The Transformer-like model と HMM による日本語単語サイレントスピーチ時における単一試行脳波の解読
Decoding single-trial EEGs during silent Japanese words by the Transformer-like and HMM

山﨑 敏正†, 赤迫 健太†, 伊藤 智恵子†
Toshimasa Yamazaki, Kenta Akasako, Chieko Ito

†九州工業大学大学院情報工学研究院
Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology
tymzk@bio.kyutech.ac.jp, akasako.kenta289@mail.kyutech.jp, ito.chieko584@mail.kyutech.jp

概要

本研究では、頭皮脳波を利用した SSI として、日本語単語 SS 時の単一試行脳波を Transformer の attention 機能と HMM によって解読する．日本語が mora（拍）言語なので、SS された単語を構成する拍の系列として解読する．本手法は（1）単一試行脳波からノイズ下で信号を復元する RNN、（2）RNN 出力と拍 ERPs の内積値から単語を構成する拍の確率値の算出、（3）拍確率値、言語モデル、HMM による拍系列の生成、から構成される．2〜7 拍から成る単語の解読結果を示す．

キーワード：頭皮脳波, SSI (silent speech interface), Transformer, 拍, RNN, HMM

1. はじめに

Silent speech interface（SSIs）とは、音声生成やそうしようとした時に生じる非音響的な生体信号からその音声を解読することにより、口頭会話を復元するための補助的手段である[1][2]．音声に関連した生体信号の様々なセンシング様式の中で、音声生成に関連する脳領域の神経活動を直接的におよび間接的に捉えられるのが脳波である．

脳波を利用した silent-speech-to-text アプローチは、初めて Suppes, Lu & Hau[3]により 7 つのサイレント英単語で調べられ、後に音素[4][5]、音節[6]へ進んでいった．日本語については、近年、2-mora（後述）単語をサイレントスピーチ（silent speech、SS）した時の頭皮脳波の解読[7]、2-mora 以上の単語を SS した時の単一試行脳波の解読[8]が試みられている．より最近では、侵襲的にはヒアリングや発話時に記録された ECoG（electrocorticography）の解読[9][10][11]、非侵襲的には MEG（magnetoencephalography）[12]や fMRI[13][14]の解読が高精度な結果を得ている．

本研究では、非侵襲的、コンパクトで portable な SSI として、単一試行頭皮脳波を利用した研究[8]を発展させる．本手法の neural network 構造は the Transformer [15]に類似しており HMM（hidden Markov model）を含む．

2. 実験と方法

2.1 実験方法

被験者は 22〜23 歳右利き学生ボランティア 6 名（内、女性 1 名）である．尚、実験手順は「九州工業大学大学院情報工学研究院等における人を対象とする研究審査委員会」で承認された．

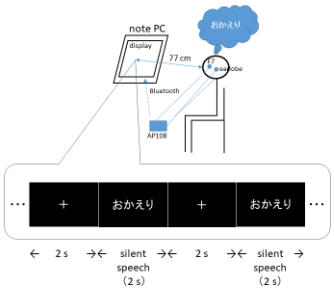

図1 本実験装置の全体像

最初に、被験者から 77 cm 前方に位置するディスプレイ上に注視点”＋”が 2s 間提示される．次に、SS すべき日本語単語がひらがなで 2s 間提示される．単語が提示されたら、被験者は出来る限り早く SS する．この計 4 秒間を 1 trial とする（図 1）．日本語は”mora”（拍）言語と呼ばれ[16]、mora は日本語単語生成の音韻的な単位である[17]．即ち、日本語単語はすべてこの拍で表現することが出来る．図 2 は日本語拍すべてを示している[18]．同図の第 1 列目が日本語母音を示しており、最後の列 “ん”、“っ”、“—” は、それぞれ、撥音、促音、長音と呼ばれ、最後の 2 拍は単独では発音が出来ない．また、日本語拍は（）内を除くと 109 ある．

2.2 脳波計測

1 つのアクティブ電極 (AP-C151-015, ミユキ技研,) が国際 10-20 システムに従い F7 に設置された．2 つの耳朶電極の平均を参照とした．これらの電極で記録された脳波はワイヤレス生体信号アンプ(Polymate Mini

AP108, ミユキ技研) へ送られ、60 Hz の notch filter をかけ 10,000 倍に増幅された. サンプリング周波数は 500 Hz とした. エポック区間は単語提示の前後 2 s とした. オンラインで A/D 変換された脳波データは Bluetooth を介してすぐにパソコンに転送され、パソコン内のハードディスクに格納された (図 1 を見よ).

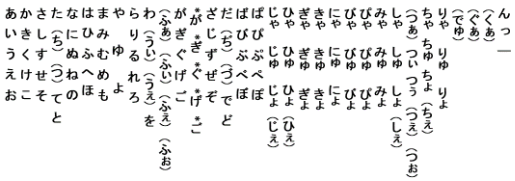

図 2　日本語の拍. () 内の拍は外国語と感嘆詞. ＊はが列の鼻音.

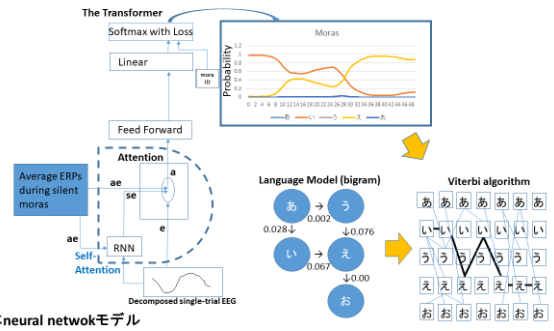

図3　本neural netwokモデル

2.3 脳波データ解析方法

本研究の脳波解析モデルは図 3 の通り、the Transformer に類似し、HMM を伴う. まず初めに、得られた単一試行脳波データは時間的に overlap するブロックに分解される. ブロック間は 2 ms ずつずれていき、ブロック数は 50 とした. 1 番目のブロックは 0〜200 ms、最後のブロックは 100〜300 ms となる. 各ブロックは 2 つのルートに分岐し、1 つは後述する RNN へ、もう 1 つはそのまま上位へ行く.

加算平均の原理は単一試行脳波＝信号＋ノイズ（平均 0）を前提とする. このノイズ下で記憶すべきパターンを復元できる RNN（recurrent neural network）が知られている[19]（考え方の詳細は[8]参照）. この信号および記憶すべきパターンを、各拍を SS した時の加算平均後の ERPs（event-related potentials）とし、ノイズ＝単一試行脳波－信号として RNN に入力すると、RNN の出力はノイズ下で各サイレント拍脳波の復元となる（拍が 106 個であれば 106 個の波形が得られる）.

次に the Transformer の attention 機能を施す. Scoring は RNN 出力である各サイレント拍脳波と単一試行脳波の内積として求まる. こうして、正規化された内積値から成る、拍数次元ベクトル $\boldsymbol{a}$ が得られる（単一試行脳波の中に各サイレント拍脳波がどの程度含まれるかを定量化する）. 更に、”Feed Forward”層と” Linear” 層を経て、教師信号（mora ID）を利用して” Softmax with Loss” 層で backpropagation により拍確率値を学習させた. こうして、各ブロックにおける拍確率値のプロットが得られる. 尚、内積計算の際、ブロックとサイレント拍 ERPs の間でも試した.

最後に、上記の確率値と言語モデル（ここでは拍の bigram[20]）を利用して、Viterbi アルゴリズム[21]により、HMM の状態遷移として、サイレント単語を構成するであろう拍の系列が求まる

3. 結果

2〜7 拍から成る日本語単語（文節を含む）12 個の解読結果を示す. 具体的には、「はい」、「いいえ」、「おかえり」、「くるま」、「すごい」、「とうきょう」、「よろしく」、「うたごえ」、「えきたい」、「ろうそく」、「おはよう」、「よかったなあと」である. 候補拍は、単語を構成する拍に加えて、母音を含み、聞こえ度[22]が高いものを選んだ. 尚、認識率を定量化するために以下で定義される MER（mora error rate）を算出した. The Transformer まででは、

$$MER = \frac{L_{coincide}}{L}$$

但し、L はブロック数、$L_{coincide}$ は各ブロックにおいて確率値最大の拍が教師信号拍と一致した総数とする. HMM までの場合も同様である.

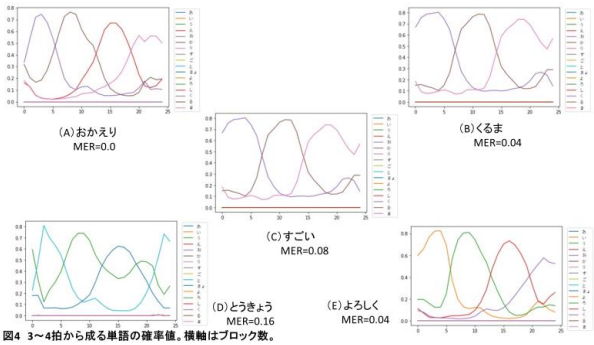

図4　3〜4拍から成る単語の確率値。横軸はブロック数。

Training performance

図 3 には、RNN 出力を利用した「いいえ」の確率値（the Transformer 出力）プロットが示されている. このグラフから良く解読されているようである. 実際、MER=0.1176 であった. 学習データ数は 5、候補拍は 5

母音である.

図 4 には、拍 ERP を利用した（A)「おかえり」、(B)「くるま」、(C)「すごい」、(D)「とうきょう」、(E)「よろしく」の確率値が示されている. 学習データ数は 99、候補拍は 17 とした．その結果、MER はそれぞれ 0.00, 0.04, 0.08, 0.16, 0.04 であった.

RNN を利用した「うたごえ」、「えきたい」、「ろうそく」ではそれぞれ MER=0.95, 0.4251, 0.70 であった．但し、学習データ数 20、候補拍 10、ブロック数 40 とした．拍 ERP に変更すると、それぞれ MER=0.275, 0.275, 0.225 に改変された.

一方、拍 ERP を利用した「よかったなあと」と「おはよう」では MER=0.72 と 0.40 であった．更に、「よかったなあと」の HMM までは MER=0.80 であった．但し、「っ」ERP＝「あっ」ERP－「あ」ERP とした．また、ブロック数は 50、学習データ数は 1 のみ、候補拍の数は 11 とした.「おはよう」では、ブロック数が 25、学習データ数が 10、候補拍の数が 17 であった.

4. 考察—おわりにかえて—

本研究では、頭皮脳波を利用した SSI として、日本語単語 SS 時の単一試行脳波の解読を試みた. 日本語単語がすべて拍によって表現されることに注目し、単語を構成する拍の系列として解読した．そのために本手法は以下のステップを要する：（1）加算平均の原理：単一試行脳波＝信号＋ノイズに着目し、ノイズ下で信号を復元する RNN を採用；（2）単一試行脳波（ノイズを含む)、拍 ERPs（信号)、RNN 出力を利用して the Transformer の self-attention 機能を適用；（３）the Transformer の出力(拍確率値)、言語モデル(拍の bigram)を使って HMM により単語を構成する拍の系列を推定.

まだ脳波データの数が少ないが、拍の最大確率値の時間的な推移を見る限り、良く解読出来ているようである. しかしながら、脳波データが多くなると以下に示すハイパーパラメータを十分に検討しなければならない．現時点で考えられるだけでも

- 学習脳波データ数、
- 内積をとる脳波データの区間および長さ、
- ブロック数、
- Feed Forward 層の多層化、
- 単語を構成する拍の数、
- RNN 出力／拍 ERPs
- RNN パラメータ（ノイズレベルなど）
- 拍 ERPs の加算回数、
- (理想的には 109 個あるいはそれ以上であるが）候補拍の数、
- 教師信号としての mora ID（ブロック数を拍数で等分配して良いのか)、
- (単独では発音出来ない）「っ」、「ー」ERPs の求め方、

などである. 例えば、内積をとる脳波データの区間および長さについては、発話と SS の ERPs で共通する成分[23][24]、N1 と P2 を含むようにしなければならない. また、MER を見る限り、言語モデルを含む HMM の利用は再考を要する.

最後に、今後、単語数を増やして本格的な testing を実施していく.

文献

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, (2010) “Silent speech interfaces”, Speech Commun., Vol.52, No.4, pp.270-287.

[2] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martín Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, (2020) “Silent speech interfaces for speech restoration: A review”, IEEE Access, Vol.8, pp.177995-178021.

[3] P. Suppes, Z.-L. Lu, and B. Han, (1997) “Brain wave recognition of words”, Proc. Natl. Acad. Sci., Vol.94, pp.14965-14969.

[4] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, (2009 “Single-trial classification of vowel speech imagery using common spatial patterns”, Neural Networks, Vol.22, pp.1334-1339.

[5] M. Matsumoto, and J. Hori, (2014) “Classification of silent speech using support vector machine and relevance vector machine”, Applied Soft Computing, Vol.20, pp.95-102.

[6] M. D’Zmura, S. Deng, T. Lappas, S. Thorpe, and S. Srinivasan, (2009) “Toward EEG sensing of imagined speech, in Human-Computer Interaction”, Part I, HCII 2009, LNCS 5610, ed. J. A. Jacko, pp.40-48, Springer-Verlag Berlin Heidelberg.

[7] H. Yamaguchi, T. Yamazaki, K. Yamamoto, S. Ueno, A. Yamaguchi, T. Ito, S. Hirose, K. Kamijo, H. Takayanagi, T. Yamanoi, and S.Fukuzumi, (2015) “Decoding silent speech in Japanese from single trial EEGs: Preliminary results”, Journal of Computer Science Systems Biology, Vol.8, No.5, pp.285-292.

[8] 山﨑敏正、赤迫健太、森田寛伸、德永由布子、上村旺生、柳橋奎、柏田倫孝、(2023) “Decoding single-trial EEGs during silent Japanese words by the Transformer-like model”, 言語処理学会第 29 回年次大会　予稿集.

[9] G. K. Anumanchipalli, J. Chartier, and E. F. Cchang, (2019) “Speech synthesis from neural decoding of spoken sentences”, Nature, Vol.568, pp.493-498.

[10] F. R. Willett, D. T. Avansino, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy, (2021) “High-performance brain-to-text communication via handwriting”, Nature, Vol.593, pp.249-253.

[11] D. A. Moses et al., (2021) “Neuroprosthesis for decoding speech in a paralyzed person with anarthria”, New England Journal of Medicine, Vol.385, pp.217-227.

[12] D. Dash, P. Ferrari, and J. Wang, (2021) “Decoding imagined and spoken phrases from non-invasive neural (MEG) signals”, Front. Neurosci., Vol.14, 290.

[13] F. Pereira, et al., (2018) “Toward a universal decoder of linguistic

meaning from brain activation", Nat. Commun., Vol.9, No.1, 963.
[14] J. Tang, A. LeBel, S. Jain, and A. G. Huth, (2022) "Semantic reconstruction of continuous language from non-invasive brain recordings", bioRxiv preprint doi: https//doi.org/10.1101/2022.0929.509744.
[15] A. Vaswani, et al., (2017) "Attention is all you need", arXiv preprint arXiv: 1706.03762v5.
[16] Kubozono, (2002) "2. Mora and Syllable", The Handbook of Japanese Linguistics, ed. N.Tsujimura, pp.31-61, Blackwell Publishers Inc., Massachusetts.
[17] B. G. Verdonschot, S. Tokimoto, and Y. Miyaoka, (2019) "The fundamental phonological unit Japanese word production: An EEG study using the picture-word interference paradigm", Journal of Neurolinguistics, Vol.51, pp.184-193.
[18] 金田一春彦、(2017), 日本語　新版（上）、岩波書店、第 53 刷.
[19] R. Laje, and D. V. Buonomano, (2013) "Robust timing and motor patterns by taming chaos in recurrent networks", Nat. Neurosci., Vol.16, No.7, pp.925-933.
[20] 今栄国晴、(1960) "日本語の digram の相対頻度とその特性", 心理学評論、Vol.4, pp.85-100.
[21] S. Furui, (2015) Digital Speech Processing, Synthesis, and Recognition, 2nd ed., rev. and expanded, Eastern Hemisphere Distribution, New York, NY 10016.
[22] 窪園晴夫、(2018), 音声学・音韻論、くろしお出版、第 11 刷.
[23] R. C. Pratap, (1987) "The speech evoked potential in normal subjects and patients with cerebral hemispheric lesions", Clinical Neurology and Neurosurgery, Vol.89, No.9, pp.237-242.
[24] S. Tsukiyama, and T. Yamazaki, (2019) "Discriminability among Japanese vowels using early components in silent-speech-related potentials", IEICE Technical Report, SP2019-31, WIT2019-30 (2019-10), pp.81-86.