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*

☆矢野彩緒里, 滝口哲也, 有木康雄 (神戸大), 添田喜治 (産総研), 中川誠司 (千葉大/産総研)

1 はじめに

発話困難な障害者は，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．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，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]” の開発例は多い．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使用者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．我々は，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，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．

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非負値行列因子分解 (Non-negative Matrix Factorization: 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，以下の方法で声質変換を実現する：(1) 入力話者の音声辞書 (入力話者辞書) と出力話者の音声辞書 (出力話者辞書) からなる同一発話内容のパラレル辞書を構築，(2) 構築された辞書を固定し，入力音声を入力辞書の少量の基底からなるスパース表現に変換，(3) 得られた入力辞書の基底毎の重み係数 (アクティビティ) に基づいて，入力話者辞書の基底を出力辞書内の基底と置き換え，線形結合することで，出力話者の音声へと変換

NMF を用いた声質変換は従来の声質変換のように統計的モデルを用いない Exemplar-based 手法であるため過学習がおこりにくい．加えて，高次元スペ クトルを用いて変換するため，自然性の高い音声へと変換可能であると考えられる．

本研究では，脳磁界データを入力，音声を出力として，それぞれの時間周波数特徴量についてパラレルな辞書を作成し，前述の手法で脳磁界データから音声への変換を試みた．

2 脳磁界データの計測と特徴量抽出

2.1 脳磁界データ計測

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 [6].

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時間も音声呈示から 800 ms とした．

脳磁界計測には，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．得られた単一試行波形データに対して，独立成分分析 (Independent Component Analysis : ICA) を適用し，眼球運動に伴うアーティファクトを除去した．

---

*Speech reconstruction from brain magnetic fields using Non-negative Matrix Factorization, YANO, Saori, TAKIGUCHI, Tetsuya, ARIKI, Yasuo (Kobe Univ.), SOETA, Yoshiharu (AIST), NAKAGAWA, Seiji (Chiba Univ./AIST).

2.2 特徴量抽出

脳磁界データ，音声データの特徴量抽出の概要を Fig.1 に示す．

本実験では単一試行波形に加え 5 回加算波形，および 10 回加算波形の 3 種類の脳磁界データを用いた．まず，左右側頭部聴覚野を覆う計 36 チャネルに対して，文字刺激呈示から 0-800 ms の区間，単一試行波形または加算波形に対して連続ウェーブレット変換 (Continuous Wavelet Transform : CWT) による時間周波数特徴量の抽出をおこなった． CWT 関数 W は以下の式 (1) で示される．

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

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．

CWT 関数 W には 1 チャネルごとに周波数帯域 1-100 Hz で 1 Hz ごとの 100 次元，時間シフトは 5 ms ごとに 0-800 ms の 160 次元を用いた．

音響特徴量には STRAIGHT [7] を用いて原音声をスペクトル包絡，F0，非周期成分に分解したものを使用した．時間は 5 ms ごとに 0-800 ms の 160 次元，スペクトル包絡 1025 次元・非周期成分 1025 次元・F0 1 次元の計 2051 次元を用いた．

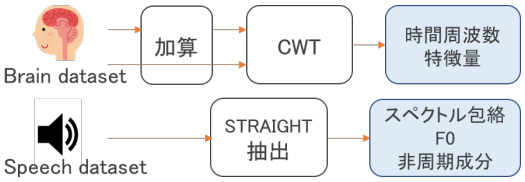

Fig. 1 How to extract features.

3 NMF を用いた音声復元

3.1 NMF を用いた音声復元手法

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．

$$\mathbf{v}_l \approx \sum_{j=1}^{J} \mathbf{w}_j,\ \mathbf{h}_{j,l} = \mathbf{W}_{\mathbf{h}_l} \qquad (2)$$

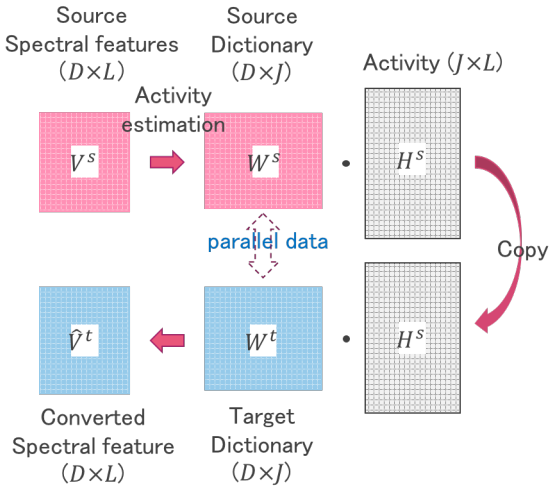

Fig. 2 Approach of NMF-Sound Reconstruction

式 (2) 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 (2) は二つの行列の内積で表される．

$$\mathbf{V} \approx \mathbf{WH} \qquad (3)$$

$$\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_L],\ \mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_L] \qquad (4)$$

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.2 に示す．$\mathbf{V}^s$ は脳磁界データの特徴量，$\mathbf{W}^s$ は脳磁界データの辞書，$\mathbf{W}^t$ は出力音声辞書，$\hat{\mathbf{V}}^t$ 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脳磁界データの辞書 $\mathbf{W}^s$ と出力音声辞書 $\mathbf{W}^t$ からなる辞書の対を用いる．

3.2 SVM を用いた基底選択

本実験では辞書選択の際誤りを抑制するため，SVM による分類をおこなっている．各実験データを時間長 100 ms ごとに分割し，時系列ごとに 3 単語の分類をおこなった．

学習，評価に用いる脳磁界データの時間周波数特徴量は，各チャネル 5 ms ごとに 100 ms の

time [ms]	0-100	100-200	200-300	300-400	400-500	500-600	600-700	700-800	Ave.
Sub.1	37.9	37.9	38.8	34.6	35.0	34.6	35.8	32.9	35.9
Sub.2	36.2	37.9	36.2	33.8	35.0	32.9	39.6	35.4	35.9
Sub.3	44.2	42.1	39.2	38.8	35.8	44.2	37.9	38.3	40.1
Sub.4	40.8	38.8	36.7	37.5	31.2	27.5	32.9	36.7	35.3
Sub.5	35.0	38.3	37.5	40.8	36.2	35.8	30.4	30.8	35.6
Sub.6	39.6	37.9	39.6	38.8	36.2	38.8	35.0	32.1	37.2
Sub.7	32.9	37.5	36.7	34.6	32.9	33.8	40.4	34.2	35.4
Sub.8	35.0	44.6	35.8	40.8	39.2	45.0	39.2	42.9	40.3
Ave.	37.7	39.4	37.6	37.4	35.2	36.6	36.4	35.4	37.0

Table 1　Single-trial classification accuracy.

time [ms]	0-100	100-200	200-300	300-400	400-500	500-600	600-700	700-800	Ave.
N = 1	37.7	39.4	37.6	37.4	35.2	36.6	36.4	35.4	37.0
N = 5	53.1	54.8	54.7	55.2	53.3	49.6	46.5	49.1	52.0
N = 10	66.5	67.9	68.0	68.9	67.9	65.6	62.0	60.5	65.9

Table 2　Classification accuracy of N = 1, 5, 10.

20 次元，1Hz ごとに 1-100 Hz の 100 次元であるが，主成分分析により 10 次元に圧縮した．

各被験者ごとの特徴量の抽出及び識別には，各被験者において，単語ごとに 400 試行分の脳磁界波形を用いた．80 試行を評価データ，残りの 320 試行を学習データとした． 学習データを 5 fold Cross Validation をおこなった．グリッドサーチから，分類精度の高いものを，評価時の時系列ごとのハイパーパラメータとして決定した．識別には RBF カーネルを用いて，one-versus-one のマルチクラス分類をおこなった．

3.3　音声生成

SVM の分類に従って，出力音声辞書の基底選択をおこない，時系列毎に出力音声特徴量を得る．出力音声特徴量として生成されたスペクトル包絡，F0，非周期成分を STRAIGHT により合成することで復元音声を生成した．

4　結果と考察

Table 1 に，SVM による 単一試行波形の識別率を被験者ごと，時系列ごとに示す．全体の識別率は 37.0 % と chance rate を 3.7 % ほど上回る精度であった．全被験者の平均値では，100-200 [ms] において 39.4 % と最も高い識別率が示され，700 [ms] 以降に最も低い識別率が示された．これは音声刺激呈示初期に被験者の音声への注意が高くなり，その後音声刺激が継続するにつれて音声への注意が逓減するためではないかと予測される．

単一試行波形および加算波形のにおける 8 人の被験者の平均を時系列ごとに Table 2 に示す．表における N は加算回数 (5 or 10) である．単一試行波形では全体の識別率の平均が 37.0 % であるのに対し，5 回加算波形では 51.8 %，10 回加算波形では 66.0 % を示した．この結果から，加算回数を増やすほど S/N 比が上昇し誘発反応を取り出すことができているといえる．

Fig.3 には実際に復元した音声のスペクトログラムを示す．単一試行波形の 450-600 [ms] の区間で大きな誤りが生じている．これは，単一試行波形では分類精度が不十分であるため SVM でおこなった辞書の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．一方，加算回数を増やすと基底選択の正解率も上昇し，10 加算波形では安定して正解に近いものを得ることができた．

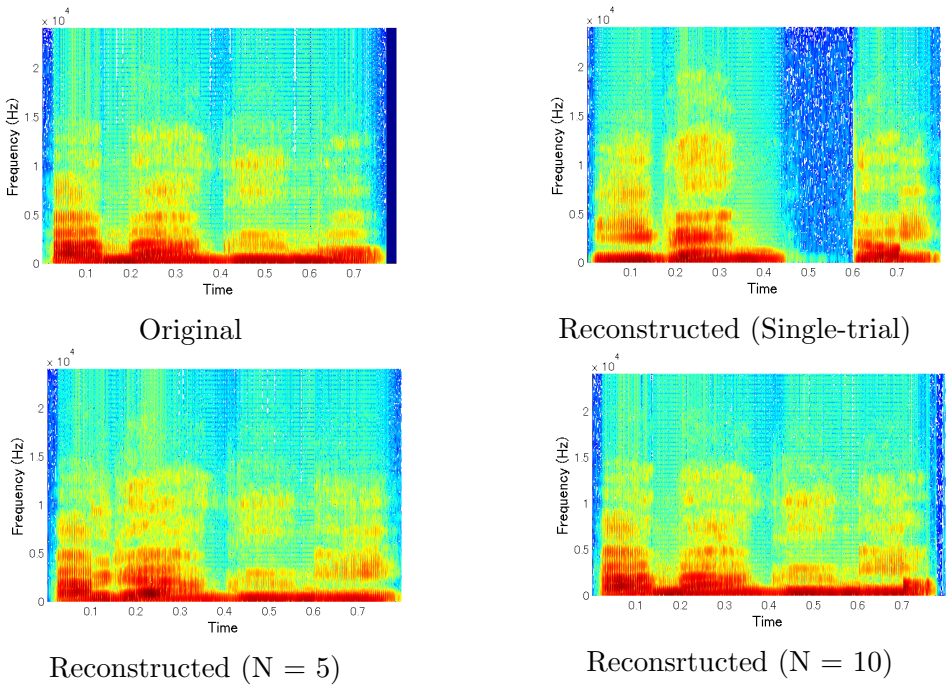

Fig. 3　Spectrogram of reconstructed sound “amagumo”.

5　おわりに

分類に基づいて，加算波形を用いると 60 % 以上の精度で音声の復元をすることができたが，単一試行波形については特徴量抽出・次元圧縮法のさらなる改善をおこなう必要がある．また，本実験は分類に基づいた音声復元をおこなったが，より汎用性の高い音声復元を実現させるため，脳磁界データから直接音声を復元する手法も検討していきたい．

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

[7] 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.