2017 年度

卒 業 論 文

# Deep Non-negative Matrix Factorization を用いた脳磁界データから音声の復元

神戸大学工学部情報知能工学科

矢野 彩緒里

指導教員　　滝口 哲也　教授

2018 年　2 月 13 日

# Deep Non-negative Matrix Factorization を用いた脳磁界データから音声の復元

矢野 彩緒里

## 要 旨

障害者支援を主な目的に，近年，脳活動を用いて機械制御をおこなうブレイン・コンピュータ・インターフェイス (BCI) の利用が期待されている．しかし，現在主に利用されている BCI は，予め用意された選択肢の中からユーザの意思を判別するものであり，自由度の高い意思伝達はおこなわれない．そこで，ユーザの意思をより汎用的に他者に伝達するために，脳活動から直接音声を復元する意思伝達手法の開発を試みた．

本研究では，Deep learning と Non-negative Matrix Factorization (非負値行列因子分解：NMF) を組み合わせた Deep Non-negative Matrix Factorization (Deep NMF) を用いた．本手法の Deep NMF モデルは脳磁界データから得られる特徴量を入力，聴取させた音声の音響特徴量を出力としている．モデルの学習を教師ありでおこない，未知の脳磁界データを入力すると，復元音声が出力されるというシステムになっている．特徴量と活性化関数は全て非負で実装しており，出力層には重み更新をしない辞書行列を用いているので，Exemplar-based な音響特徴量が得られる．また，出力層の 1 層前では辞書の基底毎の重み係数 (アクティビティ) が出力される．

以上のモデル構成により Exemplar-based な変換をすることで，より自然性の高い音声を復元することを示す．

# 目次

# 第 1 章

# 序論

## 1.1 研究背景

近年，音声による機械制御の技術開発が盛んに行なわれており，我々の日常生活の様々な場面で利用されている．しかし一方で，発話困難な障害者など，音声認識による機械制御を有効に活用することができないユーザも一定数存在する．そこで，脳活動を用いて機械制御をおこなうブレイン・コンピュータ・インターフェイス (BCI) の利用が期待される．

これまでの BCI 開発では，ユーザが注意を向けた低頻度刺激に対してのみ出現する誘発反応 (P300) によって，限定的な意思伝達をおこなう “P300 スペラー型 [1]” の開発例が多い．しかしながら，P300 スペラー型では予め用意された選択肢の中からユーザの意思が判別されるため，自由度の高い意思伝達はできない．本研究では，ユーザの意思をより汎用的に他者に伝達するために，脳活動から直接音声を復元する意思伝達手法の開発を試みた．

これまでにも，脳活動からの音声復元を試みた例は存在する．例えば，fMRI データからの音声復元 [2] では，音声スペクトログラムの概形の推定に成功しているが，完全な音声の復元には至っていない．音声の時間変化をより詳細に捉えるには，より時間分解能の高い脳機能計測が有効であると思われる．一方，電極を細胞に差し込んで細胞内電位を計測する Local field potential (LFP) [3] は，時間分解能に優れ，高感度での脳信号計測が可能であるが，侵襲的であり人体に負担がかかる．そのため，日常的な手段としては現実的ではない．本研究では，時間分解能に優れ人体に不可逆的な影響を与えない非侵襲的な脳磁界計測を用いた．

脳磁界データから音声復元を実現するにあたって，非負値行列因子分解 (Non-negative

Matrix Factorization: NMF) [4] を用いた信号変換に着目した．我々はこれまで，従来の統計的手法とは異なる，スパース表現に基づく NMF を用いた Exemplar-based 声質変換手法 [5] を提案してきた．この手法では，次の手順で声質変換を実現する．(1) 入力・出力話者それぞれにおいてパラレルな辞書を用意する， (2) 入力音声を話者性に依存した辞書行列とその辞書の基底毎の重み係数 (アクティビティ) に分解する，(3) 入力辞書の基底を出力辞書の基底に置き換える．

この手法において，入力・出力辞書は少量データによって作成することが望ましい．しかしながら脳磁界データには多く雑音が含まれ，同じ刺激を与えた場合でも試行によって得られる脳磁界波形は大きく異なる．そのため，多くのデータを学習に用いたほうが，刺激に対する脳活動の特徴を捉えることができると考える．また，大量データを用いた信号変換において，近年では Deep learning の利用が盛んである．

そこで本研究では音声出力にのみ NMF を用い，学習に Deep learning を用いることで， Exemplar-based であるため音声の再現性の高い NMF と，データ量を生かしてより精度を上げることが期待される DNN 双方の利点を生かせることを期待し， Deep Non-negative Matrix Factorization のモデル構築を試みた．

## 1.2 本論文の構成

本論文は全 6 章で構成される．まず第 1 章では，研究の背景と本研究の目的について述べた．第 2 章では，非侵襲的手法における脳磁界計測システムの位置付けとそのシステムの概要について述べる．第 3 章では，脳磁界データと音声の特徴量抽出法について述べ，第 4 章では，Deep learning の基礎と NMF 声質変換手法について紹介したのち， Deep Non-negative Matrix Factorization のモデル構成を述べる．第 5 章では音声復元実験の結果と考察を述べる．最後に，第 6 章では本論文のまとめを行う．

# 第 2 章

# 脳磁界計測

## 2.1 脳機能計測手法の比較と脳磁界計測の利点

従来の脳機能計測法は，開頭して脳の構造を調べたり，直接電気的反応を観察するために脳に電極を差し込んだりして人体に不可逆的な影響を与える可能性のある侵襲計測 (invasive measurement) が主であり，技術的観点と倫理的観点から人の脳活動を測定することは困難であった．しかし近年では，人体に不可逆的な影響を与えない非侵襲計測技術 (noninvasive measurement) の進歩により，人に対してより安全な脳活動の計測を行うことが可能になった．非侵襲脳機能計測には体外から与えた放射線や磁場等の反射や透過を計測する受動 (passive) 的手法や，脳活動に伴う神経発火による発火電位や脳磁場等を計測する能動 (active) 的手法などが存在する [6, pp. 180-183].

代表的な手法として，X 線 CT (Computed tomography : コンピュータ断層画像)，MRI (Magnetic resonance imaging : 核磁気共鳴画像)，PET (Positron Emission Tomography : 陽電子崩壊断層画像)，fMRI (機能的核磁気共鳴画像: functional Magnetic Resonance Imaging)，NIRS (Near Infrared Spectrophotometory : 近赤外分 光法)，EEG (Electroencephalography : 脳波計測)，MEG (Magnetoencephalography : 脳磁界計測) などが挙げられる [7, pp. 136-139].

以上で挙げた非侵襲計測の中でも脳磁界計測は人体に対してより安全な手法である．X 線 CT は放射線を照射する，PET は放射性同位元素を体内に注入する，f-MRI は高磁場を生体に加える手法であり，完全に非侵襲とは言い切れないからである．一方で脳磁界計測は，脳の神経活動によって頭外に発生する微弱な磁場を非常に高精度なセンサで検出する手法であり，完全非侵襲な手法と言える．また，脳磁界計測は f-MRI のように機器が騒音を発することはなく，非常に静粛である．このため本研究のような聴覚を対象とした

計測に最適であると言える．

また， 脳機能計測手法の特徴を述べる上で重要な概念として，時間分解能と空間分解能がある．時間 / 空間分解能とは，近接した時間 / 場所における脳活動の違いを区別する能力のことであり，Fig.2.1 は代表的な非侵襲計測手法の時空間分解能について示している．

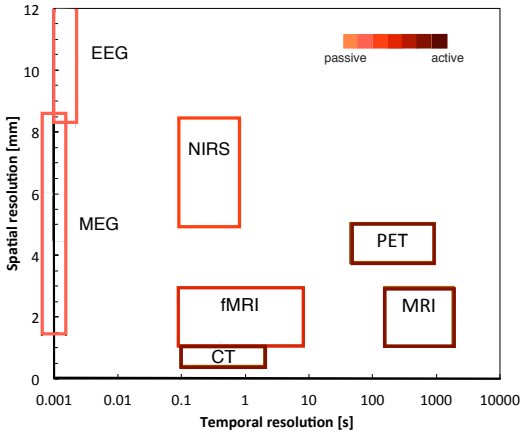



Fig. 2.1. Time and spatial resolution in each method of measurement.

MEG は，時間分解能，空間分解能ともに優れており，特に時間分解能に関しては ms のオーダを持ち，他の手法と比べても大変優れていることがわかる．MEG の欠点としては超電導を用いて微弱な磁界を測定するための大掛かりな設備，システムが必要であるため，可搬性がなく，導入からランニングコストまで非常に高価である点がある．

MEG と同じく時間分解能が高く完全非侵襲，かつ可搬性に優れる装置に EEG が存在する．しかし EEG は頭皮上から脳神経電位計測を行う手法であり，頭蓋骨や頭皮などの電気伝導率が異なるものを介して計測するため，観測電場に歪みが生じて電位発生源の位置に誤差が生じる．よって神経の活動をそのまま捉えることができない．以上のことから本研究のように音声やその高次機能の時空間特徴を扱う場合には，脳磁界計測は適当な手法であると言える．

## 2.2　MEG システムの概要

生体磁気の強さは，地上に存在する様々な磁気と比べると非常に弱い．とりわけ，MEG で計測対象の脳磁気の強さは $10^{-12}$ T 以下であり，地磁気の 1 億分の 1，都市環境ノイズの 100 万分の 1 以下である．(Fig.2.2).

このように大変微弱な磁気信号を検出するためには，環境雑音を遮断するための磁気シールドルームや高感度なセンサが必要となる [6, pp. 113-118] [7, pp. 15-41]．MEG

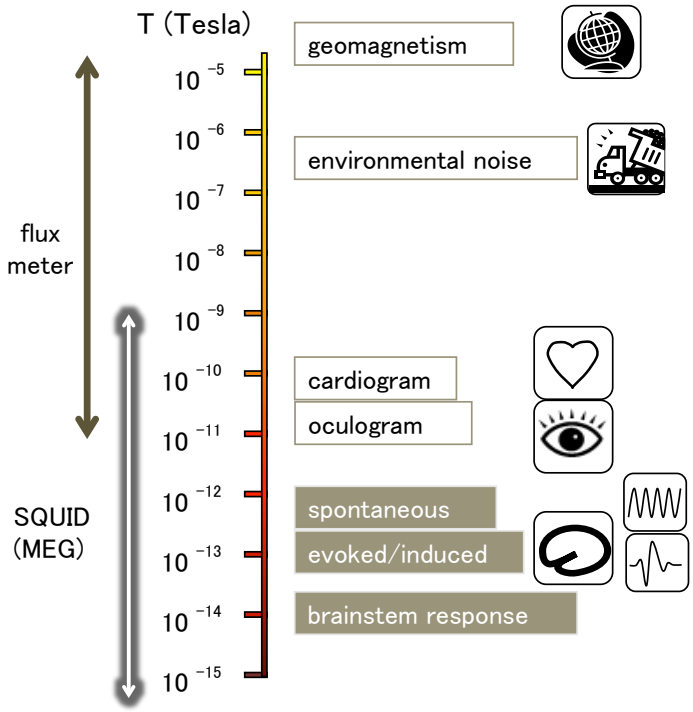



Fig. 2.2. Intensity of biomagnetism.

システムに導入されている高感度磁束計の多くは，ジョセフソン接合による SQUID (Superconducting quantum interference device：超電導量子干渉素子）磁束計であり，磁束の変化を電圧の変化として捉えることができる．加えて，磁場の効率的な検出のためにセンサコイルが付属する．

本研究で用いる 122 ch 全頭型脳磁界計測システム (Neuromag - 122$^{TM}$: Neuromag, Ltd.) (Fig.2.3) では，Fig.2.4 に示すとおり，平面型グラジオメータコイル (planar coil) が交差するように頭の接平面方向に 61 組配置されており，計 122 個のコイルによる脳磁界が得られる．

これらのコイルが計測する脳磁界は，主に脳溝の脳神経活動を表している．頭蓋内を球であると仮定した場合，頭の表面に垂直な方向の脳磁界は互いにより打ち消されるため，MEG で検出されるのは，頭の接平面方向の磁界成分である．そのため，脳の表層より発生する頭蓋垂直平面方向は観測されず，一方で脳の部位を分ける脳溝は溝の表面に対して垂直方向の磁界が発生するため，MEG 信号として捉えることが可能である．

Fig. 2.3. Neuromag - 122$^{TM}$ , Neuromag Ltd, Finland.

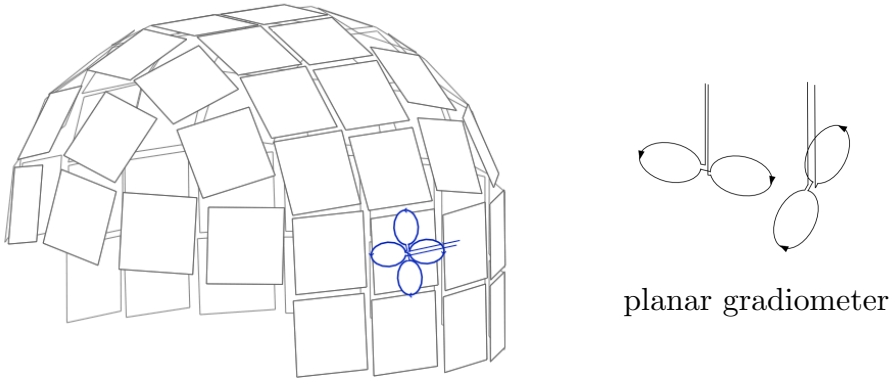



Fig. 2.4. Coil location of Neuromag - 122$^{TM}$ .

## 2.3　脳の構造と機能地図

大脳皮質は，大きく 4 つの部位に分けられる．前から順に前頭葉 (frontal lobe) ，側頭葉 (lateral lobe) ，頭頂葉 (parietal lobe)，後頭葉 (occipital lobe) と呼ばれ，前頭葉と側頭葉，頭頂葉の間にはそれぞれ大きな溝が存在する [6, pp. 19-21]．2.5 に主な脳機能

を示す．

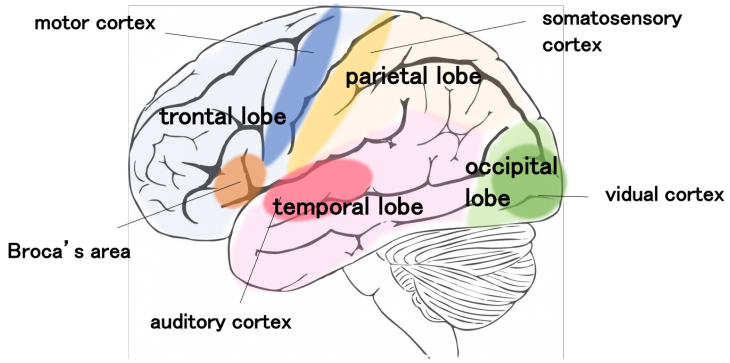



Fig. 2.5. Brain function map.

前頭葉と側頭葉の間の大きな溝であるシルヴィウス溝の奥に，音の最も基本的な知覚処理を行う 1 次聴覚野が存在する．聴覚野の後ろには言語の認知に関連するヴェルニッケ野が存在し，シルヴィウス溝を挟んで向かい側の前頭葉には，左半球に言語の発話に関連するブローカ野が存在する．後頭葉は主に視覚野の領域であり，後頭極を 1 次視覚野とし，高次視覚野が広がる．前頭葉と頭頂葉の間にある大きな溝は中心溝と呼ばれ，その前壁には筋肉をコントロールする運動野が，後壁には皮膚感覚などの身体感覚に関わる体性感覚野がある．このように主要な感覚皮質は脳溝内に存在しており，2.2 節で述べたように MEG 計測の対象となる．

# 第 3 章

# 特徴量抽出

本章では，モデルに入力する脳磁界データと音声データそれぞれの特徴量抽出法について述べる．

## 3.1 脳磁界データの特徴量抽出

### 3.1.1 眼電除去

MEG により計測される脳磁界には，脳神経群の自発的な発火によって観測される自発脳磁界と，何らかの刺激に応じた脳神経群の一斉発火によって観測される誘発脳磁界に区分される．誘発脳磁界は，2 章で述べたように，自発脳磁界や他の生体磁気よりも弱く，通常はそれらの生体磁気や外部環境の磁気がノイズとなるため観測が困難である．本研究では，眼球運動により発生する脳磁界がノイズとして影響することを防ぐために，以下の手法でアーティファクトの除去を行った．

まず，磁界計測によって得られた raw データに対し，独立成分分析 (Independent component analysis : ICA) を適用する．時刻 t における m 次元脳磁界データ $\boldsymbol{x}(t)$ は，n 個の独立的な活動源信号 $\boldsymbol{s}(t)$ が混合して存在するものとして仮定すると，分離行列 $\boldsymbol{A}$ により以下の関係が得られる．

$$\boldsymbol{s}(t) = \boldsymbol{A}\boldsymbol{x}(t) \tag{3.1}$$

分離行列 $\boldsymbol{A}$ を推定し，得られた分離信号 $\boldsymbol{s}(t)$ から眼球運動と思われる信号だけを取り除いたものを $\boldsymbol{y}(t)$ とし，以下の式を用いて復元する．

$$\boldsymbol{X}(t) = \boldsymbol{A}^{-1}\boldsymbol{y}(t) \tag{3.2}$$

### 3.1.2 連続ウェーブレット変換

脳磁界データの時間周波数特徴量の抽出法として，連続ウェーブレット変換 (Continuous Wavelet Transform : CWT) を用いた． ウェーブレット変換は，小さい波 ( wavelet ) を拡大縮小、平行移動して足し合わせることで，与えられた入力の波形を表現しようとする手法であり．ある信号が与えられた時に、時間的に局在した周波数成分を知りたい場合，フーリエ解析においては，サイン波とコサイン波を拡大縮小して足し合わせることで入力を表現しようとするが，波が局在化していないため，時系列の情報が失われてしまう．ウェーブレット変換では，周波数に合わせてウェーブレットの幅が変化するので，周波数解像度が高くなるため，脳磁界データの特徴量として適当であると言える．

CWT 関数 $W$ は以下の式 (1) で示される．

$$W(a,b) = \frac{1}{\sqrt{a}} \int x(t)\psi(\frac{t-b}{a})dt \tag{3.3}$$

ここで， $x(t)$ は脳磁界の時系列波形である．$\psi(t)$ はマザーウェーブレットであり，本実験では Morlet ウェーブレットを用いた．$a$ はスケール，$b$ は時間シフトを表すマザーウェーブレットのパラメータである．

本研究では脳磁界の raw データに眼電除去を施し，CWT をかけたものを脳磁界データの特徴量として用いる．

## 3.2 音響特徴量抽出

### 3.2.1 STRAIGHT による特徴量抽出

音響特徴量の抽出には，音声変換合成方式 STRAIGHT (Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram) [8] を用いた．STRAIGHT は Vocoder と呼ばれる音声を分析し再合成するシステムの一種であり, その品質の高さから，声質変換や音声合成システムで広く用いられている．

まず, 入力音声は基本周波数 (Basic Frequency), スペクトル包絡 (Spectrum envelope), 非周期成分 (Aperiodic component) に分離される．

基本周波数は，声の高さに相当するものである．スペクトル包絡は，音声の音色を表す要素である．STRAIGHT スペクトルは，短時間フーリエ変換のみならず，時間周波数分析，wavelet 分析，spline 関数理論など，様々な理論に基づいて推定される．人の声においては，音韻情報またはその人の話者性を表すものとされており，話者変換のタスクが主な声質変換においては重要な要素である．非周期成分は, スペクトルの上側包絡と下側包

絡の差から求められ, 発話音源に含まれる雑音成分の大きさを表す.

### 3.2.2　メル周波数ケプストラム係数

従来の音声認識において，メル周波数ケプストラム係数（ Mel Frequency Cepstral Coefficient： MFCC ）が音声特徴量として用いられてきた．Mel とはメル尺度を表し, 人間の音の高さの知覚特性から定められている．具体的には，1,000 Hz, 40 dB の純音を基準の音 1,000 mel とし，マグニチュード測定法などで測定された基準音から 2 倍の高さに知覚される音を，2,000 mel，1/2 の高さに知覚される音を 500 mel と定めた.

MFCC ではメル尺度フィルタバンクの短時間対数エネルギー出力系列に対して，離散コサイン変換 (Discrete Cosine Transform： DCT) を適用しケプストラムが得られる．そして音声のスペクトル包絡成分に対応する低次ケプストラムのみを抽出し，特徴量とする.

本研究において, 提案手法の特徴量は STRAIGHT から抽出されたスペクトル包絡のメルケプストラムを用いる.

# 第 4 章

# 学習モデル

本研究では，学習に Deep learning モデルを用いている．4.1 節で Deep learning の基礎について述べ，4.2 節では NMF を用いた声質変換手法について述べる．4.3 節では本研究の提案手法に用いた Deep NMF モデル構造について述べる．

## 4.1 Deep learning の基礎

### 4.1.1 Deep learning モデルの基本構造

Deep Learning とは，機械学習の 1 種である「ニューラルネットワーク（ Neural Networks ）」の階層を深めたアルゴリズムである．ニューラルネットワークは，生物の脳の神経細胞（ニューロン）をモデルとしている．Deep learning の処理概要を，Fig.4.1 に示す．

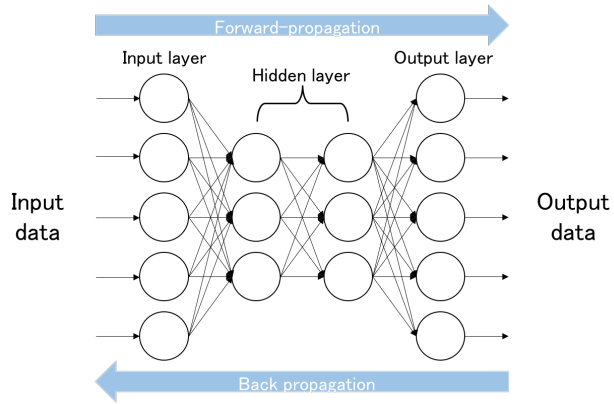



Fig. 4.1. Deep learning model.

まず，入力データとその入力に対する出力の教師データのペアを複数用意する．入力データは入力層，隠れ層，を経て出力層へと流れ，最後の出力から推論結果が得られる．学習では，推論された出力データと教師データの誤差を最小化するよう重み更新を行い，最適な重みを求める．

Fig.4.2 に示すように，ニューラルネットワークにおいて第 $l+1$ 層の $k$ 次元目の出力 $x_{(l+1),k}$ は次式のように定義される．

$$v_{l,k} = \sum_j w_{l,j,k} x_{l,j} + b_l \tag{4.1}$$

$$x_{(l+1),k} = a_l(v_{l,k}) \tag{4.2}$$

ここで，第 $l$ 層において $x_{l,j}$ は $j$ 次元目の出力，$w_{l,j,k}$ は第 $l$ 層の $j$ 次元目から第 $l+1$ 層の $k$ 次元目の重み，$b_l$ はバイアス，$a_l$ は活性化関数を示す．ニューラルネットワークの回帰モデルにおいて，主な活性化関数に relu，sigmoid などがある．

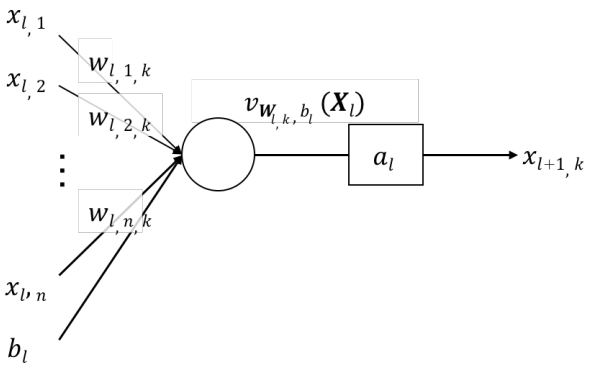



Fig. 4.2. Perceptron.

Deep learning における重みの更新には勾配降下法 (Gradient descent) を用いる．回帰モデルにおける損失関数には式 (4.3) で表される最小二乗誤差 (Minimum square error) を用いることが多い．

$$E(\mathbf{w}) = \frac{1}{2} \sum_n \{t_n - \mathbf{w}^T \mathbf{y}_n\} \tag{4.3}$$

$t_n$ は $n$ 次元目の教師データ，$y_n$ は出力を示す．$\frac{\partial E_n}{\partial v_j}$ を $\delta$ とおくと，目的の勾配は

$$\frac{\partial E_n}{\partial w_l, i, j} = \delta_{l,j} x_{(l-1),i} \tag{4.4}$$

と計算され，この $\delta_{l,j}$ は上位の $\delta_{(l+1),k}$ が与えられれば

$$\delta_l, j = \sum_k \delta_{(l+1),k} \left( w_{(l+1),j,k} a'(v_{l,j}) \right) \tag{4.5}$$

と求められる．つまり，出力層から辿っていけば任意の層の $\delta$ を求めることができる．これを誤差逆伝搬 (Back propagation) という．

### 4.1.2 Long Short-Term Memory

再帰型ニューラルネットワーク (Recurrent Neural Networks : RNN) は，音声やテキストなどの系列データを扱えることで，データ間の依存関係を学習できるネットワークである．RNN は内部に有向閉路を持つネットワークで，時系列データの過去の情報を一時的に保管できる．

RNN の中間層を LSTM (Long Short-Term Memory) に置き換えることでさらなる精度向上を図ることができる．LSTM は入力ゲート，出力ゲート，忘却ゲートを持つ．LSTM ブロックの概念図を Fig.4.3 に示す [9].

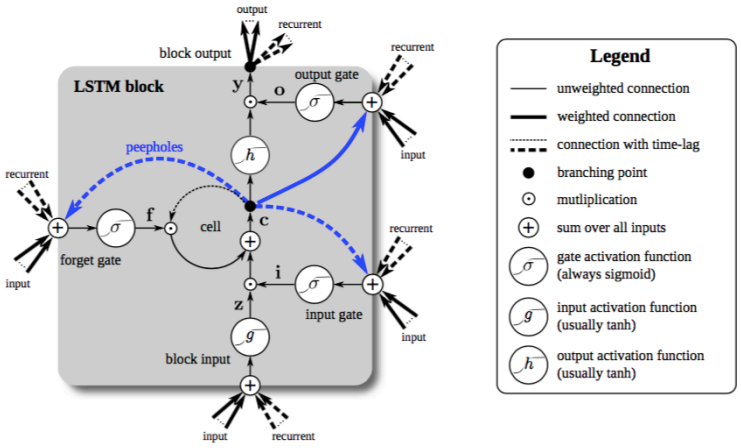



Fig. 4.3. LSTM block as used in the hidden layers.

LSTM を用いることで，必要に応じて誤差を消失させることなく過去の情報を長期保存することができるようになった．

## 4.2 NMF を用いた声質変換

スパース表現の考え方において，与えられた信号は少量の学習サンプルや基底の線形結合で表現される．

$$\mathbf{v}_l \approx \sum_{j=1}^{J} \mathbf{w}_j \, \mathbf{h}_{j,l} = \mathbf{W}_{\mathbf{h}_l} \tag{4.6}$$

式 (4.6) において，$\mathbf{v}_l$ は観測信号の $l$ 番目のフレームにおける $D$ 次元の特徴量ベクトルを表す．$\mathbf{w}_j$ は $j$ 番目の基底を表し，$\mathbf{h}_{j,l}$ はその結合重みを表す．基底を並べた行列 $\mathbf{W} = [\mathbf{w}_1 ... \mathbf{w}_J]$ を “辞書” と呼び，重みを並べたベクトル $\mathbf{h}_l = [\mathbf{h}_{1,l} ... \mathbf{h}_{J,l}]^T$ を “アクティビティ” と呼ぶ．このアクティビティベクトル $\mathbf{h}_l$ がスパースであるとき，観測信号は重みが非ゼロである少量の基底ベクトルのみで表現されることになる．フレーム毎の特徴量ベクトルを並べて表現すると式 (4.6) は二つの行列の内積で表される．

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \tag{4.7}$$

$$\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_L],\ \mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_L] \tag{4.8}$$

ここで $L$ はフレーム数を表す．また，本手法では $\mathbf{W}$ は固定される．

本手法の概要を Fig.4.4 に示す．

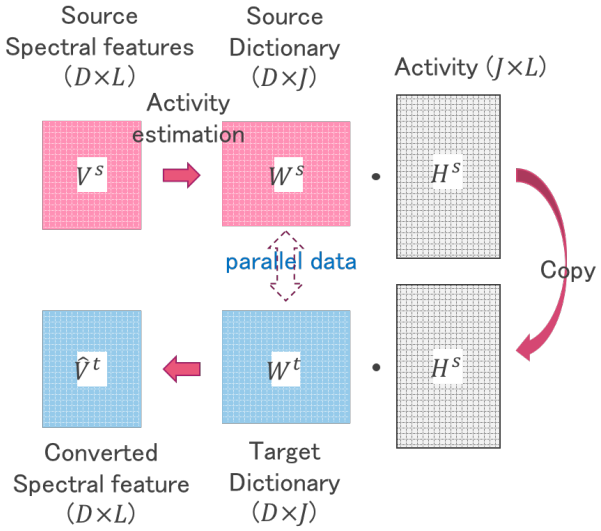



Fig. 4.4. NMF-voice conversion.

$\mathbf{V}^s$ は入力話者の音響特徴量，$\mathbf{W}^s$ は入力話者辞書，$\mathbf{W}^t$ は出力話者辞書，$\hat{\mathbf{V}}^t$ は変換された音声特徴量，$\mathbf{H}^s$ は入力特徴量から 推定されるアクティビティを表す．$D$，$J$ はそれぞれ特徴量の次元数，辞書の基底数である．この手法では，パラレル辞書と呼ばれる入力話者辞書 $\mathbf{W}^s$ と出力話者辞書 $\mathbf{W}^t$ からなる辞書の対を用いる．入力音響特徴量 $\mathbf{V}^s$ を，話者性に依存した辞書行列 $\mathbf{W}^s$ とアクティビティ $\mathbf{H}$ に分解し，入力辞書の基底を出力辞書の基底 $\mathbf{W}^t$ に置き換えることで，出力音響特徴量 $\mathbf{V}^t$ を得る．

## 4.3 提案手法

### 4.3.1 音声復元手法

脳磁界データからの音声復元システムの概要を Fig.4.5 に示す.

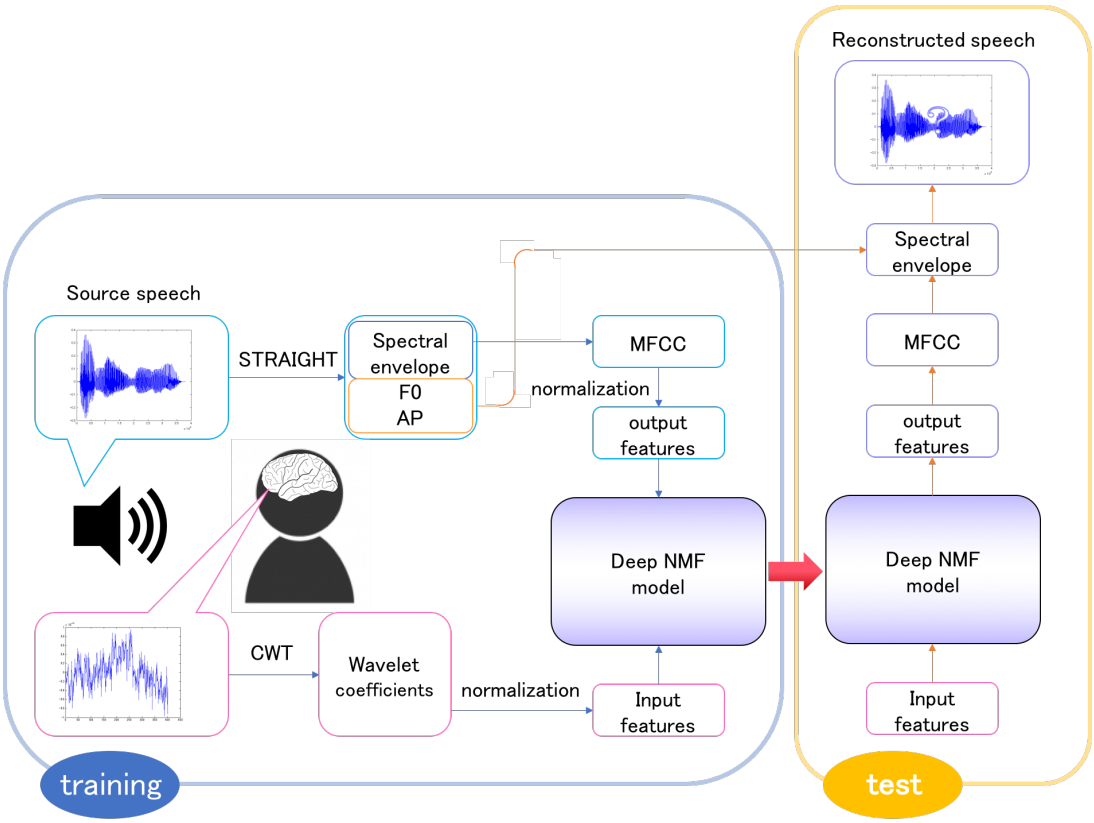



Fig. 4.5. Speech-reconstruction system.

はじめに，音声と，音声聴取時の脳磁界波形のパラレルデータを用意する．得られた脳磁界波形には眼電除去を施したのち，CWT にかけることによって時間周波数特徴量を取り出す．さらに正規化したものを入力の脳磁界データ特徴量とする．音声は STRAIGHT 抽出によりスペクトル包絡，F0，非周期成分に分解し，得られたスペクトル包絡から MFCC を推測する．正規化したものを出力の音響特徴量とする．

入力データと教師データのペアによりモデルのパラメータ学習を行う．学習を終えたモデルに未知の脳磁界データ特徴量を入力すると，音響特徴量の推測値が得られる．そこから得られた MFCC からスペクトル包絡を復元し，もとの F0，非周期成分を用いて再合成し，復元音声を得る．

以上の方法で，脳磁界データから復元音声を生成する．

### 4.3.2　Deep Non-negative Matrix Factorization

本研究で用いた Deep Non-negative Matrix Factorization (Deep NMF) のモデル構造を Fig.4.6 に示す.

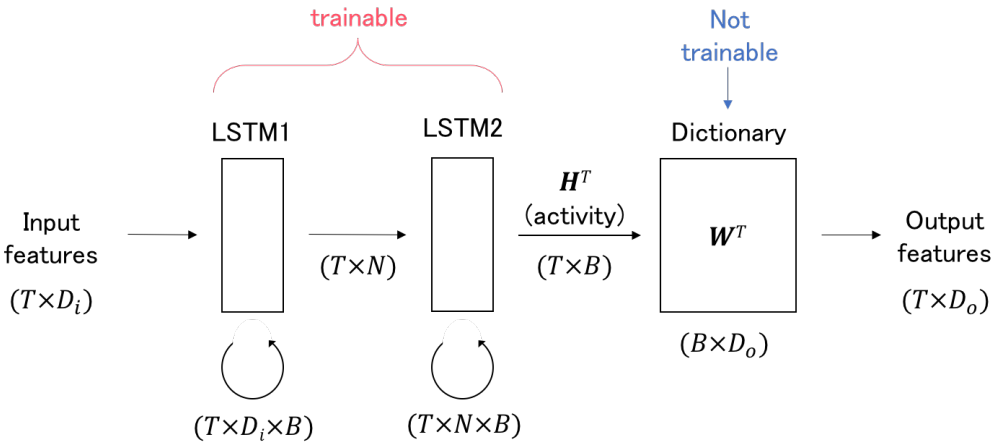



Fig. 4.6. Deep Non-negative Matrix Factorization model.

ここで，$D_i$ は入力データの次元数，$D_o$ は出力データの次元数，$T$ は 1 データあたりのフレーム数，$N$ は隠れ層のユニット数，$B$ は辞書行列の基底数を表す．また，$\mathbf{W}$ は音響特徴量の辞書行列であり，$\mathbf{H}$ はアクティビティを表す．特徴量は全て非負になるよう min-max 正規化をおこなっている．活性化関数には relu を用いている．

入力データは 2 層の LSTM を経て，重み行列が $\mathbf{W}^T$ で固定された全結合の最終層に入力される． 最終層の重みを固定することで，出力 $\mathbf{V}^t$ は式 (4.9) のように，非負の $\mathbf{W}^T$ と $\mathbf{H}^T$ の内積で表されるため，モデル学習により NMF がおこなわれ，アクティビティ $\mathbf{H}$ を得ることができる．

$$\mathbf{V}^t \approx \mathbf{H}^T\mathbf{W}^T \tag{4.9}$$

辞書行列を重みとして固定することによって Exemplar-Based な出力となるため，一般的な Deep learning 回帰モデルと比べ，音質の高い出力が得られることが期待される．

# 第 5 章

# 脳磁界データから音声の復元

## 5.1 実験条件

聴覚健常者 8 名 (男性 7 名，女性 1 名) に対し，3 パターンの単語音声 (“あまぐも”，“いべんと”，“うらない”) の音声を呈示した際の脳磁界データを計測した [10].

音声刺激には，親密度音声データベース (FW03, NTT-AT) に含まれる女性話者音源 (fto) を利用した．刺激呈示時間は 800 ms であり，解析対象の脳磁界データの解析対象も音声呈示から 800 ms とした．

脳磁界計測には，2.2 節で紹介した 122 ch 全頭型脳磁界計測システム (Neuromag - 122$^{\mathrm{TM}}$: Neuromag, Ltd.) を用いた．計測した脳磁界データは 0.03-100 Hz のアナログフィルタを適用した後，サンプリング周波数 400 Hz で A/D 変換をおこなった．

Deep NMF の入力特徴量には，脳磁界計測で得られたデータに眼電除去を施し，左右聴覚野周辺の 2 チャネルに対し，フレーム長 50 ms，1Hz から 25 Hz の範囲で 1 Hz ごとに CWT を適用したもの用いる．1 フレームあたりの次元数は 50 次元であり，フレーム数は 16 ，フレームシフトは用いていない．

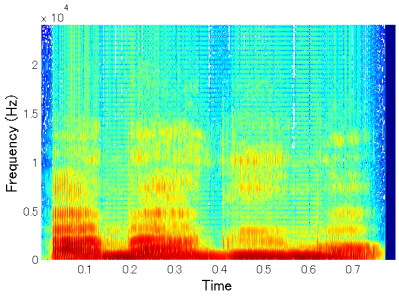
出力の音響特徴量として， 5 ms ごとのスペクトル包絡 1025 次元から推測される MFCC 24 次元を用いた．50 ms ごとにセグメントをとり，240 次元を 1 フレームあたりの特徴量とした．フレーム数は 16 であり，フレームシフトは用いていない．

辞書行列には 3 種類の発話について，出力の音響特徴量と同様に MFCC のセグメントをとったものを用いた．1 基底あたり 240 次元，基底数は 48 である．

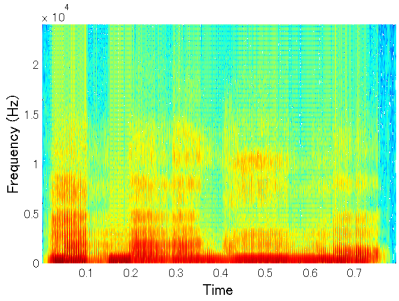
4.3.2 節の Deep NMF モデルにおける LSTM の中間層を 50 とし，エポック数は 1000 とした．最適化アルゴリズムには Adam [11] を使用した．学習データ 960 個，テストデータ 240 個を用いて実験をおこなった．

## 5.2　実験結果

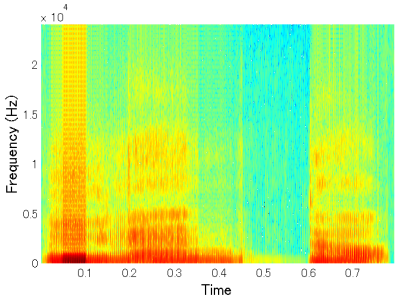
Fig.5.1 に “あまぐも” の原音声と，既知データ (Closed)，未知データ (Open) それぞれそ入力した時の復原音声のスペクトログラムを示す．未知データに対する復元音声スペクトログラムを見ると，元の音声がうまく復元できていないことが目視でも確認できる．特に 450-600 [ms] の区間で大きな誤りが生じている．これは，脳磁界データの信号が微弱であり雑音を多く含み，分類精度が不十分であるため，基底選択に誤りが生じやすく，それに伴い復元音声に誤りが生じてしまったためであると考えられる．このような基底選択の誤りが，テストデータの多くに見られた．




Original




Reconstructed (Closed)

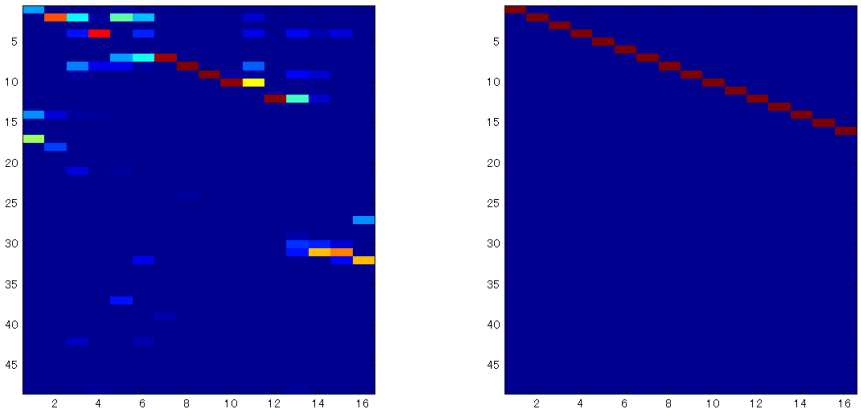



Reconstructed (Open)

Fig. 5.1. Spectrogram of reconstructed sound “amagumo”.

アクティビティ行列の中身を見ることで，入力データに対して正しい辞書基底選択ができているか評価することができる．既知データを入力としたとき，推測されたアクティビティ行列のヒートマップと，正解のアクティビティ行列のペアを Fig.5.2 に示す．

以下の手順においてアクティビティ行列の評価をおこなう．まず，推測されたアクティビティ行列において，時系列ごとに最大値を示す基底を取り出す．正解の基底なら 1，不正解であれば 0 とする．被験者ごとに全データの全ての時系列において同様の評価をし，その平均を被験者ごとの正解率とする．




Predicted activity. Correct activity.

Fig. 5.2. Activity matrix.

被験者のアクティビティの正解率を Table 5.1 に示す.

| | Closed | Open |
|---|---|---|
| Sub.1 | 0.928 | 0.371 |
| Sub.2 | 0.979 | 0.338 |
| Sub.3 | 0.975 | 0.371 |
| Sub.4 | 0.998 | 0.358 |
| Sub.5 | 0.874 | 0.333 |
| Sub.6 | 0.775 | 0.400 |
| Sub.7 | 0.885 | 0.379 |
| Sub.8 | 0.988 | 0.338 |
| Ave. | 0.925 | 0.361 |

Table 5.1. Accuracy of activity matrix.

辞書基底数は 48 であり，chance rate は 2.0 % である．既知データにおいては，全被験者の平均で 90 % の正解率であった．未知データにおける基底正解率は約 36 % と，未知データの半分以下となってしまった．

脳磁界データの微弱な誘発反応を単一試行データから取り出すのは困難であり，うまく分類が行われていないことが理由であると考えられる．精度を向上させるには脳磁界データの雑音除去法，特徴量抽出法のさらなる工夫が必要であると考える．次に，復元音声の

聞き取り実験を主観評価実験としておこなった．比較手法として，重み固定の辞書行列を用いない LSTM での復元音声を用いた．被験者は聞こえた音声を回答し，聞き取りやすさについて 5 段階での評価をおこなった．Fig.5.3 に音韻正解率と聞き取りやすさの評価についてのグラフを示す．Deep NMF モデルは両者において LSTM より高い値を示したが，正解率においては 50 % 弱であり，実用段階に至るためにはさらなる改善が必要である．


Phonetic accuracy.　　Ease of listening.

Fig. 5.3. Subjective Evaluation.

# 第 6 章

# まとめ

本研究では，Deep Non-negative Matrix Factorization を用いて Exemplar-based な変換をすることで，脳磁界データから，より自然な音声を復元することができた．しかしながら，実用段階に達するには基底選択の正解率を向上させることが必要不可欠である．脳磁界データの雑音除去法や特徴量抽出法の改善に加え，データ数を増大させることで精度向上が期待される．よって一単語あたりのデータ個数を増やすためのデータ収録実験が必要である．

また，今回は 3 単語という少ない語彙数で実験を行ったが，より汎用性のあるシステム開発に向けて，データの種類を増やした収録も行いたい．

さらに今後の課題としては，Deep learning モデルの見直し，試行錯誤をすることで実用的な精度の得られるシステムの実現を目指したい．

# 謝辞

本研究に取り組むにあたって，独立行政法人産業技術研究所にて研究活動を行う機会を与えて下さり，また，数多くのご指導，ご助言を賜りました指導教員の滝口哲也教授に厚く御礼申し上げます．

研究に対する心構えや，研究生活で必要な教訓，知識を数多く頂きました有木康雄名誉教授に厚く御礼申し上げます．

本研究の方針や，実験に関わる装置の扱い方に対して多数のご助言を頂きました産業技術総合研究所バイオメディカル研究部門の添田喜治研究員に厚く御礼申し上げます．

本研究の遂行にあたり，研究の方針，論文の執筆に対して多大なるご指導，ご助言を頂きました千葉大学の中川誠司教授に厚く御礼申し上げます．

本研究を行うにあたり，データを提供してくださり多くのアドバイスを頂きました修士課程二回生の字澤志保美先輩に深く感謝致します．

最後に，研究生活を支えて頂いた滝口研究室の皆様，実験にご協力頂いた被験者の方々に心より感謝を申し上げます．

# 参考文献


[1] R. Fazel-Rezai *et al.*, “P300 brain computer interface : current challenges and emerging trends,” Frontiers in Neuroengineering, pp. 1-15, 2012.

[2] R. Santoro *et al.*, “Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns,” Proc. Natl. Acad. Sci. USA, vol. 114 no. 18, pp. 4799-4804, 2017.

[3] M. Yang *et al.*, “Speech reconstruction from human auditory cortex with deep neural networks,” INTERSPEECH 2015, pp. 1121-1125.

[4] D. D. Lee and H. S. Seung, “Algorithms for nonnegative matrix factorization, ” Neural Information Processing System, pp. 556-562, 2001.

[5] R. Aihara *et al.*, “Individuality-Preserving Voice Conversion for Articulation Disorders Based on Non-negative Matrix Factorization,” ICASSP2013, pp. 8037-8040.

[6] 武田常広，“脳工学”，コロナ社，2003.

[7] 原宏，粟城真也，“脳磁気科学 - SQUID 計測と医学応用 -”，オーム社，1997

[8] H. Kawahara *et al.*, “Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneousfre-quencybased F0 extraction: possible role ofa repetitive structure in sounds, ” Speech Com-munication, vol. 27, no. 3-4, pp. 187-207, 1999.

[9] K. Greff *et al.*, “LSTM: A search space odyssey,” IEEE Transactions on Neural Networks and Learning Systems, pp. 2222-2232, 2016.

[10] S. Uzawa *et al.*, “Spatiotemporal Properties of Magnetic Fields Induced by Auditory Speech Sound Imagery and Perception,” IEEE EMBC2017, pp. 2542-2545.

[11] D. P. Kingma *et al.*, “Adam : a method for stochastic optimization,”International Conference on Learning Representations 2015.

# 発表論文

## 全国大会

1. 矢野彩緒里，滝口哲也，有木康雄，添田喜治，中川誠司，“非負値行列因子分解を用いた脳磁界データから音声の復元”, 日本音響学会 2018 年春季研究会講演論文集, 1-Q-40, 2018-03 (投稿済)

# 質疑応答リスト

QA. 1　質問者:江口先生

Q: アクティビティをスパースにする制約はしているか？

A: ．現段階ではしていないが学習によりある程度スパースなアクティビティは得られている．しかし今後，精度向上のため導入が必要だと考えている．

QA. 2　質問者:和泉先生

Q: モデルは被験者依存か，被験者共通か？

A: 被験者依存である．現段階では被験者依存のモデルの精度向上が必要であると考える．

Q: なぜ加算波形ではなく単一試行波形を用いるのか？

A: 実用性重視のため単一試行波形での実験をおこなったが，加算波形を用いると精度が向上することは予備実験で確認している．