Vol.2015-MUS-107 No.32
2015/5/23

連想記憶を用いた線形分離行列の推定法に関する検討

大町 基[1,a)]　小川 哲司[1]　小林 哲則[1]　藤枝 大[2]　片桐 一浩[2]

概要：聴感上違和感を与えるノイズが発生しない音源分離の実現を目指し，ニューラルネットワークに基づく連想記憶モデルと線形分離行列を用いたブラインド音源分離手法を提案している．この手法は，音源信号を分離信号から連想記憶モデルを用いて推定する処理と，分離信号が推定された音源信号に近づくように分離行列を補正する処理とを交互に繰り返す．本研究では，提案法において特に，連想記憶モデルの構造に関する検討を行った．連想記憶モデルとして，音声のスペクトルを復元する際に局所特徴の位置情報を考慮した Convolutional autoencoder を構築し，その有効性を検証した．

キーワード：マルチチャネルブラインド音源分離，線形分離行列，連想記憶モデル，Denoising autoencoder，Convolutional autoencoder

1. はじめに

ニューラルネットワークに基づく連想記憶を用いた線形分離行列の推定法 [1] において，スペクトルを復元する際に局所特徴の位置情報を考慮した Convolutional autoencorder(CAE) を連想記憶モデルとして用いることを検討し，その有効性を示す．

音源分離の手法として，独立成分分析（ICA） [2] や独立ベクトル分析（IVA） [3], [4] などが提案されている．これらは，線形分離行列に基づく音源分離を行うため聴感上違和感を与える非線形歪が発生しないという優れた性質をもつ．しかしながら，分離行列の推定には，音源が互いに独立な分布から生成されるという仮定が広く用いられているため，音源が互いに独立でない場合に分離性能が劣化する．また，音源の事前分布を考慮して，音声らしい分離音を出力する工夫 [5], [6] がいくつか提案されているものの，より直接的に音声らしい分離音を出力する枠組みを作ることは難しい．

この問題に対して筆者らは，ニューラルネットワークに基づく連想記憶を用いた線形分離行列の推定法を提案した [1]．提案法は，分離信号から音源信号を連想記憶モデルを用いて推定する処理と，分離信号が推定された音源信号に近づくように分離行列を補正する処理とを交互に繰り返すことにより，分離行列を推定する．この枠組みにより，音源の独立性の仮定を置くことなく，音声らしい分離音を再現することが可能となる．

前報では分離行列の更新方法を中心にアルゴリズムの概要を報告したが，本研究では，連想記憶モデルの構造について検討することで分離性能の改善を目指す．提案法で用いる連想記憶モデルは，分離信号のスペクトルから音声らしい特徴を抽出するデコード，抽出した特徴から音源信号のスペクトルを復元するエンコードの処理からなる．ここではエンコードあるいはデコードの際に，スペクトルの局所的な構造を考慮するかという観点で，ニューラルネットワークの構造を比較した．

Denoising autoencoder (DAE) は，歪が含まれる音声から無歪の音声を推定する問題に対して効果的であることが報告されている [7], [8]．しかしながら，エンコード・デコード時に局所的な構造を考慮していない．Convolutional neural network (CNN) [9] は，エンコード時に音声スペクトルの局所的な特徴を抽出する構造を有するため，局所的に存在する歪が低減されることが期待できる．さらに，畳み込み層の出力に対してプーリングの処理 [10] を適用することにより，話者変動などに起因する時間周波数方向の変動に対して頑健な特徴を抽出することができる．しかし，プーリングにより得られた特徴から音源信号のスペクトルを復元する場合には，元音声の性質が保持されない可能性がある．そこで本研究では，デコード時に局所特徴の位置に関する情報を考慮した CAE を検討する．デコード時に，プーリング適用時に選択された局所特徴の位置の情報を用いてスペクトルを復元することにより，元音声らしさが保持された音声が推定されることを期待する．

[1] 早稲田大学, Waseda University, 27 Waseda-machi, Shinjyuku-ku, Tokyo 162-0042, Japan.
[2] 沖電気工業株式会社，Oki Electric Industry Co., Ltd., 1-16-8, Chuo, Warabi-shi, Saitama, 335-8510, Japan.
[a)] omachi@pcl.cs.waseda.ac.jp

ⓒ 2015 Information Processing Society of Japan

以降，2 で提案法のアルゴリズムについて概観し，3 でニューラルネットワークの構造に関して検討した結果について述べ，4 でまとめと今後の課題について述べる．

2. 連想記憶を用いた分離行列推定法

N_m 個の観測信号から N_s 個の音源信号（$N_s \leq N_m$）を再現する問題を想定し，提案法のアルゴリズムを概観する．

音源信号のスペクトルを $\mathbf{S}(\omega,\tau) \in \mathbb{C}^{N_s\times 1}$ 混合行列を $\mathbf{H}(\omega) \in \mathbb{C}^{N_m\times N_s}$ とすると，観測信号のスペクトル $\mathbf{Z}(\omega,\tau) \in \mathbb{C}^{N_s\times 1}$ は，以下のような線形変換によって表すことができる．

$$\mathbf{Z}(\omega,\tau) = \mathbf{H}(\omega)\mathbf{S}(\omega,\tau)\ . \tag{1}$$

ω，τ は離散周波数およびフレームを表す．線形分離フィルタを用いた音源分離では，$\mathbf{Z}(\omega,\tau)$ に対して線形分離行列 $\mathbf{W}(\omega) \in \mathbb{C}^{N_s\times N_m}$ による線形変換を適用することにより，$\mathbf{Y}(\omega,\tau) \in \mathbb{C}^{N_s\times N_m}$ を再現する．

$$\mathbf{Y}(\omega,\tau) = \mathbf{W}(\omega)\mathbf{Z}(\omega,\tau) = \mathbf{W}(\omega)\mathbf{H}(\omega)\mathbf{S}(\omega,\tau)\ . \tag{2}$$

ここでの目的は，$\mathbf{H}(\omega)$ の影響を打ち消す $\mathbf{W}(\omega)$ を推定することである．

提案法では，$\mathbf{W}(\omega)$ を推定するために音源信号に類似する信号（以降では，参照信号とよぶ．）$\hat{\mathbf{S}}(\omega,\tau)$ を導入する．$\hat{\mathbf{S}}(\omega,\tau)$ は，事前に音源のスペクトルを学習させた連想記憶モデルにより推定する．$\hat{\mathbf{S}}(\omega,\tau)$ を $\mathbf{S}(\omega,\tau)$ とみなし，$\mathbf{Y}(\omega,\tau)$ が $\hat{\mathbf{S}}(\omega,\tau)$ に近づくように $\mathbf{W}(\omega)$ を補正する．この枠組みにより，線形分離行列で再現しうる音の中で最も音源に近い分離信号が得られることが期待できる．本手法の流れを図 1 に，アルゴリズムを Algorithm 1 に示す．本手法は，参照信号推定部および線形分離行列更新部からなる．参照信号推定では，$\mathbf{W}(\omega)$ により得られた分離信号 $\mathbf{Y}(\omega,\tau)$ から，連想記憶モデルを用いて $\hat{\mathbf{S}}(\omega,\tau)$ を推定する．また，線形分離行列の更新では，前段で推定された参照信号のスペクトル $\hat{\mathbf{S}}(\omega,\tau)$ と $\mathbf{Y}(\omega,\tau)$ との誤差 J が小さくなるように $\mathbf{W}(\omega)$ を補正する．J が収束するまで参照信号の推定と線形分離行列の更新を繰り返すことにより，分離信号のスペクトル $\mathbf{Y}(\omega,\tau)$ を求める．

2.1 線形分離行列の更新

IVA により求めた線形分離行列 $\mathbf{W}^{(IVA)}(\omega)$ を初期値として定め，その出力を $\mathbf{Y}^{(0)}(\omega,\tau)$ とする．ここでは，$\mathbf{Y}^{(0)}(\omega,\tau)$ を参照信号のスペクトル $\hat{\mathbf{S}}(\omega,\tau)$ に近づくよう補正する行列 $\mathbf{M}(\omega) \in \mathbb{C}^{N_s\times N_s}$ を推定する．$\mathbf{M}(\omega)$ により補正されたスペクトル $\bar{\mathbf{Y}}(\omega,\tau)$ は，

$$\bar{\mathbf{Y}}(\omega,\tau) = \mathbf{M}(\omega)\mathbf{Y}^{(0)}(\omega,\tau) = \mathbf{M}(\omega)\mathbf{W}^{(IVA)}(\omega)\mathbf{Z}(\omega,\tau) \tag{3}$$

となる．すなわち，$\mathbf{M}(\omega)$ を求めた上で $\mathbf{M}(\omega)\mathbf{W}^{(IVA)}(\omega)$ を計算することにより分離行列を更新する．ここで，$\mathbf{M}(\omega)$

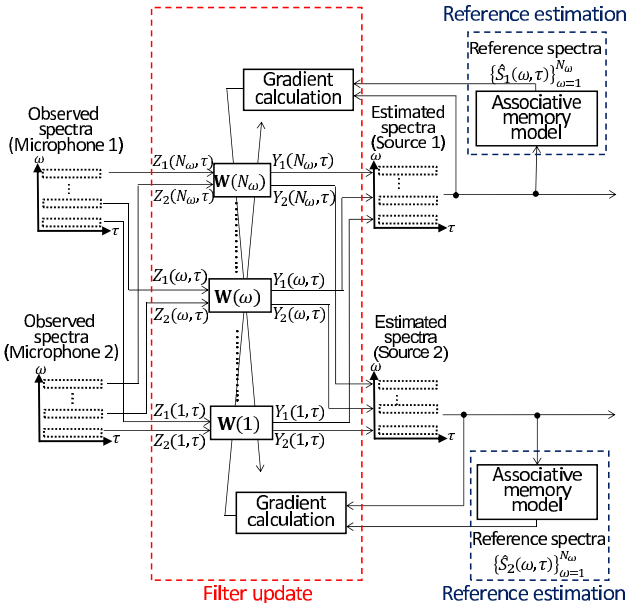


図 1 *Schematic diagram of proposed method under situation that two sources are estimated from two observations. Proposed method consists of reference spectrum estimation and linear separation filter update.*

Algorithm 1 Algorithm for separation matrix optimization based on associative memory.

Require: Observed signal $\mathbf{Z}(\omega,\tau)$
Require: Initial separation matrix $\mathbf{W}^{(0)}(\omega)$
Require: #epochs for reference signal update N_R, #epochs for filter update N_M, learning rate μ

1: $\mathbf{M}^{(0)}(\omega) = \mathbf{I}$ ($\mathbf{I}$: identity matrix).
2: $\mathbf{Y}^{(0)}(\omega,\tau) = \mathbf{W}^{(0)}(\omega)\mathbf{Z}(\omega,\tau)$.
3: Estimate $\hat{\mathbf{S}}^{(0)}(\omega,\tau)$ from $\mathbf{Y}^{(0)}(\omega,\tau)$ using Associative Memory Model.
4: **for** $i = 0 : N_R$-1
5: **for** $j = 0 : N_M$-1
6: Calculate gradient $\mathbf{G}^{(j)}(\omega)$ using $\hat{\mathbf{S}}^{(i)}(\omega,\tau)$ and $\mathbf{Y}^{(i)}(\omega,\tau)$.
7: $\mathbf{M}^{(j+1)}(\omega) = \mathbf{M}^{(j)}(\omega) - \mu\mathbf{G}^{(j)}(\omega)/||\mathbf{G}^{(j)}(\omega)||$.
8: **end for**
9: $\bar{\mathbf{M}}(\omega) = \mathbf{M}^{(N_M)}(\omega)$.
10: $\mathbf{Y}^{(i+1)}(\omega,\tau) = \bar{\mathbf{M}}(\omega)\mathbf{W}^{(\mathrm{IVA})}\mathbf{Z}(\omega,\tau)$.
11: Estimate $\hat{\mathbf{S}}^{(i+1)}(\omega,\tau)$ from $\mathbf{Y}^{(i+1)}(\omega,\tau)$ using AMM.
12: $\mathbf{M}^{(0)}(\omega) = \bar{\mathbf{M}}(\omega)$.
13: **end for**

Output: $\bar{\mathbf{W}}(\omega) = \bar{\mathbf{M}}(\omega)\mathbf{W}^{(\mathrm{IVA})}(\omega)$.

を推定するために，以下のコスト関数 $J(\omega)$ を定義する．

$$J(\omega) = \sum_{\tau=1}^{N_\tau} |\log|\hat{\mathbf{S}}(\omega,\tau)|^2 - \log|\mathbf{M}(\omega)\mathbf{Y}(\omega,\tau)|^2|^2. \tag{4}$$

$\mathbf{M}(\omega)$ は以下に示すような勾配法により求める．

$$\mathbf{M}^{(n+1)}(\omega) \leftarrow \mathbf{M}^{(n)}(\omega) - \mu\frac{\mathbf{G}(\omega)}{||\mathbf{G}(\omega)||}, \tag{5}$$

ⓒ 2015 Information Processing Society of Japan

$$\mathbf{G}(\omega) = \partial J(\omega)/\partial \mathbf{M}^{*}(\omega), \quad (6)$$

μ，n はそれぞれ学習係数および更新回数のインデックスを，* は複素共役を表す．

2.2 参照信号の推定

連想記憶モデルを用いて，歪を含む音声から無歪の音声（参照信号）を推定する．ここでの連想記憶モデルは，歪を含む音声スペクトルから音声の性質を保持した特徴を抽出するエンコード，抽出した特徴から無歪の音声スペクトルを合成するデコードの処理からなる．本研究では，それぞれの処理において局所的な構造を扱うことの効果を調べた．DAE に基づく構造をもつ連想記憶モデル（**Full-Full**），CNN に基づく構造をもつ連想記憶モデル（**Conv-Full**），そして CAE に基づく構造をもつ連想記憶モデル（**Conv-Full**）の 3 種類の連想記憶モデルを構築し，それらの性能を比較した．連想記憶モデル **Full-Full** は音声を表現するのに不要な成分を取り除く効果が期待できるが，局所的な構造を扱わない．連想記憶モデル **Conv-Full** はエンコード時のみ局所的な構造を扱う．局所的な構造を考慮することで，局所的に存在する歪に対して頑健な特徴が得られることが期待できる．また，Max-pooling を行うことで，時間周波数方向の変動に対して頑健な特徴が抽出されることが期待される．しかし，元音声らしさが失われてしまう可能性がある．連想記憶モデル **Conv-Conv** はエンコード・デコード時共に局所構造を扱う．デコード時に，局所特徴の位置情報を考慮することで，元音声らしさを保持したスペクトルが合成されることが期待できる．以降では，それぞれの連想記憶モデルについて概観する．

DAE に基づく連想記憶モデル（Full-Full）

図 2 に構造を示す．分離音声の対数パワースペクトルから 513 ビン × 10 フレームの 2 次元パターン $\mathbf{I}$ を 5 フレームずつずらしながら切り出す．$\mathbf{I}$ をベクトル化し，平均が 0，分散が 1 となるよう標準化したものを入力として用いる．

エンコード処理では，音声らしさを保持した 2,048 次元のベクトル $\mathbf{h}$ を，重み $\mathbf{w}_1$ およびバイアス $\mathbf{b}_1$ を用いて下式により抽出する．

$$\mathbf{h} = \tanh\left(\mathbf{w}_1\mathbf{I} + \mathbf{b}_1\right) . \quad (7)$$

デコード処理では，$\mathbf{h}$ から無歪の音声のスペクトルの時間周波数パターン $\mathbf{O}$ を，重み $\mathbf{w}_2$ およびバイアス $\mathbf{b}_2$ を用いて，下式により復元する．

$$\mathbf{O} = \mathbf{w}_2\mathbf{h} + \mathbf{b_2} . \quad (8)$$

$\mathbf{O}$ の大きさを 513×10 に変更し，重畳加算法を行うことで無歪音声のスペクトルを合成する．予備実験として，ノード数が 2,048 の中間層の層数を 3，5 と増やした構造も検討したが，SIR・SDR ともに上述の構造と同程度の性能であった．

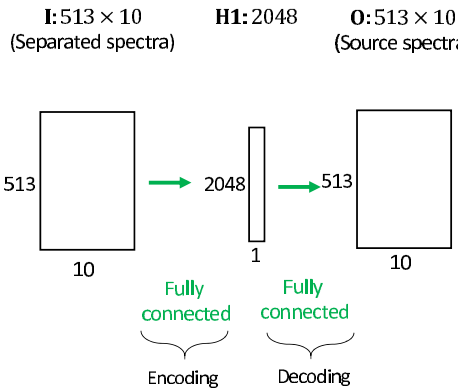


図 2 *DAE-based AMM (***Full-Full***).*

CNN に基づく連想記憶モデル（Conv-Full）

図 3 に構造を示す．**Full-Full** と同様の方法で $\mathbf{I}$ を切り出し，平均が 0，分散が 1 となるよう標準化したものを，入力として用いる．

エンコード処理ではまず入力に対して，50 種類の 30 ビン ×5 フレームのフィルタ $\mathbf{w}_1 = \{\mathbf{w}_1^1, \cdots, \mathbf{w}_1^{50}\}$ を周波数・フレーム方向ともに 1 ずつシフトさせながら重畳することで，音声スペクトルの時間周波数パターンにおける局所的な特徴 $\mathbf{c} = \{\mathbf{c}^1, \cdots, \mathbf{c}^{50}\}$ を抽出する．このフィルタは誤差逆伝播法によって学習される．f 番目のフィルタ $\mathbf{w}_1^f$ に対する応答 $\mathbf{c}^f$ は，

$$\mathbf{c}^f = \tanh\left(\mathbf{w}_1^f \otimes \mathbf{I} + \mathbf{b}_1^f\right) , \quad (9)$$

と計算される．ここで，$\mathbf{b}_1^f$ はバイアス，$\otimes$ は 2 次元の畳込みを表す．次に，$\mathbf{c}^f$ に対して max-pooling を適用することにより，時間周波数方向のずれに対して頑健な 109 ビン ×7 フレームの特徴 $\mathbf{p}^f$ を抽出する．max-pooling は $\mathbf{c}^f$ を，763 個の 5 ビン× 2 フレームの小領域に分割し，各領域における最大値を抽出することにより実現される．それぞれの $\mathbf{p}^f$ をベクトル化し結合することで，38,150 次元のベクトル $\mathbf{p}$ を生成する．そして $\mathbf{p}$ より，異なる位置間の関係を表す 2,048 次元の高次特徴 $\mathbf{h}$ を以下の式を用いて抽出する．

$$\mathbf{h} = \tanh\left(\mathbf{w}_2\mathbf{p} + \mathbf{b}_2\right) , \quad (10)$$

$\mathbf{w}_2$，$\mathbf{b}_2$ は，それぞれ重みおよびバイアスを表す．エンコードの時点で音声スペクトルの局所的な構造を陽に扱うため，**Full-Full** と比較してより音声スペクトルらしい特徴が抽出されることが期待できる．

デコード処理では，以下の式により $\mathbf{h}$ から，5,130 次元の無歪の音声の対数パワースペクトル $\mathbf{O}$ を復元する．

$$\mathbf{O} = \mathbf{w}_3\mathbf{h} + \mathbf{b}_3 , \quad (11)$$

$\mathbf{w}_3$ および $\mathbf{b}_3$ は，重みおよびバイアスを表す．**Full-Full** と同様に，$\mathbf{O}$ の大きさを 513 ビン ×10 フレームに変更し，重畳加算法を行うことで無歪音声のスペクトルを合成する．

ⓒ 2015 Information Processing Society of Japan

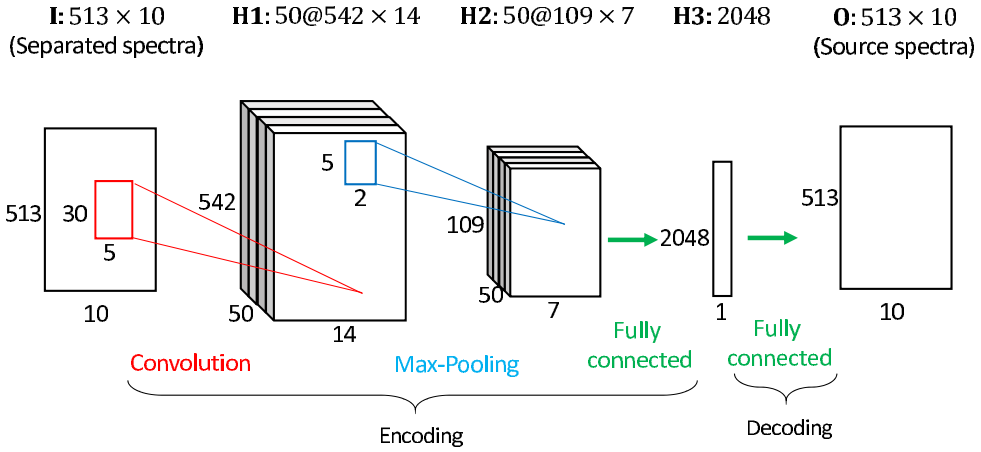


図 3 *CNN-based AAM* (**Conv-Full**).

CAE に基づく連想記憶モデル（Conv-Conv）

Conv-Full では，max-pooling を用いることにより，時間周波数方向のずれに対して頑健な特徴を抽出する．しかし局所特徴の位置に関する情報が失われる．そのため，復元されたスペクトルには元の音声らしさが含まれていない可能性がある．そこで，図 4 のように，max-pooling で抽出された局所特徴の位置情報を用いて，無歪の音声を推定する CAE を検討した．

エンコード処理は，**Conv-Full** と同様の方法で行う．このとき，max-pooling の処理において抽出された局所特徴の位置情報 $\mathbf{i}$ を保持する．

デコード処理では，まず $\mathbf{p}$ に対応する 38,150 次元の特徴 $\hat{\mathbf{p}}$ を以下の式によりデコードする．

$$\hat{\mathbf{p}} = \mathbf{w}_3\mathbf{h} + \mathbf{b}_3 \tag{12}$$

ここで，$\mathbf{w}_3$，$\mathbf{b}_3$ はそれぞれ，重みおよびバイアスを表す．次に，$\hat{\mathbf{p}}$ を 50 種類の 109 ビン ×7 フレームの特徴 $\{\hat{\mathbf{p}}^1, \cdots, \hat{\mathbf{p}}^{50}\}$ に分割する．そして，$\mathbf{i}$ に基づき $\hat{\mathbf{p}}^f$ をアップサンプリングすることにより $\mathbf{c}^f$ に相当する 542 ビン ×7 フレームの特徴 $\hat{\mathbf{c}}$ を生成する．アップサンプリングは $\hat{\mathbf{c}}^f$ のうち，$\mathbf{p}^f$ の要素が存在していた座標に $\hat{\mathbf{p}}^f$ の値を埋め，それ以外の座標の値は 0 を埋めることにより実現する．そして，513 ビン ×10 フレームの無歪音声の対数パワースペクトルの時間周波数パターン $\mathbf{O}$ を，各 $\hat{\mathbf{c}}^f$ に対して 30 ビン ×5 フレームのフィルタ $\mathbf{w}_4^f$ を重畳したものを，重ね合わせることにより復元する．

$$\mathbf{O} = \sum_{f=1}^{50} \left(\mathbf{w}_4^f \otimes \hat{\mathbf{c}}^f + \mathbf{b}_4^f \right) \tag{13}$$

ここで，$\mathbf{b}_4^f$ はバイアスを表す．復元された $\mathbf{O}$ を用いて重畳加算法を適用することにより，無歪のスペクトルを合成する．デコード時に，max-pooling により失われた局所構造の位置情報を陽に扱うことにより，**Conv-Full** と比較して，より元音声らしいスペクトルが推定されることを期待する．

3. 音源分離実験

図 5 の環境を想定し，シミュレーション実験により，提案法の分離性能を評価した．

3.1 連想記憶モデル学習条件

連想記憶モデルのパラメータを，**clean-clean** および **separated-clean** の 2 種類のパラレルコーパスを用いて誤差逆伝播法 [11] により学習した．

clean-clean は，無歪の音声の対数パワースペクトルを入力データ・教師データ双方に用いるものである．ここでは，無歪の音声として ATR 音素バランス文 [12] のセット B に含まれる 1,800 発話 (女性 4 話者，各話者 450 発話) を用いた．**separated-clean** は，分離音声の対数パワースペクトルを入力データ，対応する目的信号 (無歪音声) の対数パワースペクトルを教師データに用いるものである．分離音声は，ATR 音素バランス文のセット B に含まれる女性 4 話者から 12 組の話者対を作成し，そのうち 9 組を学習データ，3 組を開発セットして用いた．表 1 の条件で混合信号を作成し，補助関数法に基づく IVA を適用することにより分離音声を生成した．分離音声 3,600 発話 (9 組 ×2 話者 ×50 発話 ×4 条件) を入力として，対応する目的音声 3,600 発話を教師として用いた．また，早期終了に用いる開発データとして表 1 の条件で作成した分離音声 312 発話 (3 組 ×2 話者 ×52 発話 ×1 条件) を用いた．

学習を行う際には，Vanishing gradient の問題に対応するため，貪欲学習 [13] のアルゴリズムを採用した．学習時のミニバッチサイズは 256，学習係数は初期値を 0.1 とし，new-bob 法により動的に制御した．

3.2 音源分離条件

評価データはシミュレーションにより生成した．インパルス応答は図 5 の環境で収録した．本実験では，反射や残響の影響を取り除くために，直接波到来時刻から初期反射

ⓒ 2015 Information Processing Society of Japan

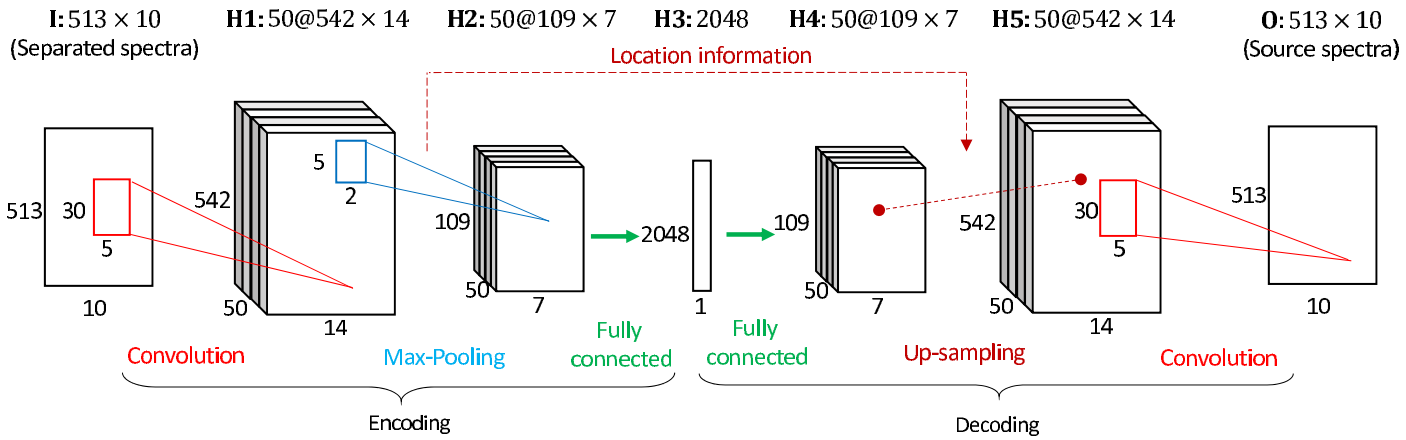


図 4 *CAE-based AMM (***Full-Full***).*

表 1 Direction of target and interference sources.

data set	**source direction of** (θ_1, θ_2)
training	(-15,15), (-45,45), (-75,75), (-90,90)
development	(-60,60)
testing	(-30,30)

到来時刻までの区間のみを切り出して用いた．JNAS 新聞読み上げコーパス [14] より無作為に抽出した音声にインパルス応答を畳み込み，2 話者同時発話 30 発話 (女性話者 10 組，各組 3 発話) を作成した．こうして得た混合信号に対して以下の分離手法を適用し，その性能を評価した．

- **IVA(Baseline)**： IVA に基づく音源分離
- **AMM-Oracle**： 音源信号を参照信号として推定した分離行列による分離
- **AMM-Full-Full**： DAE に基づく連想記憶モデル（**Full-Full**）を用いて推定した分離行列による分離
- **AMM-Convl-Full**： CNN に基づく連想記憶モデル（**Conv-Full**）を用いて推定した分離行列による分離
- **AMM-Conv-Conv**：局所特徴の位置情報を考慮した CAE に基づく連想記憶モデル（**Conv-Conv**）を用いて推定した分離行列による分離

AMM-Oracle は，連想記憶モデルが音源信号を正しく推定することできた場合の分離性能を想定しており，推定された分離行列による SIR・SDR が **AMM-Oracle** に近いほど，連想記憶の性能が高いものとして考える．

参照信号の更新回数の上限は 30 回とした．また，得られた参照信号に対する式 (5) に基づく分離フィルタの更新回数の上限は 5000 とした．学習係数 μ は初期値を 0.0001 とし，new-bob 法により動的に制御した．評価尺度として signal-to-interference ratio (SIR) および signal-to-distortion ratio (SDR) [1] を用いた．SIR は分離信号に含まれる目的音源以外の成分に対する目的音源の成分の比を表し，SDR は分離処理により生じた歪に対する目的音源の成分の比を表す．

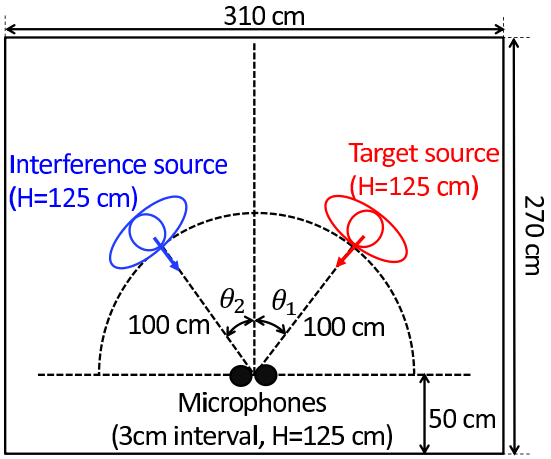


図 5 *Experimental environment.*

表 2 SIR and SDR averaged 30 utterances along with their standard deviations.

BSS methods	**SIR** [dB]		**SDR** [dB]	
	mean	stddev	mean	stddev
IVA(Baseline)	32.8	2.66	10.1	0.36
AMM-Oracle	37.3	2.36	13.9	0.45
AMM-Full-Full	32.3	3.01	12.0	1.03
AMM-Conv-Full	32.7	3.32	11.4	1.14
AMM-Conv-Conv	33.3	3.06	11.8	1.54

3.3 実験結果

表 2 に，SIR・SDR の平均値および標準偏差を示す．

エンコード・デコード共に局所構造の位置情報を扱わない DAE を用いて推定した分離行列による分離は，**IVA** に対して SDR が 1.9 dB の改善したものの，SIR が 0.5 dB 程度下回った（**AMM-Full-Full**）．また，エンコード時のみに局所構造の位置情報を考慮した CNN を用いて推定した分離行列による分離においても，同様の傾向がみられた（**AMM-Conv-Full**）．一方，エンコード・デコード共に局所構造の位置情報を陽に扱う CAE を用いて推定した分離行列による分離は，**IVA** に対して，SIR が 0.5 dB，SDR

ⓒ 2015 Information Processing Society of Japan

が 1.7 dB 改善した（**AMM-Conv-Conv**）．

このことより，連想記憶で音源信号を推定する際にはエンコード・デコード共に局所構造の位置情報を考慮することが効果的であると考えられる．

4. まとめ

連想記憶を用いた線形分離行列推定法の性能改善を目指して，連想記憶モデルの構造に関する検討を行った．連想記憶モデルの構造として，デコード時にスペクトルの局所特徴の位置情報を扱う CAE を構築し，その有効性を検証した．局所特徴の位置の情報を陽に扱うことで，元音声らしさを保持したスペクトルを再現することが可能となる．2 話者同時発話音声に対する音源分離実験により，連想記憶で音源信号を推定する際には，局所特徴の位置を考慮することが有効であることが確認された．

参考文献

[1] 大町他：“連想記憶と線形分離フィルタを用いたブラインド音源分離,” SIG-SLP105, pp.1–6, 2015 年 2 月.
[2] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” Neurocomputing, vol. 22, No. 1, pp.21–34, Nov. 1998.
[3] T. Kim *et al.*, “Blind source separation exploiting higher-order frequency dependencies,” IEEE Trans. Audio, Speech Language Proces., vol. 15, no. 1, pp. 70-79, Jan. 2007.
[4] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” Proc. of WASPAA, pp. 189-192, Oct. 2011.
[5] I. Lee *et al.*, “Adaptive independent vector analysis for the separation of convoluted mixtures using EM algorithm,” Proc. of ICASSP2008, pp.145–148, Mar. 2008.
[6] Y. Liang *et al.*, “Independent vector analysis with a generalized multivariate gaussian source prior for frequency domain blind source separation,” Signal Proces., vol. 105, pp.175-184, May 2014.
[7] T. Ishii *et al.*, “Reverberant speech recognition based on denoising autoencoder,” Proc. of INTERSPEECH2013, pp,2352–2516, Aug. 2013.
[8] Y. Xu *et al.*, “An experimental study on speech enhancement based on deep neural networks,” IEEE Signal Proc. Let., vol.21, no.1, pp.65–68, Jan. 2014.
[9] L. Deng *et al.*, “A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion,” Proc. ICASSP2013, pp.6669–6673, May. 2013.
[10] D. Scherer *et al.*, “Evaluation of pooling operations in convolutional architectures for object recognition.” Proc. of ICANN, pp. 92–101, Sept. 2010.
[11] D. E. Rummelhart *et al.*, “Learning representations by back-propagating errors,” Nature, vol.323, no.9, pp.533–536, Oct. 1986.
[12] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol.9, pp.357–363, Aug. 1990.
[13] G. Hinton *et al.*, “Deep neural networks for acustic modeling in speech recognition,” IEEE Signal Proces. Magazine, vol.29, no.6, pp.82–97, Nov. 2012.
[14] K. Itou *et al.*, “The design of the newspaper-based Japanese large vocabulary continuous speech recogonition corpus,” Proc. ICSLP, pp.3261-3264, Nov. 1998.

ⓒ 2015 Information Processing Society of Japan