNet の畳み込み層に 2 層の全結合層を接続したものを用いた．VAE のデコーダーには畳み込み演算の代わりに転置畳み込み演算を用い，エンコーダーと対称的な構造にした．潜在変数の次元数は 32 とした．

2.2.3　CVAE

CVAE は，VAE の近似事後分布 q_ϕ とデコーダの分布 $p_{\boldsymbol{\theta}}$ を，さらにデータ $\boldsymbol{x}$ に対応するラベル y で条件付けたものである [11]．このため CVAE の構造は VAE のエンコーダー及びデコーダーにクラスラベル入力を追加した構造になる．本研究では，クラスラベルを one-hot 符号化の後，4 または 8 次元ベクトルに線形変換し，エンコーダーの第 2 全結合層の入力及びデコーダーの入力に結合した．

2.3　音声想起時の MEG 分類

音声想起時の MEG を分類する EEGNet を被験者ごとに学習した．以降では，このモデルを被験者内モデルと呼ぶ．学習された被験者内モデルを 10 分割交差検証によって評価した．交差検証の各分割に含まれる試行は時間的に連続するようにした．評価データ以外のデータは，時間的に最初の 80%を学習データに，残りの 20%を検証データに分割した．

次に，EEGNet を複数の被験者のデータを用いて学習した．以降では，このモデルを被験者間モデルと呼ぶ．被験者間モデルでは，1 人の被験者のデータを評価データとし，残りの被験者のデータをモデルの学習及び検証に用いた．学習及び検証に用いるデータは，被験者ごとに時間的に最初の 80%を学習データ，残りの 20%を検証データとした．

学習されたモデルの評価指標には macro F1 score を用いた．この指標はクラスごとに算出した F1 score を平均したものである．EEGNet は最大 50 エポック学習させ，検証データセットに対し macro F1 score が最大となるモデルを選択した．

2.4　転移学習

まず，複数被験者の音声想起時の MEG を用いて VAE の学習を行った．学習及び検証に用いたデータは，EEGNet の被験者間モデルと同様である．また，1 回目の聴取と想起時の MEG を用いた VAE の学習も行った．学習データ数の増加によって，より良い特徴表現の学習が期待される．VAE は最大 500 エポック学習を行い，検証データに対する損失関数が最小となるエポックのモデルを選択した．

次に，学習された VAE の畳み込み層を EEGNet の畳み込み層にコピーし，転移学習を行った．畳み込

み層のコピー後，全結合層のみ，もしくは，ネットワーク全体のファインチューニングを行った．転移学習したモデルの学習，検証及び評価に用いるデータは VAE の学習に用いられていない被験者のデータを用い，被験者内モデルと同様に分割した．

転移学習で転移させる層は他の被験者のデータで学習されたものであるため，その層から得られる特徴表現が評価対象の被験者に適したものとは乖離している可能性がある．このため，転移学習する前に評価対象となっている被験者のデータを用いて VAE を適応させた．これにより，転移学習後のモデルの精度向上が期待される．

2.5　データ拡張

まず，VAE と同様に，複数被験者の音声想起時の MEG を用いて CVAE の学習を行い，学習に用いていない評価対象の被験者のデータを用いて適応を行った．次に，評価対象の被験者の学習データを CVAE に入力し，同数のデータを合成した．合成したデータのうち 25%，50%，75%及び 100%をランダムに取り出し，元の学習データと合わせて新たな学習データとした後，拡張された学習データを用いて EEGNet の学習を行った．

3　結果と考察

3.1　転移学習

音声想起時の MEG から学習した VAE から転移学習した EEGNet の分類結果を Fig. 2 に示す．6 人の被験者において，転移学習をしなかった場合（w/o trans）と比較して，転移後に全結合層のみファインチューニングした場合（w/ trans, fc-ft），もしくはネットワーク全体をファインチューニングした場合（w/ trans, whole-ft）の分類精度が向上した．このうちの 4 人の被験者において，各被験者のデータに VAE を適応させた後に転移学習を行うことでさらに分類精度が向上した（w/ adapt trans）．これらの結果は，複数の被験者の MEG データを利用して学習した VAE を転移学習に用いることや，VAE の被験者への適応が音声想起時の MEG の分類にも有効であることを示している．

被験者間モデルを用いた時の音声想起時の MEG の分類結果を Table 1 に示す．被験者間モデルの分類精度は被験者内モデルよりも高い傾向が見られた．また，被験者間モデルの精度のばらつきは被験者内モデルの精度のばらつきよりも小さかった．この理由として，分類精度が 3 クラス分類のチャンスレート（33.3%）に近く，個人差の影響よりも分類器の学習データ数が増えたことの恩恵の方が大きかったことが考えられる．

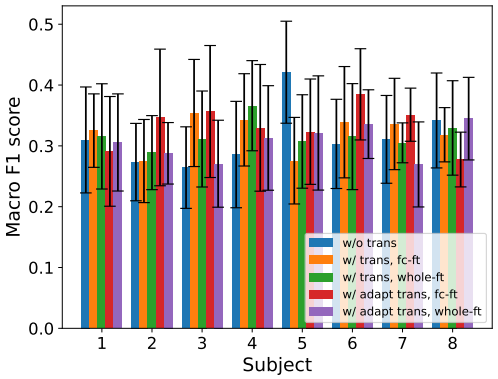

Fig. 2　Macro F1 scores of imagined speech classification with transfer learning using the VAE model trained with only speech imagery MEGs. Each errorbar indicates the standard deviation.

Table 1　Macro F1 score of the between-subject models for imagined speech classification

Test subject	Macro F1 [%]
1	36.1
2	33.2
3	34.7
4	37.2
5	35.9
6	37.6
7	31.4
8	36.2
Mean ± SD	35.3 ± 2.1

音声想起時及び聴取時の MEG を用いて学習した VAE を転移学習させた EEGNet の分類結果を Fig. 3 に示す．学習データ数が増加したにも関わらず，転移学習や VAE の適応による精度向上はあまり見られなかった．聴取時の MEG データが想起時 MEG の分類器の学習に悪影響を与えた可能性が考えられる．

そこで，音声想起時及び聴取時の MEG の分類のしやすさの違いを調査した．1 回目の音声聴取時の MEG から聴取していた音声を識別する EEGNet を学習した．また，聴取時及び想起時の MEG データを用いて学習した VAE を用いて転移学習も行った．Fig 4 に音声聴取時の MEG の分類結果を示す．聴取時の MEG の分類精度は約 50%程度と想起時の MEG 分類結果よりも高く，転移後にネットワーク全体をファインチューニングすることで精度が向上した．全結合層だけをファインチューニングした場合は精度が低下したが，これは評価対象の被験者のデータが VAE の学習に用いられておらず，転移させたエンコーダーから得られる特徴表現が評価対象の被験者のものと乖離

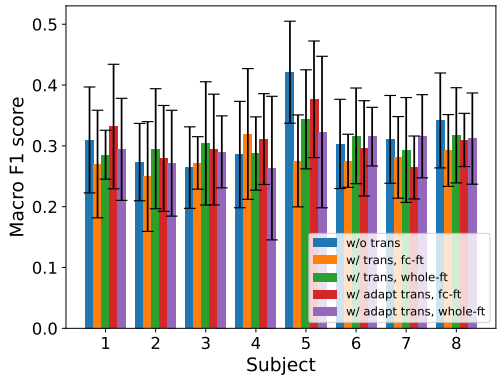

Fig. 3　Macro F1 scores of imagined speech classification with transfer learning using the VAE model trained with MEGs during speech imagery and listening to speech. Each errorbar indicates the standard deviation.

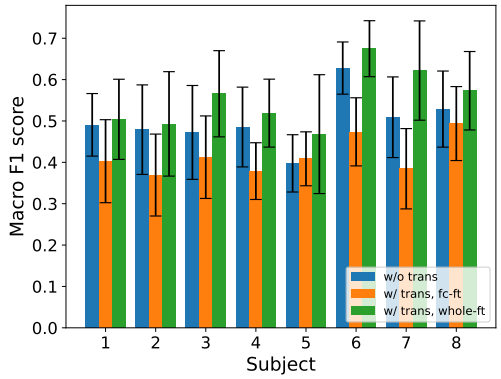

Fig. 4　Macro F1 scores of classification of MEG while listening to speech with transfer learning. Each errorbar indicates the standard deviation.

していたためだと考えられる．

音声想起時の MEG に比べて聴取時の MEG の方が分類が容易であることから，VAE の学習においても聴取時の MEG の方が特徴表現を獲得しやすいと考えられる．聴取時及び想起時の MEG データを学習に用いた VAE モデルの転移学習が EEGNet の分類精度をあまり向上させなかった原因として，聴取時 MEG の分類に有利なエンコーダーが学習され，想起時 MEG の分類器の学習にはあまり寄与しなかった可能性が考えられる．

3.1.1　データ拡張

Fig 5 にデータ拡張した時の EEGNet の分類精度を示す．多くの被験者の結果において，データ拡張を行った場合に分類器の性能が向上した．複数の被験者のデータを用いた学習と評価対象の被験者への適応を行った CVAE を用いて学習データを生成したことで，学習データの数と多様性が増し，分類器の汎化性能が向上されたと考えられる．

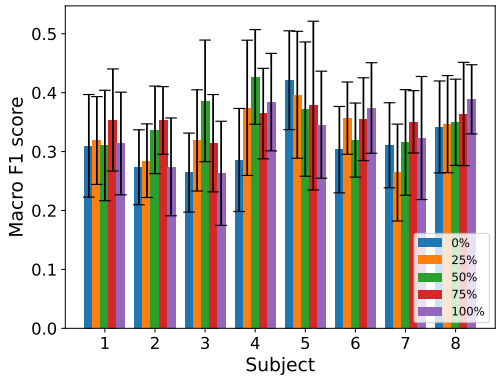

Fig. 5　Macro F1 scores of imagined speech classification with data augmentation. Each errorbar indicates the standard deviation.

4　おわりに

本研究では，VAE を用いて音声想起時の MEG から特徴表現の学習を行った．VAE のエンコーダーの重みを用いて EEGNet を転移学習することによって音声想起時の MEG の分類精度が向上した．また，条件付き VAE を用いて音声想起時の MEG データを生成し，拡張した学習データを用いて分類器を学習した結果，データ拡張を行わなかった場合と比べて分類精度が向上した．これらの結果は，より多くの被験者のデータを用いて学習した VAE の特徴表現を用いることで，利用できるデータ数が限られた新たな個人の想起音声の分類タスクでも分類精度を向上できることを示している．

謝辞　本研究の一部は，JSPS 科研費 JP22K18626 の支援を受けて実施された．

参考文献

[1] C. H. Nguyen *et al.*, J. Neural Eng., 016002, 2018.
[2] S. Uzawa *et al.*, IEEE EMBC, 2542–2545, 2017.
[3] 矢野ら，音講論（春），507–510，2020.
[4] 山名ら，音講論（春），517–520，2023.
[5] J. Devlin *et al.*, arXiv:1810.04805, 2018.
[6] K. He *et al.*, CVPR, 16000–16009, 2020.
[7] A. Baevski *et al.*, NeurIPS, 2020
[8] D. P. Kingma, M. Welling, ICLR, 2014.
[9] M. Dai *et al.*, Sensors, 19(3), 551, 2019.
[10] L. Bollens *et al.*, IEEE ICASSP, 1256–1260, 2022.
[11] K. Sohn *et al.*, NIPS, 2015
[12] V. J. Lawhern *et al.*, J. Neural Eng., 15, 056013, 2018.