

# 脳磁図 (MEG) による聴覚刺激再構成システムの提案と評価

山下 正人<sup>1,a)</sup> 中沢 実<sup>1</sup>

**概要：**現在の Brain Computer Interface(BCI) は、スクリーン上からの画像刺激や方向意図 (移動, クリック) を脳活動からの推定により, 文字の入力を行う研究が行われている。しかし, それらは画面に焦点を当て続けなければならない, 入力速度も遅い。そのため, 極めて非効率である。画像刺激以外からの BCI の手法として, 硬膜下皮質表面電位 (ECoG) の電気活動を利用して聴覚刺激から高品質な音声を再構成する研究が行われている。しかし, ECoG は侵襲型であるため, 手術が必要である。本研究では, 非侵襲型の脳磁図 (MEG) を用いて健常者を対象に聴覚刺激を与えた際の脳活動を計測し, 聴覚刺激を再構成するシステムの提案を行い, システムについて評価を行う。中間表現を取得する AutoEncoder(AEC) の学習では, 客観評価の ESTOI:平均 0.87, 主観評価では, 数字の認識度:98.88%, 単位の認識度:97.22%, 音声の品質評価:平均 4.72 を得た。しかし, 脳活動から音声を再構成する DNN の学習結果, 客観評価の ESTOI:平均-0.001 となり, 音声再構成されなかった。今後, 手法を見直し, 精度を向上させる必要がある。

## 1. はじめに

近年では, Brain Computer Interface(BCI) に関する研究が多く行われている。BCI の研究では, スクリーン上からの画像刺激や方向意図 (移動, クリック) を脳活動からの推定により, 文字の入力を行う研究が盛んに行われている [1][2]。しかし, それらは画像に焦点を当て続けなければならない, 入力速度も遅い。画像刺激や方向意図 (移動, クリック) を用いた手法では, 生成された文章は離散単位で復元したものである。そのため話者の特徴や感情といった情報を付与することができない。これらの問題点から, 画像刺激以外の手法を用いて, 容易に音声を再構成することができる BCI の研究開発が求められている。

画像刺激以外の手法として, 聴覚刺激を用いた手法 [3] や, 被験者が音声を読み上げた際の脳活動を用いた手法 [4][5] などがある。音声を読み上げた際の脳活動を用いた手法では, 被験者自身が発話を行い, その際に発生した脳活動を用いることによって, 音声の再構成を行っている。被験者が障害のない状態で測定を行っているため, 被験者が障害を発症してしまった際には, 筋電を動かすことができなくなってしまう。それに伴い発話によって生じていた脳活動が変化してくる可能性があるため, 筋電を動かして発話を行うことができない方へ適応させることが難しくなると考えられる。本研究では, 非侵襲型の脳磁図 (MEG) を用い

て聴覚刺激を与えた際の脳活動を計測する。計測された脳活動から高品質な聴覚刺激を再構成することを目標に非侵襲型の脳活動測定手法を用いることによる聴覚刺激の再構成についてシステムを提案し, 評価を行う。

## 2. 関連研究・測定技術

本章では, 脳活動の測定技術及び関連研究について説明する。

脳活動に関する研究開発に用いられている測定手法は主に硬膜下皮質表面電位 (ECoG), 頭皮脳波 (EEG), 脳磁図 (MEG) の 3 つである。ECoG は侵襲型の脳活動測定手法であり, EEG と MEG は非侵襲型の脳活動測定手法である。ECoG と EEG が測定の対象にしているのは, 神経活動によって生じる電位であり, 2 つの電極間の電位差である。一方, MEG が測定の対象にしているのは, 神経活動によって生じる磁場である。3 つの手法は, いずれも神経活動の変化における一次信号を測定しているという点では共通している [6]。ECoG は手術を行い, 測定電極を硬膜下に電極を設置する。そのため, 同じ皮質電位を測定している EEG と比べ, 頭皮・頭蓋からの影響を受けずにすみ, 信号の経過的変化による信号の減衰が少なく, S/N 比が高いという利点がある。EEG は手術を行う必要がなく, 頭皮上に電極を設置して測定を行う。そのため, 比較的容易に測定が可能であり, 測定時の身体的制約が少ない。しかし, 頭皮・頭蓋からノイズの影響を強く受けるため, 測定された信号に, ノイズが多く含まれてしまうという問題点がある。MEG は磁気センサの超伝導量子干渉素子 (SQUID) を用いて測

<sup>1</sup> Kanazawa Institute of Technology,  
7-1 Ogaigaoka Nonoichi Ishikawa 921-8812, Japan

<sup>a)</sup> b6800602@planet.kanazawa-it.ac.jp

定する手法である。MEGで測定するためには、SQUIDの超伝導状態を保持する必要がある。そのため、継続的な液体ヘリウムの補充が不可欠となる。また、脳の神経活動の磁場の大きさは10～100fT(フェムトテスラ)であり、非常に微弱であるため、微弱な脳活動を安定して測定するために、シールドルームが必須になってくる[7]。各測定手法の時間分解能は、3つの手法はどれも約 $10^{-4}$ ～ $10^{-3}$ (s)と優れており、空間分解能は、ECoG: 約10(mm), EEG: 約30～40(mm), MEG: 約5～7(mm)とMEGが最も優れており、空間分解能が最も悪いのは、EEGである[8][9]。時間分解能とは、脳の同一の場所が短時間に二回活動した場合に、それぞれを時間的に独立した脳活動として計測できる最短の時間間隔である。また、空間分解能とは、脳の異なる部位が同時に活動した際に、それぞれを独立した脳活動として計測できる最小の距離である。

MEGは同じ非侵襲型の測定手法EEGと比べて、高い空間分解能を持っており、高周波成分が減衰しにくいという利点がある。そのため、ガンマ帯域などの高周波帯域の脳活動計測を行うことが可能になるという利点がある。また、基準電極との電位差を測定している手法では、基準電極の部位が活性化した場合には正確な振幅・電位分布が得られないが、MEGでは磁束密度の絶対値を計測しているため、基準電極の活性化した際の問題が生じることがない[6]。これらの利点から、本研究では、MEGを用いて聴覚刺激時の脳活動を測定する。

脳活動を用いて音声を再構成する研究に関して、既にいくつかの研究事例がある。Akbari, Hら[3]の研究では、ECoGを用いて人間の聴覚皮質からクロズドセットの理解可能な音声を再構成することを目的として研究を行なった。その中では聴覚スペクトログラムと音声合成パラメータを含む線形および非線形回帰法と再構成のターゲットとして使用される音響表現への再構成精度の依存性、低周波数と高周波数の範囲における再構成精度を調査した。結果、音声合成パラメータと非線形回帰法を用いた手法により、数字認識タスクで最高の主観的および客観的スコアを達成した。この手法は線形回帰を用いたベースライン法よりも明瞭度を65%改善している。

Wang, Jら[5]の研究では、MEGを用いて被験者が発話した時の脳活動をデコードすることによって、発話したフレーズを解読する研究を行なった。その結果デコードとして人工ニューラルネットワーク(ANN)を用いることにより、5つのフレーズに対して約94.54%の分類精度を達成した。

Patrick, Jら[10]の研究では、EEGによって測定された脳活動を用いて選択的聴覚注意(SAA)の解読を目的に4人の話者が空間的に分離され、異なるメッセージを聴取者に提示している環境において、被験者がどのメッセージに焦点を当てているかの解読を行なった。刺激再構成を用い

ることにより、平均61.1%で焦点を当てている方向を正しく分類することができた。

聴覚刺激における脳活動の反応に関して、既にいくつかの研究事例がある。

Khalighinejad, Bら[11]の研究では、EEGを用いて、連続音声における音素カテゴリへの誘発された神経反応の特性特徴付けを行なっている。音素に対する応答が音素の種類によって、音素開始後50msや400msの異なるタイミングで発生した。このことから音素の種類により複数の識別可能な神経反応が生じることを見出した。

Chananel Braimanら[12]の研究では、健常者と脳損傷患者の話し言葉の自然音声包絡線(NSE)に対するEEG反応の潜時および振幅を調査した。その結果、脳損傷患者は、健常者のEEG反応と比べて、進行性の潜時鈍化を示した。しかし、fMRIベースの精神的画像タスクを実行できる脳損傷患者は、健常者と比較した際に、NSE潜時において、統計的な有意差はなかった。この調査により、脳損傷患者の隠れた認知の検出を改善し、研究における意識障害のある患者の層別化を改善する受動的な手段の可能性を示している。

これまでの先行研究より、脳活動における聴覚刺激の再構成は有益であると考えられる。また、聴覚刺激において脳損傷患者が、被験者と有意差が生じなかったという点から、聴覚刺激を用いることは、脳損傷患者に対するアプローチとしても有益であると考えられる。本研究では、Akbari, Hら[3]の研究で用いられていた手法のターゲットパラメータを利用し、非線形回帰法を用いた復号モデルをMEGに適用したモデルとすることにより、非侵襲型の脳活動測定手法であるMEGを用いることによる聴覚刺激再構成を行う。復号モデルにおける再構成精度を調査し、MEGにおける聴覚刺激再構成について利用可能性を検証する。

### 3. 脳活動の測定手法

本章では、聴覚刺激を用いた脳活動の取得方法及び聴覚刺激について説明する。

MEGを用いた脳活動の測定には、同軸型グラジメータを用いて脳活動を測定する。チャンネル数は160chであり、サンプリング周波数は1kHzである。バンドパスフィルタを適用し、0.1Hzから200Hzの周波数帯域を取得する。測定は、金沢工業大学天池キャンパスのシールドルーム内で測定を行った。図1に聴覚刺激の測定の流れを示す。聴覚刺激を提示した際の脳活動の測定には、学習データの取得を6回、テストデータの取得を6回ずつ、各被験者に対して行う。

本研究の聴覚刺激には、JUST corpus[13]の音声データセットを用いた。JUST corpusは無響室で収録された一人の日本語女性話者の音声サンプリング周波数48kHzで

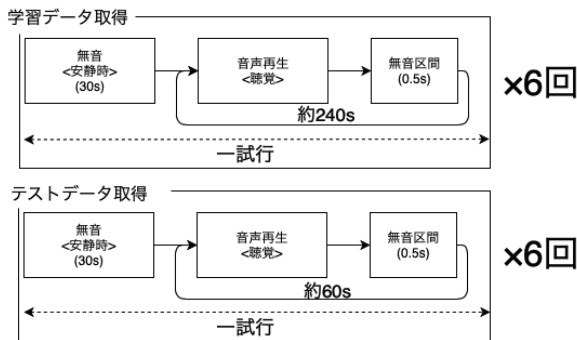


図1 測定の流れ

収録されている。音声の種類は全部で9種類あり、約10時間の音声を含んでいる。JUST corpusの音声データを後述するデータセットのサンプリング周波数と同一にするため、16kHzにダウンサンプリングしたものを利用した。学習データ用の聴覚刺激は、countersuffix26(助数詞)以外の8種類の中からランダムに選んだ音声をつなぎ合わせ、ランダムに並び替えた学習データ用の音声データ(240秒間)を作成した。テストデータ用の聴覚刺激は、助数詞の中から単位が「個、枚、本、冊、台、番」の6種類を単語ごとに分割し、ランダムに並び替えた音声データ(60秒間)を作成した。音声のつなぎ合わせには、無音区間(0.5秒間)を挿入している。各試行で利用する音声データは共通であるが、各音声再生の順番はランダムである。

被験者として男子大学院生3名に協力してもらい、聴覚刺激提示時の脳活動測定を行なった。被験者はシールドルーム内のベッドに横になり、目を開けた状態で測定を行なった。聴覚刺激の提示には、シールドルーム外から延長したエアチューブを用いて、被験者に提示を行なった。実験の開始前にエアチューブを耳につけた際に聴覚刺激を聞いてもらい、被験者が音声を聞き取りやすい音量に調整した。学習データとテストデータの取得は同一日に行なった。試行間に休憩を挟みながら実験を行なった。休憩の間は被験者はシールドルーム外に出ることはなく、ベッドに横になっている状態のままである。試行の開始と終わりはシールドルーム内に取り付けたスピーカーを介し被験者に合図を送った。

MEGによって測定された脳活動を脳信号(brain wave)として定義する。脳信号に対して同一の音声区間かつ同一のチャンネルに対して加算平均を行い、ノイズを軽減させる。その際、目視で脳信号を確認し、ノイズが多く含まれているデータを加算平均の対象から除外した。ノイズが多く含まれていた脳信号の例を図2に示す。

図2は同じ音声を聞いている時の6回分の脳信号を示している。グラフを見ると1-3-40のグラフが他の信号と大きく異なっていることが目視で確認できる。このように6回分の脳信号を目視で見比べた時に、明らかに異なる信号であると人が判断したものは、加算平均の対象から除外

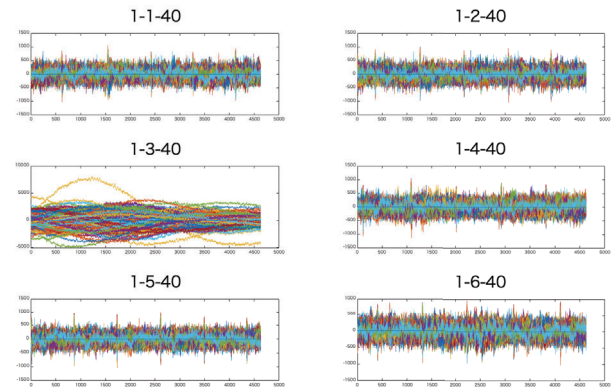


図2 ノイズの例

した。

#### 4. 提案手法

本章では、聴覚刺激再構成システムについて説明する。図3にシステムの概要を示す。

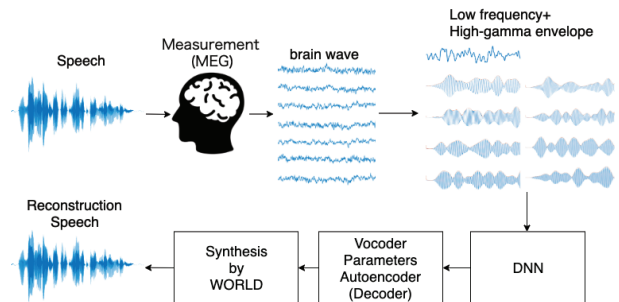


図3 システム概要

本研究では、低周波数(Low frequency)と高ガンマ包絡線(High-gamma envelope)の信号を特徴量として聴覚刺激の再構成を行う。

加算平均後の各チャンネルの脳信号に対して、FIRのローパスフィルタ(4次の30dBを持つチェビシェフウィンドウのローパスフィルタ)を使用して脳信号をフィルタリングすることにより、脳信号の低周波(0~50Hz)成分に減衰する。抽出された低周波数成分の形状は $t \times 160$ となる。ここで、 $t$ は音声区間の切り出す長さであり、160は測定したMEGのチャンネル数である。また、高ガンマ包絡線の抽出にはIIRフィルタ(6次のバターワースバンドパスフィルタ)を使用し、70Hzから150Hzまでの周波数帯域を10Hzごとに8分割して周波数帯域ごとに高周波成分を抽出する。抽出された8つの高周波成分に対して、ヒルベルト変換を行い、高ガンマ包絡線を得る。得られた高ガンマ包絡線の形状は、 $t \times 160 \times 8$ となる。8は分割した周波数帯域の個数である。

抽出された特徴量から音声を再構成する部分の詳細を図4に示す。

音声を再構成する際には、音響表現をターゲットパラ



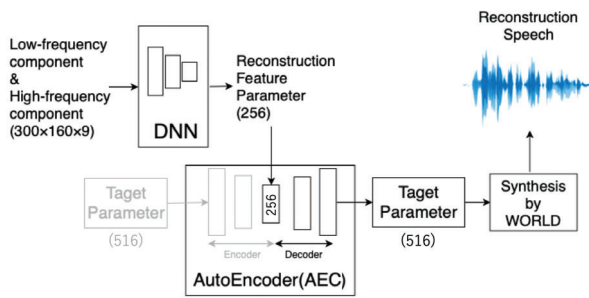


図4 音声再構成の詳細

メータとする。そこでは、音声合成システムで活用されている Vocoder パラメータを用いる。脳活動からの復号モデルとして、DNN(Deep Neural Network) と AutoEncoder(AEC)[14] の Decoder レイヤーを用いることにより、脳活動からターゲットパラメータへの復号を行う。DNN から直接ターゲットパラメータを再構成することなく、AEC の Decoder を活用する理由は、ターゲットパラメータが高次元であるため、AEC を活用することにより、ターゲットパラメータをより小さな低次元に特徴を圧縮することができ、圧縮された特徴を中間表現として利用することによって、効率的に学習を行うことができると考えたためである。そのため、AEC の学習には、Encoder と Decoder の両方を用いて行うが、音声を再構成する際には、Decoder ののみを利用して音声の再構成を行う。DNN と AEC の Decoder レイヤーから得られたターゲットパラメータを WORLD[15] で提案されている合成アルゴリズムに適用することによって、音声の再構成が行われる。以下の節にて、Vocoder パラメータの詳細、提案する AEC モデル、DNN モデルに関して、詳細に説明する。

#### 4.1 ターゲットパラメータ:Vocoder パラメータ

本節では、音声再構成におけるターゲットパラメータについて説明する。本研究でのターゲットパラメータは音声合成システム (WORLD) で使用されている Vocoder パラメータを用いる。WORLD[15] で利用されている Vocoder パラメータは主に以下に示す 4 つのパラメータで構成されている。

- (1) スペクトル包絡線 (Spectral Envelop)
- (2) 非周期性パラメータ (Aperiodic Parameter)
- (3) 基本周波数 (F0)
- (4) 有声無声励起ラベル (VUV)

これらのパラメータを音声波形から 5ms ごとに求め、パラメータを合成アルゴリズムに適用することにより、入力の音声波形に類似した波形を得ることができる。サンプリング周波数 16kHz の音声に対する、特徴量の個数はスペクトル包絡線:513 個、非周期性パラメータ:1 個、基本周波数:1 個、有声無声励起ラベル:1 個 (有声 (1) or 無声 (0)) となる。本研究では、音声データに対して、WORLD[15](D4C

edition[16]) を用いることにより、Vocoder パラメータを求める。スペクトル包絡線を推定には、CheapTrick アルゴリズム、非周期性パラメータの推定には、D4C アルゴリズム、基本周波数の推定には Harvest アルゴリズム [17] をそれぞれ用いた。有声無声励起ラベルは基本周波数から推定される。

#### 4.2 AutoEncoder

本節では、提案する AutoEncoder(AEC) モデルの詳細について説明する。AEC モデルは、Akabari ら [3] の手法で使用されていた構造や活性化関数を参考にして作成した。提案する AEC モデルの概要を図 5 に示す。

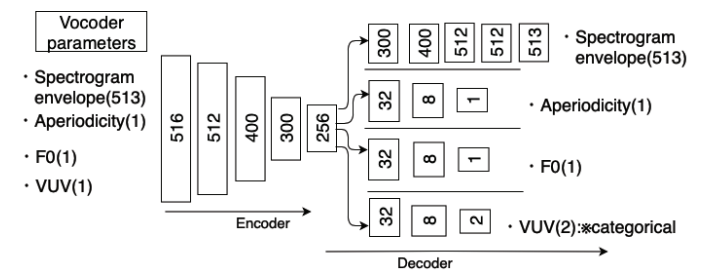


図5 AEC の概要

入力 Vocoder パラメータの 4 つをまとめて 1 つの入力とする。出力では、各パラメータがそれぞれ出力として得られる。有声無声励起ラベルだけは、入力と出力のサイズが異なっている。これは、有声無声励起ラベルの値は、無声 (0) と有声 (1) の値のみを出力としたいため、カテゴリ分類を行うため出力のサイズが 2 となる。

AEC の学習には、Speech Commands[18] と Common Voice[19] の音声データを用いて行う。Speech Commands には、サンプリング周波数 16kHz で、30 種類の短い英単語を発音した約 1 秒間のデータが含まれている。Common Voice は、オープンな音声データベースであり、言語は英語のデータセット、サンプリング周波数 48kHz で収録されている。Speech Commands と Common Voice はサンプリング周波数が異なる。そのため、Common Voice を 16kHz にダウンサンプリングして利用する。前節で示したアルゴリズムを用いて得た Vocoder パラメータを学習の入力値と正解値とする。学習の際には、Encoder レイヤーからの出力を tanh 関数によって活性化を行う。その後、Decoder レイヤーに渡す前に、ガウシアンノイズを付与する。ノイズを付与することによって、Decoder レイヤーをより堅牢なものとするのが期待できる。

Akabari ら [3] の手法では、スペクトル包絡線の出力層の活性化関数として、ReLU 関数を使用している。ReLU 関数を使用した状態で実際に学習を行うと、「dying ReLU」[20] という問題が発生してしまい、学習がうまく行われなかった。「dying ReLU」とは、学習時に負の値を持ってし

まい、常に 0 を出力する場合、負の範囲での ReLU の勾配は 0 であるため、ニューロンからの出力が負の値を持つと、学習が行われなくなってしまうという問題点がある。これらが発生する理由として学習率が高すぎる、または負の偏りがある場合などが挙げられる。学習率の変更や重みの初期化、LeakyReLU や ELU などの ReLU から派生された活性化関数を用いることにより、発生する確率を減少させることはできるが、根本的な解決とはならない。学習で「dying ReLU」が発生していた部分は、出力層であった。そのため、LeakyReLU や ELU を活性化関数を用いてしまうと出力に負の値を持ってしまう。スペクトル包絡線は、正の値のみを持つため、LeakyReLU や ELU は出力層の活性化関数には適さない。そこで、絶対値をスペクトル包絡線の活性化関数として用いることにより、学習を行う。

その他の出力層の活性化関数として、非周期性パラメータ、基本周波数には、ReLU 関数、有声無声励起ラベルには、softmax 関数を用いる。損失関数には、スペクトル包絡線:平均絶対誤差 (MAE)、非周期性パラメータ、基本周波数:平均二乗誤差 (MSE)、有声無声励起ラベル:交差エントロピーを用いた。検証用データで得られた損失値を足し合わせて、総和が最小となるように学習を行なった。その際に、スペクトル包絡線の誤差は、小数の値となってしまう。そのため、スペクトル包絡線の損失値に対して、バイアスを掛け合わせてから足し合わせを行う。今回はバイアスを  $10^7$  として設定した。

初期学習率を 0.0001 として、3epoch に渡り、検証時の損失値が改善されなかった際には、学習率を 2 で割り、学習率を減らすように学習を行なった。

### 4.3 DNN architecture

本節では、提案する DNN の詳細について説明する。DNN に対する入力、低周波数成分と高ガンマ包絡線 ( $t \times 160 \times 9$ ) であり、ウィンドウサイズ:300ms、スライドウィンドウ:5ms である。DNN での出力は、正解のターゲットパラメータを学習済みの Encoder レイヤーに渡した時に得られる中間表現である。本論文では、DNN によって得られる出力を再構成特徴パラメータ (Reconstruction Feature Parameter) として定義する。再構成特徴パラメータは、AEC の Encoder によって、ターゲットパラメータが畳み込まれた際と同じ出力となるように学習される。提案する DNN の概要を図 6 に示す。DNN は特徴抽出 (Feature extraction)、特徴要約 (Feature summary) の 2 つのフェーズによって構成されている。

特徴抽出フェーズでは、入力から得られる 9 つの高次元表現 ( $300 \times 160$ ) に対して、各高次元表現を「Conv\_model」により 1 次元 (512) の表現に畳み込み、出力として得る。特徴抽出フェーズでの活性化関数は LeakyReLU 関数を用いた。その後、特徴抽出フェーズから得られた特徴を結合

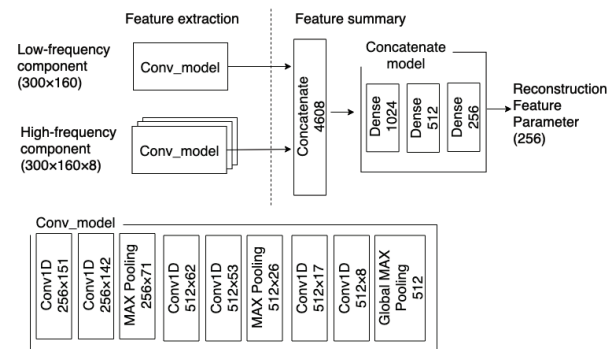


図 6 DNN の概要

し、1 次元表現 (4608) を得る。得られた 1 次元表現に対し、特徴要約フェーズの「Concatenate model」を適用し、再構成特徴パラメータ (256) を得る。「Concatenate model」での活性化関数には、1 層目と 2 層目には、ELU 関数を用い、3 層目には tanh 関数を用いた。

人の脳活動は、同じ刺激を提示しても被験者ごとに異なっていることが知られている。そのため、今回の実験でも被験者ごとに刺激による反応は共通していないということが考えられる。そこで、本研究では、同じ構造の DNN を用いて、被験者ごとのデータを用いて別々の学習を行う。DNN の出力の活性化関数には、AEC と値を共通させるために、活性化関数は tanh 関数とする。学習済みの AEC の Encoder 部分からの出力を DNN の出力の正解値とすることで、学習を行う。損失関数には、平均絶対誤差 (MAE) とし、計算時のバイアスを  $10^5$  として、計算を行う。評価関数には、決定係数の  $R^2$  を用いて、評価を行う。決定係数とは、値の当てはまりの良さを表すものである。決定係数  $R^2$  は (1) 式で計算される。

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

ここで、 $y$  は正解値、 $\hat{y}$  は予測された値、 $\bar{y}$  は正解値の平均である。

初期学習率を 0.001 として、4epoch に渡り、検証時の評価値が改善されなかった際には、学習率を 2 で割り、学習率を減らしながら学習を行なった。

## 5. 実験結果

本章では、実験の評価指標および評価結果について説明する。

### 5.1 評価指標

本節では、再構成された音声に対する精度を検証するために、使用する評価指標に関して説明する。評価指標の測定には、主観評価では、音声の認識度・音声品質を評価し、客観評価では、音声の明瞭度を ESTOI[21] によって、評価する。主観評価とは、評価者を利用した評価方法である。音声の認識度の評価は、再構成された音声を評価者に聞い

てもらい、音声から数字(1~10)の10種類と単位(個, 枚, 本, 冊, 台, 番)の6種類を正しく認識することができるかを評価する。各数字と各単位に関して、どのくらい正しく評価者に認識されたかの割合を指標とする。品質の評価は、再構成された音声を評価者に聞いてもらい、その音声は、「非常に悪い(1)」～「非常に良い(5)」のいずれかに当てはまるか評価を行なってもらい、「非常に良い(5)」とする音声は、脳活動測定時に被験者に聞いてもらった音声である。評価者によって音声品質の指標にばらつきが生じてしまう。そこで、全評価者の評点を平均した MOS(Mean Opinion Score) 値とすることで、定量化を行う。ESTOI は、音声合成技術の評価に使用されており、音声の明瞭度を定量的に評価する指標である。ESTOI によって取りうる値の最大値は1である。値が1に近い方が音声の明瞭的であるとされる。評価時には、Vocoder パラメータから WORLD の合成アルゴリズムを利用することによって再構成された音声を参照音声として使用する。

## 5.2 AEC の学習結果

本節では、AEC の学習を行なった結果を示す。学習した AEC に対して、被験者に聞いてもらったテストデータ取得時の各音声ファイルから得た Vocoder パラメータをテストデータとして検証を行なった。Vocoder パラメータを入力とし、出力のパラメータから WORLD の合成アルゴリズムを適用することによって、再構成音声を取得。再構成音声の種類は数字 10 種類(1~10)と単位 6 種類(個・枚・本・冊・台・番)の組み合わせであり、全 60 種類であった。得られた再構成音声に対して、主観評価と客観評価を行なった。主観評価として数字の認識度、単位の認識度の評価、音声の品質を評価者 3 名の方に協力してもらい認識度の結果を評価した。主観評価の結果として、数字の認識度 98.88%, 単位の認識度 97.22% となった。また、音声の品質評価の結果は、平均 4.72 となった。客観評価の ESTOI によって、再構成音声を評価した結果平均:0.8691 となった。これらの評価結果から AEC から出力された Vocoder パラメータは、人がほとんど正しく認識することができる精度のパラメータであることが示された。しかし、数字や単位の認識どちらも 100% にはなっていない。このような結果が得られた理由として、数字や単位で似ている文字(「9」と「10」や「枚」と「台」)が存在しているため、少しの聞き逃しなどによって、誤認識が発生してしまった点や、再構成音声にノイズが乗ってしまい、聞き取りにくかった点が存在していることが考えられる。

## 5.3 DNN の学習結果

本節では、各被験者ごとの脳活動を用いて DNN の学習を行なった結果を示す。被験者ごとの最良の結果が得られた際の検証時の損失値、評価値及び、それらのモデルを用

いて、音声の再構成を行なった際の平均 ESTOI を表 1 に示す。

表 1 各被験者ごとのモデル学習結果

| 被験者 | 損失値 ( $MAE * 10^5$ ) | 評価値 ( $R^2$ ) | 平均 ESTOI |
|-----|----------------------|---------------|----------|
| 1   | 54,420               | -0.051        | -0.0023  |
| 2   | 55,600               | -0.081        | 0.0026   |
| 3   | 59,104               | 0.0261        | -0.0062  |
| 平均  | 56,374               | -0.0353       | -0.001   |

表 1 から、ESTOI がかなり小さいことがわかる。これは元の音声と全く類似していない音声が生産されたことを示している。そのため、主観評価は実施しなかった。被験者ごとの各エポックごとの訓練時・評価時の損失値(loss)・評価値(evaluation)を図 7 に示す。

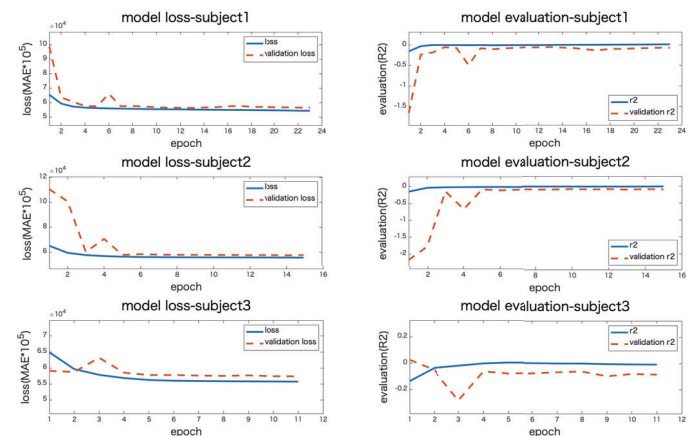


図 7 各エポックにおける損失値および評価値

学習時の訓練時・評価時の損失値・評価値を確認してみると、これらの値が訓練時と評価時で乖離しておらず、似通った値を取っていることが確認できた。そのため、提案した特徴量及び、モデル構造で学習した結果、モデルが未学習であることが確認できた。

## 6. まとめ

本研究では、非侵襲型の脳磁図(MEG)を用いて聴覚刺激を与えた際の脳活動の計測を行なった。また、計測された脳活動から高品質な聴覚刺激を再構成することを目標に非侵襲型の脳活動測定手法を用いることによる聴覚刺激の再構成についてシステムを提案した。その結果、ターゲットパラメータを圧縮した特徴に落とし込むために利用した AEC では、ある程度良好な結果(主観評価:数字の認識度 98.88%, 単位の認識度 97.22%, 音声の品質評価平均 4.72, 客観評価:ESTOI 平均 0.8691)を得ることができた。しかし、提案した DNN のモデルでは、うまく学習を行うことができなかった。被験者ごとの脳活動を用いて学習を行なった結果として、ESTOI が平均-0.001 となった。これは脳活動から再構成された音声は全く異なったものであることを



示している。

今後の課題として、DNN の構造の見直し、AEC の精度の向上が挙げられる。今回の提案した手法では、DNN がうまく学習が行われておらず、未学習の状態であることが確認できた。そのため、今後は、データ拡張、DNN の構造の見直し、学習のスケジューリング、特徴量の増加などを行い、再構成される音声の精度を向上させていきたい。AEC の学習結果では、主観評価による数字や単位の認識どちらも 100%にはなっていない。AEC によって、再構成された音声を聞いてみると、ノイズが乗ってしまっている箇所などが存在しているため、音声の主観評価の精度を下げていられる。そのため、学習量を増やすことや、正則化係数をチューニングすることによって、より汎化性能を向上させていき、ノイズが乗っていない音声を再構成することができる AEC の学習を行なっていく必要があると考える。

## 参考文献

- [1] Zhang, X., Yao, L., Sheng, Q. Z., Kanhere, S. S., Gu, T. and Zhang, D.: Converting Your Thoughts to Texts: Enabling Brain Typing via Deep Feature Learning of EEG Signals, *CoRR*, Vol. abs/1709.08820 (online), available from <http://arxiv.org/abs/1709.08820> (2017).
- [2] Pandarinath, C., Nuyujukian, P., Blabe, C. H., Sorice, B. L., Saab, J., Willett, F. R., Hochberg, L. R., Shenoy, K. V. and Henderson, J. M.: High performance communication by people with paralysis using an intracortical brain-computer interface, *eLife*, Vol. 6, p. e18554 (online), DOI: 10.7554/eLife.18554 (2017).
- [3] Akbari, H., Khalighinejad, B., L. Herrero, J., Mehta, A. and Mesgarani, N.: Towards reconstructing intelligible speech from the human auditory cortex, *Scientific Reports*, Vol. 9, p. 874 (online), DOI: 10.1038/s41598-018-37359-z (2019).
- [4] Anumanchipalli, G. K., Chartier, J. and Chang, E. F.: Speech synthesis from neural decoding of spoken sentences, *Nature*, (online), DOI: 10.1038/s41586-019-1119-1 (2019).
- [5] Wang, J., Kim, M., Hernandez-Mulero, A. W., Heitzman, D. and Ferrari, P.: Towards decoding speech production from single-trial magnetoencephalography (MEG) signals, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3036–3040 (online), DOI: 10.1109/ICASSP.2017.7952714 (2017).
- [6] 宮内 哲：脳を測る：一改訂ヒトの脳機能の非侵襲的測定一，心理学評論，Vol. 56, No. 3, pp. 414–454 (オンライン)，入手先 (<https://ci.nii.ac.jp/naid/130007436988/>) (2013).
- [7] 藤平潤一，河合 淳，樋口正法，足立善昭，小山大介，尾形久直，上原 弦：MEG 用液体ヘリウム再凝縮装置の運転，低温工学，Vol. 49, No. 7, pp. 379–384 (オンライン)，DOI: 10.2221/jcsj.49.379 (2014).
- [8] 菅田陽怜，平田雅之：脳磁図 (MEG) を利用した脳機能計測とその応用，理学療法学，Vol. 43, No. 6, pp. 514–519 (2016).
- [9] Asano, E., Juhasz, C., Shah, A., Muzik, O., Chugani, D., Shah, J., Sood, S. and T Chugani, H.: Origin and Propagation of Epileptic Spasms Delineated on Electro-corticography, *Epilepsia*, Vol. 46, pp. 1086–97 (online), DOI: 10.1111/j.1528-1167.2005.05205.x (2005).
- [10] Schfer, P. J., Corona-Strauss, F. I., Hannemann, R., Hillyard, S. A. and Strauss, D. J.: Testing the Limits of the Stimulus Reconstruction Approach: Auditory Attention Decoding in a Four-Speaker Free Field Environment, *Trends in Hearing*, Vol. 22, p. 2331216518816600 (online), DOI: 10.1177/2331216518816600 (2018).
- [11] Khalighinejad, B., Cruzatto da Silva, G. and Mesgarani, N.: Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech, *Journal of Neuroscience*, Vol. 37, No. 8, pp. 2176–2185 (online), DOI: 10.1523/JNEUROSCI.2383-16.2017 (2017).
- [12] Braiman, C., Fridman, E. A., Conte, M. M., Voss, H. U., Reichenbach, C. S., Reichenbach, T. and Schiff, N. D.: Cortical Response to the Natural Speech Envelope Correlates with Neuroimaging Evidence of Cognition in Severe Brain Injury, *Current Biology*, Vol. 28, No. 23, pp. 3833 – 3839.e3 (online), DOI: <https://doi.org/10.1016/j.cub.2018.10.057> (2018).
- [13] Sonobe, R., Takamichi, S. and Saruwatari, H.: JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, *CoRR*, Vol. abs/1711.00354 (online), available from (<http://arxiv.org/abs/1711.00354>) (2017).
- [14] Hinton, G. E. and Salakhutdinov, R. R.: Reducing the Dimensionality of Data with Neural Networks, *Science*, Vol. 313, No. 5786, pp. 504–507 (online), DOI: 10.1126/science.1127647 (2006).
- [15] MORISE, M., YOKOMORI, F. and OZAWA, K.: WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications, *IEICE Transactions on Information and Systems*, Vol. E99.D, No. 7, pp. 1877–1884 (online), DOI: 10.1587/transinf.2015EDP7457 (2016).
- [16] Morise, M.: D4C, a band-a-periodicity estimator for high-quality speech synthesis, *Speech Communication*, Vol. 84, pp. 57 – 65 (online), DOI: <https://doi.org/10.1016/j.specom.2016.09.001> (2016).
- [17] Morise, M.: Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals, *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20–24, 2017*, pp. 2321–2325 (online), available from (<http://www.isca-speech.org/archive/Interspeech.2017/abstracts/0068.html>) (2017).
- [18] Warden, P.: Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition, *CoRR*, Vol. abs/1804.03209 (online), available from (<http://arxiv.org/abs/1804.03209>) (2018).
- [19] mozilla: Common Voice, , available from (<https://voice.mozilla.org/ja>) (accessed 2019-07-3).
- [20] Lu, L., Shin, Y., Su, Y. and Karniadakis, G. E.: Dying ReLU and Initialization: Theory and Numerical Examples, *ArXiv*, Vol. abs/1903.06733 (2019).
- [21] Jensen, J. and Taal, C. H.: An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 11, pp. 2009–2022 (online), DOI: 10.1109/TASLP.2016.2585878 (2016).