

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

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: 連想記憶と線形分離フィルタを組み合わせることにより、歪が少ない高精度なブラインド音源分離方式を提案する。独立成分分析 (ICA) や独立ベクトル分析 (IVA) のような線形フィルタに基づく音源分離は、歪が少ないという特徴を持つ。しかしながら、ICA, IVA は、音源の独立性や非ガウス性を仮定するため、これが成立しないとき分離性能が劣化する。提案法は、線形分離フィルタの出力に最も近い無歪の音声を連想記憶を用いて求める処理と、連想記憶の出力に分離フィルタの出力が近づくよう分離フィルタの係数を補正する処理とを繰り返すことで分離音声を求める。これにより音源の独立性を仮定すること無く、歪の少ない分離音声を得ることができる。2 話者同時発話音声に対する音源分離実験の結果、提案法は IVA より分離精度を向上できることを確認した。

キーワード: マルチチャネルブラインド音源分離, 線形分離フィルタ, 連想記憶, Denoising Autoencoder

1. はじめに

連想記憶モデルを用いて線形分離フィルタを補正することで、歪の発生量が少なく、高精度に音源を再現可能な音源分離を実現した。

マルチチャネルのブラインド音源分離は、時間周波数マスクに基づく手法 [1]-[3] と線形分離フィルタに基づく手法 [4]-[6] とに大別できる。時間周波数マスクに基づく手法は、スペクトルの時間周波数パターンのうち、支配的に存在する音源の時間周波数成分のみを通過させる非線形なマスクを用いて音源を再現する。分離精度は高いが、非線形処理に起因する歪が発生するという欠点がある。一方、線形分離フィルタに基づく手法は、音源の混合過程の影響を取り除く線形のフィルタを用いて目的音源を再現する。このアプローチでは非線形処理に起因する歪が原理的に発生しないため、時間周波数マスクに基づく手法よりも高音質な分離信号が得られるという利点がある。

線形分離フィルタの推定には、音源が独立かつ非ガウス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。この仮定を用いた代表的な手法として、独立成分分析 (independent component analysis: ICA) や独立ベクトル分析 (independent vector analysis: IVA) がある。ICA や IVA は、音源が独立でない場合や、音源の事前分布と

実際のデータの分布が異なるときに分離性能が劣化する。この問題に対し、音源の性質を表現するのに適した事前分布を用いることで性能を改善する試みが多くなされている [7]-[9]。しかし、音源が独立でないときに分離性能が劣化する問題に対しては依然として検討の余地がある。

本研究では、線形分離フィルタに基づく音源分離において、音源に関する仮定を必要とせず高精度な音源分離を実現可能なフィルタの推定方式を提案する。提案法では、連想記憶により分離信号のスペクトルから目的音源のスペクトルを連想する処理と、連想されたスペクトルと分離フィルタ出力の誤差が最小となるように線形分離フィルタの係数を補正する処理とを繰り返すことにより、フィルタを推定する。本手法は、音源の独立性・非ガウス性の仮定を置くことなく分離が可能であり、かつ線形分離の枠組みのため分離処理による歪も少ないことが期待できる。

連想記憶は、denoising autoencoder (DAE) [10] により実現した。DAE は、歪を含む入力パターンから歪を取り除いたパターンの推定が可能であり、残響抑圧 [11] や雑音抑圧 [12] において高い性能が得られることが報告されている。音源分離によって得られる分離信号には、妨害音の消し残りや過剰な減算処理に起因する歪が含まれる。DAE に基づく連想記憶を音源分離の後段で用いることでこのような歪が低減され、目的音源に近いスペクトルが得られることを期待する。

DAE により推定されたスペクトルは、平滑化の影響を受ける。そこで DAE の出力をそのまま目的音源の推定値として用いるのではなく、フィルタを推定するために用いる

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

というアプローチが提案されている。Xia らは、ウィナーフィルタを設計するために DAE を用いている [13]。また、Huang らは、DAE の出力を用いて時間周波数マスクを設計する手法を提案している [14]。本研究は DAE の出力をマルチチャネルの線形分離フィルタの推定に用いているという点で、これらの研究とは異なる。

以降、2 で提案法の基礎となる線形分離フィルタに基づくブラインド音源分離について概観し、3 で提案法のアルゴリズムについて述べる。さらに、4 で 2 話者同時発話に対する音源分離実験により提案法の有効性を評価した結果について述べ、5 でまとめと今後の課題について述べる。

2. 線形分離フィルタに基づくブラインド音源分離

線形分離フィルタに基づく音源分離の目的は、混合過程が未知の条件下で、混合過程の影響を取り除く逆フィルタを推定することである。ここでは、 N_m 個の観測信号から N_s 個の音源を再現する問題を想定し、線形分離フィルタの推定法について概観する。ただし、 $N_s \leq N_m$ とする。

n 番目の音源信号を $s_n(t)$ 、 m 番目のマイクロホンにおける観測信号を $z_m(t)$ 、 n 番目の音源から m 番目のマイクロホンまでのインパルス応答を $h_{mn}(t)$ とすると、時間領域における混合過程は以下のように書ける。

$$z_m(t) = \sum_{n=1}^{N_s} \sum_{\tau=1}^{T_h} h_{mn}(t) s_n(t - \tau + 1) \quad (1)$$

ここで、 T_h はインパルス応答長を表す。時間領域において逆フィルタを求めることは、計算量および学習の収束性の観点において困難である。そこで、 $z_m(t)$ を T_h よりも十分に長い分析長で短時間フーリエ変換し、周波数領域で逆フィルタを求める。周波数領域では、混合過程は音源のスペクトルと伝達関数の積として以下のように書ける。

$$Z_m(\omega, \tau) = \sum_{n=1}^{N_s} H_{mn}(\omega) S_n(\omega, \tau) \quad (2)$$

ω, τ は離散周波数およびフレームを表す。また、 $Z_m(\omega, \tau)$ 、 $S_n(\omega, \tau)$ はそれぞれ観測信号のスペクトルと音源のスペクトルを、 $H_{mn}(\omega)$ は伝達関数を表す。ここで、 $\mathbf{Z}(\omega, \tau) = [Z_1(\omega, \tau), \dots, Z_{N_m}(\omega, \tau)]^T$ (T は転置を表す)、 $\mathbf{S}(\omega, \tau) = [S_1(\omega, \tau), \dots, S_{N_s}(\omega, \tau)]^T$ とすると、式 (2) は以下のように書き直すことができる。

$$\mathbf{Z}(\omega, \tau) = \mathbf{H}(\omega) \mathbf{S}(\omega, \tau) \quad (3)$$

$$\mathbf{H}(\omega) = \begin{bmatrix} H_{11}(\omega) & \cdots & H_{1N_s}(\omega) \\ \vdots & \ddots & \vdots \\ H_{N_m1}(\omega) & \cdots & H_{N_mN_s}(\omega) \end{bmatrix} \quad (4)$$

観測信号 $\mathbf{Z}(\omega, \tau)$ に対して線形分離フィルタ $\mathbf{W}(\omega)$ を適用すると、その出力 $\mathbf{Y}(\omega, \tau) = [Y_1(\omega, \tau), \dots, Y_{N_s}(\omega, \tau)]^T$ は以

下ようになる。このとき、 $Y_n(\omega, \tau)$ は推定された n 番目の音源のスペクトルを表す。

$$\mathbf{Y}(\omega, \tau) = \mathbf{W}(\omega) \mathbf{Z}(\omega, \tau) = \mathbf{W}(\omega) \mathbf{H}(\omega) \mathbf{S}(\omega, \tau) \quad (5)$$

$$\mathbf{W}(\omega) = \begin{bmatrix} W_{11}(\omega) & \cdots & W_{1N_m}(\omega) \\ \vdots & \ddots & \vdots \\ W_{N_s1}(\omega) & \cdots & W_{N_sN_m}(\omega) \end{bmatrix} \quad (6)$$

式 (5) において、 $\mathbf{W}(\omega) = \mathbf{H}^{-1}(\omega)$ であるならば、線形分離フィルタの出力と音源の信号は一致する。すなわち $\mathbf{H}(\omega)$ が既知であるならば、逆フィルタを求めることは容易である。しかし、 $\mathbf{H}(\omega)$ はマイクロホンと音源の位置関係や収音環境に依存する。そのため、 $\mathbf{H}(\omega)$ を事前情報として持つことは実用上困難である。

$\mathbf{W}(\omega)$ の推定には、音源がそれぞれ独立かつ非ガウスな分布から生成されるという仮定が広く用いられる。 $\mathbf{W}(\omega) = \mathbf{H}^{-1}(\omega)$ であるならば、分離された信号もまた音源と同じ分布から生成されたものと考えることができる。すなわち、分離された信号がお互いに独立となるような $\mathbf{W}(\omega)$ を求めればよい。ICA は、出力信号のスペクトル成分 $Y_n(\omega, \tau)$ が互いに独立となるような $\mathbf{W}(\omega)$ を求める。独立な信号を再現することが可能であるが、出力信号の順番の不定性（パーミュテーション問題）に起因する周波数間の不整合を解消する必要がある [15]。IVA は、出力信号のスペクトル成分のベクトル $[Y_n(1, \tau), \dots, Y_n(N_\omega, \tau)]^T$ が互いに独立となるような $\mathbf{W}(\omega)$ を求める。音源の事前分布として周波数間関係を考慮した多次元分布を仮定するため、周波数間の整合性は保証される。そのため、パーミュテーション問題が発生しないという利点がある。しかし、音源が独立でない場合や音源の事前分布が実際の分布と異なる場合には分離性能が低下する。また、独立性に基づき推定されたスペクトルは互いに独立であるという保証はあるものの、必ずしも音源と一致するとは限らない。

3. 連想記憶を用いた線形分離フィルタの推定

本研究では、分離信号のスペクトル $\mathbf{Y}(\omega, \tau)$ から対応する目的音源のスペクトル $\mathbf{S}(\omega, \tau)$ を推定し、 $\mathbf{Y}(\omega, \tau)$ との誤差最小化問題を解くことにより、線形分離フィルタ $\mathbf{W}(\omega)$ を求める。提案法の処理の流れを図 1 に示す。提案システムは、参照信号推定部およびフィルタ更新部から成る。参照信号推定では、線形分離フィルタ $\mathbf{W}(\omega)$ により得た分離信号 $\mathbf{Y}(\omega, \tau)$ から、事前に学習した連想記憶モデルを用いて目的信号 $\mathbf{S}(\omega, \tau)$ に相当するスペクトル $\hat{\mathbf{Y}}(\omega, \tau)$ (以降では、参照信号と呼ぶ) を求める。また、線形分離フィルタの更新では、前段で推定された参照信号のスペクトル $\hat{\mathbf{Y}}(\omega, \tau)$ と $\mathbf{Y}(\omega, \tau)$ 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フィルタ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。分離信号の時間波形は、 $\mathbf{Y}(\omega, \tau)$

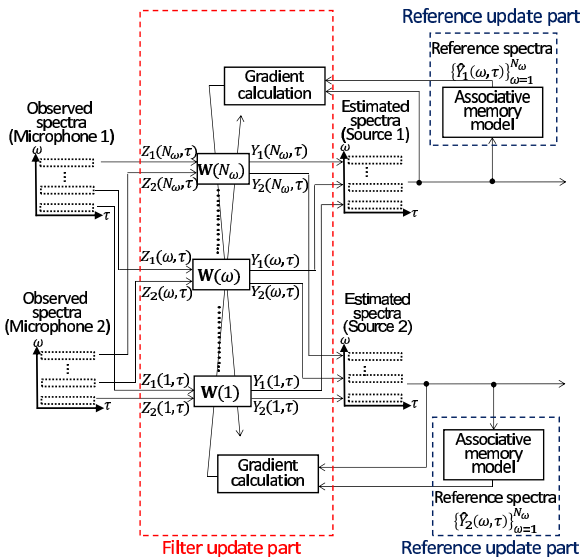


図 1 Schematic diagram of proposed method under situation that two sources are estimated from two observations. System developed consists of reference spectrum estimation and linear separation filter update.

を逆フーリエ変換した上で重畳加算法により得る。以降では、参照信号の推定と、線形分離フィルタの更新について述べる。

3.1 参照信号推定

歪を含む音声から無歪の音声を推定する連想記憶モデルを、図 2 に示すような Convolutional neural network (CNN) [16] により実現した。CNN は、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を抽出するのに適した構造を持つ。また、分離処理により生じる歪は局所的に表れる。したがって、CNN を用いて DAE を設計することで、局所的に存在する歪が低減されることを期待する。

表 1 に CNN の構造を示す。歪を含む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 次元パターンを 5 フレーム間隔で切り出す。切り出したパターンを平均が 0、分散が 1 となるよう標準化し CNN の入力とした。畳み込み層の各ユニットは、入力された 2 次元パターンに対して、30 ビン × 5 フレームのフィルタを周波数方向に 15、フレーム方向に 2 ずつシフトさせながら重畳することで得られる。つまり、畳み込み層の各ユニットは音声の時間周波数パターンにおける小さな部分領域のみから入力を受ける。また、入力層と畳み込み層の各ユニット間では同一の重みを共有する。したがって、畳み込み層では、時間周波数パターンにおける異なる位置の同一の局所パターンが検出される。ボトルネック層では、畳み込み層で得た音声の局所的な特徴から時間周波数パターンにおける異なる位置間の高次な特徴を抽出する。出力層では、音声の対数パワースペクトルにおける 513 ビン × 10 フレームの 2 次元

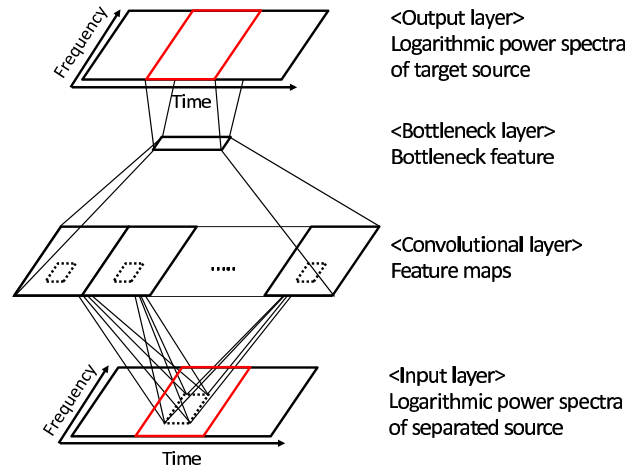


図 2 Associative memory model (AMM). AMM estimates logarithmic power spectra of target source from one of separated sources.

表 1 Structure of associative memory model.

Dimensionality of input feature	513 bins × 10 frames (8000 Hz × 160 ms)
Size of convolutional filters	30 bins × 5 frames (468.8 Hz × 80 ms)
Shift of convolutional filters	Freq.: 15 bins (234.4 Hz) Frame: 2 frames (32 ms)
# of convolutional filters	94
# of units on convolutional layer	9306
# of units on bottle-neck layer	2253

パターンの推定値を出力する。

各層の重みおよびバイアスは誤差逆伝播法 [17] によって決定する。以下の 2 種類のデータ対を平行コーパスとして使い、CNN の学習を行った。

- **Train-1:** 無歪の音声の対数パワースペクトルを入力データと教師データ双方に用いる。
- **Train-2:** 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。

Train-1 に用いる無歪の音声は、ATR 音素バランス文 [19] 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) である。AMM の学習に用いる分離音声は、表 2 の条件で作成した。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、1 組を開発セットとして用いた。**Train-2** については、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、CNN の学習における早期終了に用いる開発データとして表 2 の条件で作成した分離音声 104 発話 (1 組 × 2 話者 × 52 発話 × 1 条件) を用いた。

一般に、中間層の層数が多いニューラルネットワークを学習する際には、vanishing gradient problem とよばれる

表 2 Experimental setups for sound source separation and AMM modeling. θ_1 and θ_2 represent angle of sources 1 and 2, respectively.

	Training set	Development set
Database	ATR Japanese speech database	
Sampling frequency	16 kHz	
Number of speakers	4 females	
Number of pairs of speakers	9 pairs of females	1 pair of females
Number of sentences for each speaker	50 sentences (Randomly selected from A-I set)	52 sentences (Randomly selected from J set)
Reverberation time (RT60)	0 ms	
Microphone array	2 microphones with interval of 3 cm	
Source angles (θ_1, θ_2) [deg]	(-15,15), (-45,45), (-75,75), (-90,90)	(-60,60)
SNR	0 dB	
Separation method	Auxiliary function based IVA	

現象により、学習がうまくできないことが指摘されている。この問題に対応するため、貪欲学習 [18] により CNN を学習した。学習時のミニバッチサイズは 100、学習係数は初期値を 0.01 とし、new-bob 法により動的に制御した。

3.2 線形分離フィルタの更新

線形分離フィルタの初期値 $\mathbf{W}^{(init)}(\omega)$ を定め、その出力を $\mathbf{Y}(\omega, \tau)$ とする。ここでは、 $\mathbf{Y}(\omega, \tau)$ を連想記憶により推定された参照信号のスペクトル $\hat{\mathbf{Y}}(\omega, \tau)$ に近づくよう補正する行列 $\mathbf{M}(\omega) \in \mathbb{C}^{N_s \times N_s}$ を推定することを考える。提案システムでは、ICA や IVA などにより線形分離フィルタを推定し、その値を提案法である連想記憶モデルに基づくフィルタ推定の初期値として用いる。 $\mathbf{M}(\omega)$ により補正されたスペクトル $\bar{\mathbf{Y}}(\omega, \tau)$ は、

$$\bar{\mathbf{Y}}(\omega, \tau) = \mathbf{M}(\omega)\mathbf{Y}(\omega, \tau) = \mathbf{M}(\omega)\mathbf{W}^{(init)}(\omega)\mathbf{Z}(\omega, \tau) \quad (7)$$

となる。 $\mathbf{M}(\omega)$ および $\mathbf{W}^{(init)}(\omega)$ は線形変換であり、その積 $\mathbf{M}(\omega)\mathbf{W}^{(init)}(\omega)$ も線形変換と見なすことができる。そこで、 $\bar{\mathbf{W}}(\omega) = \mathbf{M}(\omega)\mathbf{W}^{(init)}(\omega)$ とすると、式 (7) は、

$$\bar{\mathbf{Y}}(\omega, \tau) = \bar{\mathbf{W}}(\omega)\mathbf{Z}(\omega, \tau) \quad (8)$$

となる。これは観測信号のスペクトルに、線形分離フィルタ $\bar{\mathbf{W}}(\omega)$ を適用したものと解釈できる。

そこで、 $\mathbf{M}(\omega)$ を求めた上で $\mathbf{M}(\omega)\mathbf{W}^{(init)}(\omega)$ を計算することにより、線形分離フィルタを更新する。 $\mathbf{M}(\omega)$ を推定するために、以下のコスト関数 $J(\omega)$ を定義する。

$$\begin{aligned} J(\omega) &= \sum_{n=1}^{N_s} \left| \log |\hat{\mathbf{Y}}_n(\omega, \tau)|^2 - \log |\bar{\mathbf{Y}}_n(\omega, \tau)|^2 \right|^2 \\ &= \sum_{n=1}^{N_s} \left| \log |\hat{\mathbf{Y}}_n(\omega, \tau)|^2 \right. \\ &\quad \left. - \log \left| \sum_{j=1}^{N_s} M_{nj}(\omega) Y_j(\omega, \tau) \right|^2 \right|^2 \end{aligned} \quad (9)$$

M_{nj} は $\mathbf{M}(\omega)$ の (n, j) 要素を表す。 $\mathbf{M}(\omega)$ は、以下に示す

Algorithm 1 Algorithm for separation filter update.

Require: Observed signal $\mathbf{Z}(\omega, \tau)$

Require: Initial linear separation filter $\mathbf{W}^{(init)}(\omega)$

Require: #epochs for reference signal update N_R , #epochs for filter update N_M , learning rate μ

```

1:  $\mathbf{M}^{(0)}(\omega) = \mathbf{I}$ .
2:  $\mathbf{Y}^{(0)}(\omega, \tau) = \mathbf{W}^{(init)}(\omega)\mathbf{Z}(\omega, \tau)$ .
3: Estimate  $\hat{\mathbf{Y}}^{(0)}(\omega, \tau)$  from  $\mathbf{Y}^{(0)}(\omega, \tau)$  using AMM.
4: for  $i = 0 : N_R - 1$ 
5:   for  $j = 0 : N_M - 1$ 
6:     Calculate  $\mathbf{G}^{(j)}(\omega)$  using  $\hat{\mathbf{Y}}^{(i)}(\omega, \tau)$  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:   $\mathbf{Y}^{(i+1)}(\omega, \tau) = \bar{\mathbf{M}}(\omega)\mathbf{W}^{(init)}(\omega)\mathbf{Z}(\omega, \tau)$ .
11:  Estimate  $\hat{\mathbf{Y}}^{(i+1)}(\omega, \tau)$  from  $\mathbf{Y}^{(i+1)}(\omega, \tau)$  using AMM.
12:   $\mathbf{M}^{(0)}(\omega) = \bar{\mathbf{M}}(\omega)$ .
13: end for

```

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

ような勾配法により求めることができる。

$$\mathbf{M}^{(0)}(\omega) = \mathbf{I} \quad (10)$$

$$\mathbf{M}^{(n)}(\omega) \leftarrow \mathbf{M}^{(n-1)}(\omega) - \mu \frac{\mathbf{G}(\omega)}{\|\mathbf{G}(\omega)\|} \quad (11)$$

$$\mathbf{G}(\omega) = \begin{bmatrix} \frac{\partial J(\omega)}{\partial M_{11}^*(\omega)} & \cdots & \frac{\partial J(\omega)}{\partial M_{1N_s}^*(\omega)} \\ \vdots & \ddots & \vdots \\ \frac{\partial J(\omega)}{\partial M_{N_s1}^*(\omega)} & \cdots & \frac{\partial J(\omega)}{\partial M_{N_sN_s}^*(\omega)} \end{bmatrix} \quad (12)$$

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、 $\mathbf{I}$ は単位行列を表す。また、 $*$ は複素共役を表す。分離フィルタの更新アルゴリズムを Algorithm 1 まとめる。

4. 音源分離実験

4.1 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 3 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到来時刻までの区間のみを切り出して用いた。JNAS 新聞

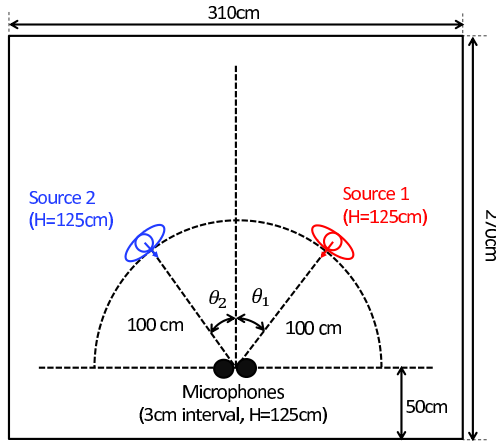


図 3 Experimental environment. Source angles (θ_1, θ_2) are $(-30, 0)$, $(-30, 30)$, $(0, -30)$, $(0, 30)$, $(30, -30)$, and $(30, 0)$.

読み上げコーパス [20] 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下記の音源分離手法を適用し分離性能を評価した。

- **IVA**: 補助関数法に基づく IVA [6]
- **IVA-AMM (提案法)**: 連想記憶を用いた線形分離フィルタの更新 (初期値: **IVA** のフィルタ)

参照信号の更新の上限は 30 回とした。また、得られた参照信号に対する式 (11) 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) 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。以下に SIR, SDR の計算式を示す。

SIR [dB]:

$$\frac{1}{N_s N_\omega} \sum_{i=1}^{N_s} \sum_{\omega=1}^{N_\omega} 10 \log_{10} \frac{\sum_{\tau=1}^{N_\tau} |S_i(\omega, \tau)|^2}{\sum_{j=1}^{N_s} \sum_{\tau=1}^{N_\tau} (1 - \delta_{ij}) |Y_{ij}(\omega, \tau)|^2}$$

SDR [dB]:

$$\frac{1}{N_s N_\omega} \sum_{i=1}^{N_s} \sum_{\omega=1}^{N_\omega} 10 \log_{10} \frac{\sum_{\tau=1}^{N_\tau} |S_i(\omega, \tau)|^2}{\sum_{\tau=1}^{N_\tau} (|S_i(\omega, \tau)| - |Y_{ii}(\omega, \tau)|)^2}$$

ただし、

$$Y_{ij}(\omega, \tau) = \sum_{m=1}^{N_m} W_{im}(\omega) H_{mj}(\omega) S_j(\omega, \tau)$$

N_τ , N_ω は、それぞれフレーム数、周波数ビン数を表す。また、 δ_{ij} は、 $i = j$ のときに 1, $i \neq j$ のときに 0 を返す。

4.2 実験結果

図 4 および 5 に SIR と SDR の平均値と標準偏差を示す。

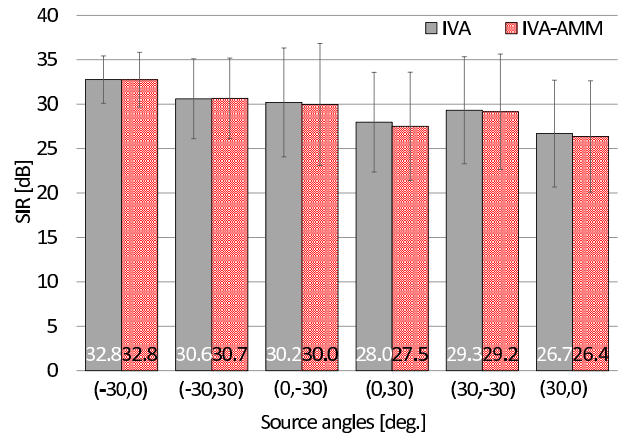


図 4 Averaged SIR for each condition. Error bar represents standard deviation.

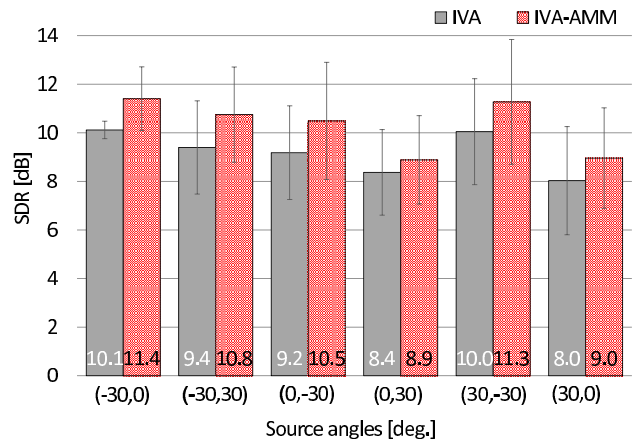


図 5 Averaged SDR for each condition. Error bar represents standard deviation.

SIR の平均値としては、提案法は IVA と同等の性能を示した。今回の実験条件では、IVA より目的音源以外の成分が十分に抑圧できていたため、提案法の効果が見えにくくなっていると思われる。

提案法の SDR は、いずれの実験条件においても IVA を上回った。提案法により IVA で生じる歪を減じ、目的音源を高精度に再現できることがわかった。

図 6 に、分離信号 (a)、参照信号 (b) および目的音源の推定値 (c) と真値 (d) のスペクトログラムを示す。例えば、1.2~1.5 s, 0~2 kHz の領域に注目すると、IVA では分離処理による影響が見て取れる (a)。一方、このスペクトルを入力として連想された参照信号のスペクトルには、この影響は含まれていない (b)。このような参照信号のスペクトルを用いて線形分離フィルタを補正することで (c)、目的音源に近いスペクトル (d) が再現されていることがわかる。

5. まとめと今後の課題

歪の発生量が少ない高精度な音源分離の実現を目指し、

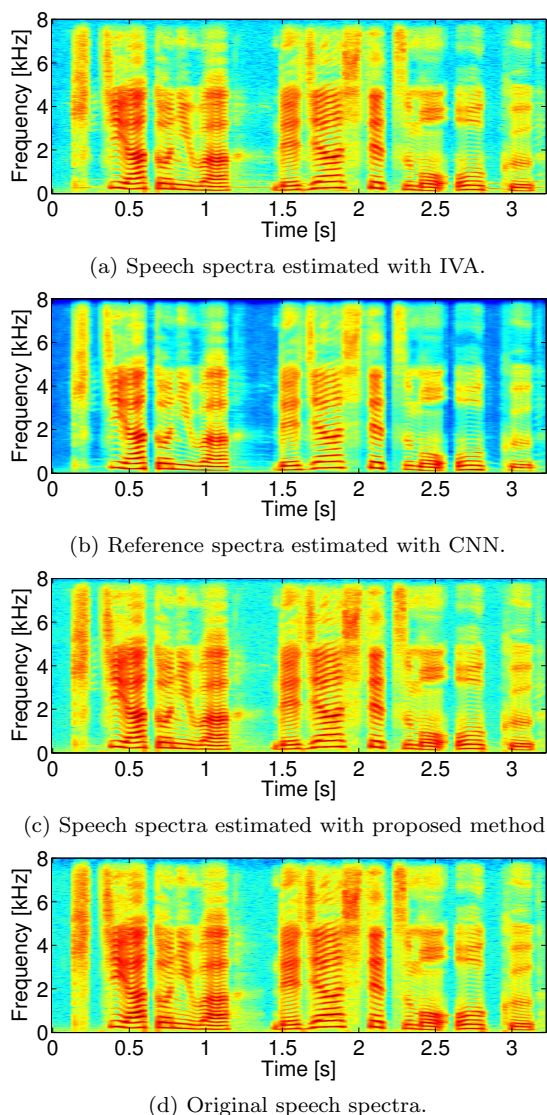


図 6 Example of logarithmic power spectra.

連想記憶モデルを用いて線形分離フィルタを推定する手法を提案した。ICA や IVA といった従来の線形分離フィルタの推定手法では、音源の独立性・非ガウス性の仮定が成立しない場合に分離性能が低下する。一方提案法は、線形分離フィルタの出力を、無歪音声のスペクトルを学習した連想記憶の出力に近づける。そのため、音源に関する仮定を必要とせず、線形分離フィルタの枠組みの中で、音声らしいスペクトルを再現することができる。2 話者同時発話音声に対する音源分離実験により、提案法は IVA によって生じる歪を低減できることを確認した。

今後は、実環境における提案法の有効性、特に残響や雑音の影響に対する頑健性を調査する。

参考文献

- [1] H. Sawada *et al.*, “Under-determined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp.516-527, Mar. 2011.
- [2] S. Araki *et al.*, “Simultaneous clustering of mixing and spectral model parameters for blind sparse source separation,” *Proc. of ICASSP2010*, Mar. 2010.
- [3] A. Alinaghi *et al.*, “Integrating binaural cues and blind source separation method for separating reverberant speech mixtures,” *Proc. of ICASSP2011*, May 2011.
- [4] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, No. 1, pp.21-34, Nov. 1998.
- [5] 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.
- [6] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” *Proc. of WASPAA*, pp. 189-192, Oct. 2011.
- [7] E. Moulines *et al.*, “Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models,” *Proc. of ICASSP1997*, pp. 3617-3620, Apr. 1997.
- [8] 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.
- [9] 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.
- [10] P. Vincent *et al.*, “Extracting and composing robust features with denoising autoencoders,” *Proc. of ICML2008*, pp. 1096-1103, Jul. 2008.
- [11] T. Ishii *et al.*, “Reverberant speech recognition based on denoising autoencoder,” *Proc. of INTERSPEECH2013*, Aug. 2013.
- [12] Y. Xu *et al.*, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal proces. letters*, vol.21, no.1, pp.65-68, Jan. 2014.
- [13] B. Xia and C. Bao, “Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification,” *Speech communication*, vol.60, pp.13-29, Feb. 2014.
- [14] P.-S. Huang *et al.*, “Deep learning for monaural speech separation,” *Proc. ICASSP2014*, pp.1562-1566, May. 2014.
- [15] H. Sawada *et al.*, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” *IEEE Trans. Speech and Audio Proces.*, vol.12, no.5, pp.530-538, Sept. 2004.
- [16] 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.
- [17] D. E. Rummelhart *et al.*, “Learning representations by back-propagating errors,” *Nature*, vol.323, no.9, pp.533-536, Oct. 1986.
- [18] G. Hinton *et al.*, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal Proces. Magazine*, vol.29, no.6, pp.82-97, Nov. 2012.
- [19] 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.
- [20] K. Itou *et al.*, “The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus,” *Proc. ICSLP*, pp.3261-3264, Nov. 1998.