

論文 / 著書情報
Article / Book Information

論題(和文)	雑音軽減マスクを用いた劣決定性音源分離の改良
Title(English)	
著者(和文)	永井友梨, 西原明法
Authors(English)	Yuri Nagai, AKINORI NISHIHARA
出典(和文)	第23回信号処理シンポジウム, , No. B6-2,
Citation(English)	, , No. B6-2,
発行日 / Pub. date	2008, 11
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2008 Institute of Electronics, Information and Communication Engineers.

雑音軽減マスクを用いた劣決定性音源分離の改良

Improvement of underdetermined blind source separation using noise reduction mask

永井友梨

西原明法

東京工業大学 理工学研究科 集積システム専攻

Yuri NAGAI

Akinori NISHIHARA

Dept. of Communications and Integrated Systems,

Graduate School of Science and Engineering, Tokyo Institute of Technology

アブストラクト スパース性に基づく音源分離による推定ミスを、マスク処理により修正する手法を提案する。音源数 $n > \text{マイク数 } m$ であるとき、混合行列推定法を用いて線形計画問題による推定を行うと推定ミスが発生する。本研究では、混合行列推定法を用いた得られた推定信号に、さらに修正マスクを適用することで線形計画問題による推定ミスを改善する。実験により、特に推定ミスが多く見られる原信号が 0 に近い部分で正確に分離された信号数が 1.5～2.0 倍に改善された。また、SDR(Signal to Difference Ratio) が改善され、全体的に推定ミスによるノイズが消去された。

Abstract We propose a method to reduce estimation errors in underdetermined blind source separation in sparse representation using noise reduction mask.

Errors occur when we estimate original signals by using linear programming problem with estimated mixing matrix if the number of original source n is greater than that of observed signals. We apply noise reduction mask on the signal obtained from the linear programming problem to reduce the errors. From our experiments, we got 1.5 to 2.0 times more number of correct signals especially when the signal is very small, where more errors tend to occur than other points. We also got improved SDR(Signal to Difference Ratio) and noise was reduced in whole signals.

1 はじめに

音源数 $n \leq \text{マイク数 } m$ の場合のブラインド音源分離(BSS)手法として、独立成分分析(ICA)が高い性能を上げている[1]。しかし $n > m$ の場合は難しく、DUET[2]に代表される信号のスパース性に基づくマスク手法や、混合行列を推定する手法[3]が提案されている。しかしマスク手法を用いた場合、信号が完全なスパースで

なければ性能が低く、混合行列推定手法の場合、線形計画問題による信号推定の誤りが発生する。本研究では、完全なスパースでない信号において混合行列推定法によって得られた信号の推定誤り問題を解決する方法を提案する。

2 スパース性を用いた混合行列推定法の概要

長さ K 、音源数 m 、観測点数 n の信号は、

$$X=AS \quad (1)$$

と書ける。ここで、 $A \in R^{n \times m}$ は混合行列、 $S = [s(1), \dots, s(K)] \in R^{m \times K}$ は原信号、 $X = [x(1), \dots, x(K)] \in R^{n \times K}$ は観測信号である。混合行列推定法は、1. 混合行列 A の推定、2. 音源信号 S の推定の 2 つのステップに分かれている[3]。

2.1 混合行列 A の推定

時間領域で観測した信号 X をさらにスパースにするためウェーブレットパケットを適用し、時間周波数領域にて信号を取り扱おうと、

$$\tilde{X} = A\tilde{S} \quad (2)$$

と書ける。ここで X, S は時間周波数領域での観測信号、音源信号である。今、 S のうち L 個の列において、1 行目のみに非零値が存在するとする。このとき、

$$[\tilde{x}(i_1), \dots, \tilde{x}(i_L)] = A[\tilde{s}(i_1), \dots, \tilde{s}(i_L)] \quad (3)$$

$$= [a_1\tilde{s}_1(i_1), \dots, a_1\tilde{s}_1(i_L)] \quad (4)$$

となる。

次に、 $q \in 1, \dots, n$ について信号比行列を考えると(4)より、

$$\begin{bmatrix} \frac{\tilde{x}_1(i_1)}{\tilde{x}_q(i_1)} & \dots & \frac{\tilde{x}_1(i_L)}{\tilde{x}_q(i_L)} \\ \vdots & \vdots & \vdots \\ \frac{\tilde{x}_n(i_1)}{\tilde{x}_q(i_1)} & \dots & \frac{\tilde{x}_n(i_L)}{\tilde{x}_q(i_L)} \end{bmatrix} = \begin{bmatrix} \frac{a_{11}}{a_{q1}} & \dots & \frac{a_{11}}{a_{q1}} \\ \vdots & \vdots & \vdots \\ \frac{a_{n1}}{a_{q1}} & \dots & \frac{a_{n1}}{a_{q1}} \end{bmatrix} \quad (5)$$

となり a_1 は、

$$a_1 = a_{q1} \left[\text{mean} \left(\frac{\tilde{x}_1(i_j)}{\tilde{x}_q(i_j)} \right) \quad \cdots \quad \text{mean} \left(\frac{\tilde{x}_n(i_j)}{\tilde{x}_q(i_j)} \right) \right]^T \quad (6)$$

と推定することができる。ここで、 $\text{mean}(\tilde{x}_1(i_j))/(\tilde{x}_q(i_j)) = (1/L) \sum_{j=1}^L (\tilde{x}_1(i_j))/(\tilde{x}_q(i_j))$ である。

実際には、時間周波数領域で信号比行列

$$\begin{bmatrix} \frac{\tilde{x}_1(1)}{\tilde{x}_q(1)} & \cdots & \frac{\tilde{x}_1(K)}{\tilde{x}_q(K)} \\ \vdots & \ddots & \vdots \\ \frac{\tilde{x}_n(1)}{\tilde{x}_q(1)} & \cdots & \frac{\tilde{x}_n(K)}{\tilde{x}_q(K)} \end{bmatrix} \quad (7)$$

を求め、各行のすべての値を M_0 個のビンに分ける。次に、エントリーの少ないビンに属する信号が存在する列は消去することによって、ある 1 信号のみが有効である列を推定する。同じビンに存在するビンの平均が推定混合行列 $\tilde{A}$ の一列となる。

2.2 原信号 S の推定

以上のように推定された混合行列を用いて、原信号 S は

$$\min \sum_{i=1}^m \sum_{j=1}^K |s_i(j)|, s.t. AS=X \quad (8)$$

の線形計画問題を解いて得られる。ここで、 $s_i(j) (i = 1, \dots, m, j = 1, \dots, K)$ は S のエントリーである。もし与えられた信号が十分スパースでない場合は、時間周波数領域にて同様に

$$\min \sum_{i=1}^m \sum_{j=1}^K |\tilde{s}_i(j)|, s.t. \tilde{A}\tilde{S}=\tilde{X} \quad (9)$$

の線形計画問題となる。 $\tilde{s}_i(j) (i = 1, \dots, m, j = 1, \dots, K)$ は $\tilde{S}$ の時間周波数表現である。

3 従来の混合行列推定法の問題点

以上の方法によって、図 1 に示す音声の混合音声 (図 2) の音源分離を行ったのが図 3 である。原信号数 $m = 4$ 、観測点数 $n = 3$ である。

図 1 と図 3 を比べると、特に値が 0 に近いところで復元誤りが多く存在していることがわかる。これらは線形計画問題による推定誤りであるが、値が 0 に近い場合はノイズとなって顕著に表れる。図 4、図 5 はそれぞれ図 1 の音源信号と図 3 に示す推定分離信号の一部 (6601-6800 サンプル) をフーリエ変換したものである。窓関数は Hann 窓を用いた。

図 1 と図 4 より、時間領域でも時間周波数領域でもスパースではない場合が存在することがわかる。完全なスパースではない信号の場合、推定信号に誤りが発生する。

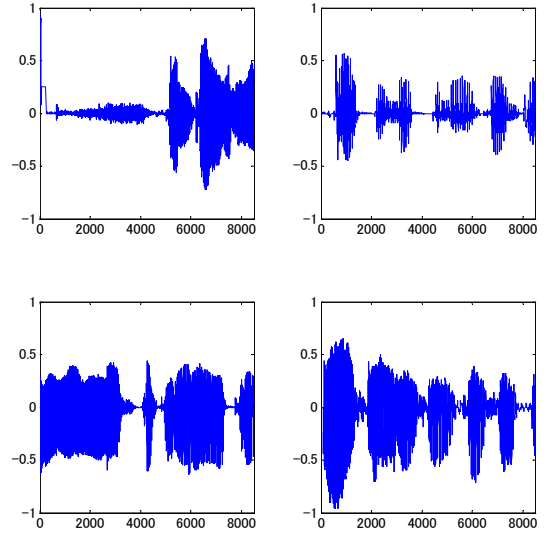


図 1: 原信号 $m = 4$

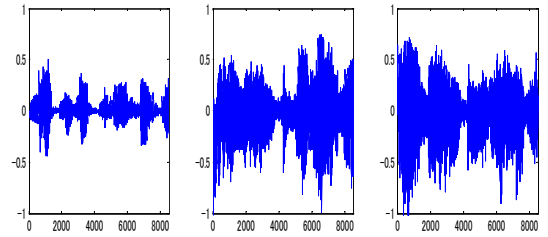


図 2: 観測信号 $m = 3$

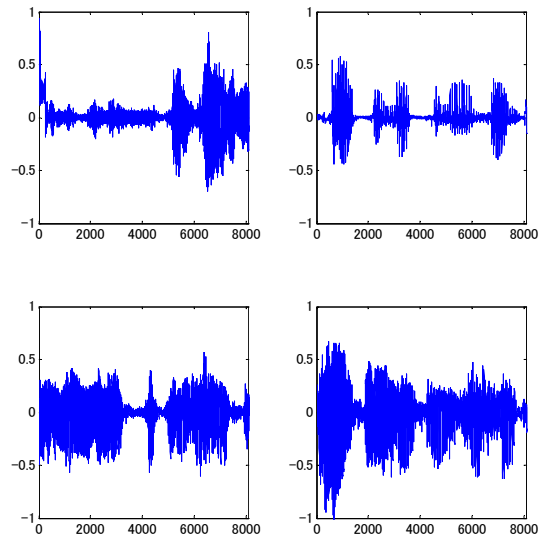


図 3: 分離された信号

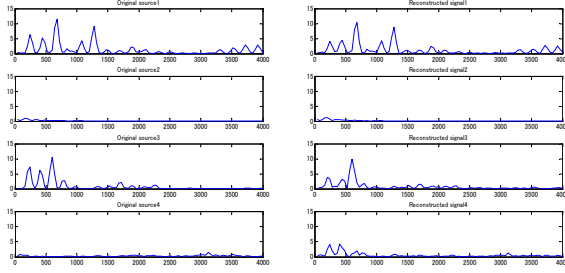


図 4: スペクトラム
原信号 (6601-6800)

図 5: スペクトラム
推定信号 (6601-6800)

6600 サンプル付近を比べると、図 1 の時間領域でも、図 4 の時間周波数領域でも信号の重なるポイントが存在する。図 4、5 を比較すると、4 つ目の信号で推定信号に誤りが生じていることがわかる。図 3 の推定信号 4、6600 サンプルあたりが図 1 の音源信号 4 を比べると復元誤りが生じていることがわかる。そこで、このような完全なスパースでない混合信号における時間周波数雑音軽減マスクを提案する。

4 提案する雑音軽減マスク

提案法では、推定混合行列を用いて得られた推定音源信号 $\bar{S}$ の時間周波数領域に対して、線形計画問題の復元誤りを修正するためのマスクを適用する。従来の劣決定性音源分離のマスクングでは各時間周波数の点で 1 つのアクティブ信号が存在すると仮定しているが、提案手法では各フレーム毎のアクティブである信号数によってマスクを決定していく。ここでフレームとは t 番目の L サンプルを含む。マスクが決定されたら、新たな推定音源 $\tilde{y}_J(f, t)$ を

$$\tilde{y}_J(f, t) = M_J(f, t) \tilde{s}_J(f, t) \quad (10)$$

により求め、最後に時間領域に戻す。ここで、 $\tilde{s}_J(f, t)$ は混合行列を用いて最適化問題を解いて得た推定音源信号 $\bar{S}$ の時間周波数領域の J 番目の値である。 $J = 1, \dots, m$ 、 m は音源数である。 t はフレーム番号である。雑音軽減マスクの決定方法は 1. フレーム毎のアクティブ信号数推定、2. 各時間周波数点でのマスク決定の 2 ステップに分かれている。

まずステップ 1 では、各フレーム毎にアクティブである信号数 $N(t)$ を推定する。フレーム毎の推定混合行列 $\bar{A}_f(t)$ を求め、すでに得られている混合推定行列 $\bar{A}$ と各列を比較し、近い値の列数がアクティブ信号数 $N(t)$ と推定する。ステップ 2 では、推定信号 $\tilde{s}_J(f, t)$ の各信号で値が重なっている点において、アクティブ信号数を $N(t)$ 以下に修正する。このとき、そのフレームでの各信号の平

均値の小さいものから優先的にノンアクティブ化させる。提案法の大まかな流れを図 6 に示す。

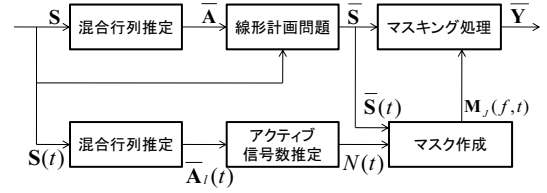


図 6: 提案法の大まかな流れ

4.1 Step1. アクティブ信号数の推定

各フレーム毎にアクティブ信号数を推定する。先に示した推定混合行列 $\bar{A} \in R^{n \times m}$ と同様の方法を用いてフレーム毎の推定混合行列 $\bar{A}_l(t) \in R^{n \times k}$ を求めると、 $k \leq m$ で、 $\bar{A}_l(t)$ は $\bar{A}$ の部分行列である。

しかし実際は、 $\bar{A}_l(t) \in R^{n \times k_r}$ 、 $k \leq k_r$ であり、必ずしも $\bar{A}_l(t) \subset \bar{A}$ になっていない場合が多い。そこで、 $\bar{A}_l(t)$ の各列 $\bar{a}_i(t) \in R^n$ の中で、

$$\bar{a}_i(t) \subset \bar{A}, i = 1, \dots, k \quad (11)$$

である列の数 $N(t) = k$ が、そのフレームでのアクティブな信号の数と推定できる。

4.2 Step2. マスクの決定法

以上の方法で各フレーム t におけるアクティブ信号数 $N(t)$ を推定する。時間周波数領域のある点 f, t でのマスク $M_J(f, t)$ は $N(t)$ によって決定する。推定混合行列によって得られた推定原信号 $\bar{S}$ の時間周波数領域のある点 f, t における提案するマスク決定アルゴリズムを以下に示す。

1. 各時間周波数領域の振幅比と各時間の信号のスペクトラムの平均値を求める。

$$RatioVector(f, t) = \begin{bmatrix} \frac{|\tilde{s}_1(f, t)|}{|\tilde{s}_Q(f, t)|} \\ \vdots \\ \frac{|\tilde{s}_m(f, t)|}{|\tilde{s}_Q(f, t)|} \end{bmatrix} \quad (12)$$

$$MeanVector(t) = \frac{1}{N_f} \begin{bmatrix} \sum_f \tilde{s}_1(f, t) \\ \vdots \\ \sum_f \tilde{s}_m(f, t) \end{bmatrix} \quad (13)$$

ここで、 $\tilde{s}_Q(f, t)$ は f, t における信号の最大値である。

2. $\frac{|\tilde{s}_J(f, t)|}{|\tilde{s}_Q(f, t)|} > \epsilon$ である個数 η を調べる。

3. η とフレーム t におけるアクティブな信号数 $N(t)$ によって $M_J(f, t)$ を決定する。

- (a) $\eta \leq N(t)$ のとき $M_J(f, t) = 1 \quad J=1, \dots, m$
(b) $\eta > N(t)$ のとき

$$M_J(f, t) = \begin{cases} 1 + \alpha_{on} & \frac{|\tilde{s}_J(f, t)|}{|\tilde{s}_Q(f, t)|} > \epsilon \cap \\ & MeanVector_J(t) \geq \zeta \\ \alpha_{off} & otherwise \end{cases} \quad (14)$$

ここで $0 \leq \alpha_{on}, \alpha_{off} \leq 1$ 、 ζ はアクティブである信号の中で $MeanVector_J(t)$ が $N(t)$ 番目に大きな値である。つまり、スペクトラムの平均値が $N(t)+1$ 番目以降に 小さな値である $\tilde{s}_J(f, t)$ を特に修正する。

4.

$$\tilde{y}_J(f, t) = M_J(f, t) \tilde{s}_J(f, t) \quad (15)$$

を計算し、逆フーリエ変換することで推定音源信号を得る。

5 実験

音源数 5、観測点数 3 として、ノイズのない場合と 13dB のノイズを加えた場合の実験を行った。

図 7 に示す音源数 $m = 5$ の音声信号に、式 (16) に示す混合行列 A を用いて観測点 $n = 3$ として実験を行った。サンプリング周波数は 8kHz、信号長は 2 秒である。

$$A = \begin{bmatrix} 0.8412 & -0.0298 & -0.0750 & -0.1735 & 0.7240 \\ -0.5025 & 0.8305 & -0.8294 & 0.3621 & 0.4088 \\ 0.1197 & 0.5563 & 0.5536 & 0.9158 & -0.5556 \end{bmatrix} \quad (16)$$

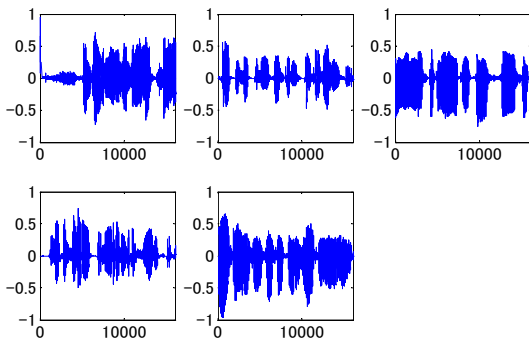


図 7: 実験に用いた原信号 $m = 5$

5.1 ノイズのない場合

図 8 に実験に用いた原信号の各フレーム毎の、 $l_0 - norm$ の数の集計を示す。 $l_0 - norm$ とは、各時間周波数領域の点における実際のアクティブである信号数を表しており、雑音軽減マスク推定の際には η として推定される。図 8 では、各フレームに含まれるサンプル数 $L = 200$ としている。このとき各フレームに対するフーリエ変換のポイント数は 200 とし、各フレームの各周波数ポイント毎に $l_0 - norm$ を求めた。図 8 の横軸はフレームの番号を表し、縦軸はフーリエ変換後のすべての周波数ポイントの $l_0 - norm$ の集計である。 $l_0 - norm = 0$ が多い場合、そのフレームではアクティブでない点が多かったことになる。実験に用いた原信号の場合、 $l_0 - norm$ は最高 5 である。図では、各フレームで全周波数点での $l_0 - norm$ を集計したものを表しており、 $l_0 - norm = 0, 1$ の数が多いフレームほどスパース性が高いと言える。各フレームの $l_0 - norm$ の平均が大きくなるほどそのフレームのアクティブ信号数が多いことになり、雑音軽減マスクを決定する際の各フレームにおける推定アクティブ信号数 $N(t)$ が大きく推定される。図より $l_0 - norm$ が 2 以上の点が多く存在していることがわかり、実験で用いた信号が完全なスパースではないことがわかる。

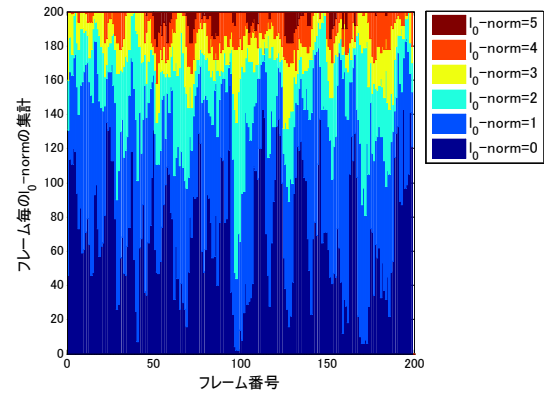


図 8: フレーム毎の $l_0 - norm$ の集計

推定混合行列は、

$$A = \begin{bmatrix} 0.8327 & -0.0458 & -0.0456 & -0.2198 & 0.7757 \\ -0.5014 & 0.8582 & -0.8424 & 0.3073 & 0.3394 \\ 0.1960 & 0.5113 & 0.5063 & 0.9207 & -0.5265 \end{bmatrix}$$

線形計画問題適用後の推定音源を図 9 を示す。さらに、提案法である雑音軽減マスクを適用した。このとき、 $\alpha_{on} = 0.5, \alpha_{off} = 0.5, \epsilon = 1/2$ とした。新たな推定信号が図 10 である。

音源信号の各点について、信号に対する原信号との差の比率である SDR(Signal to Difference Ratio) を求め、 ϵ を変えながら信号毎に平均を求めたものが図 11 である。

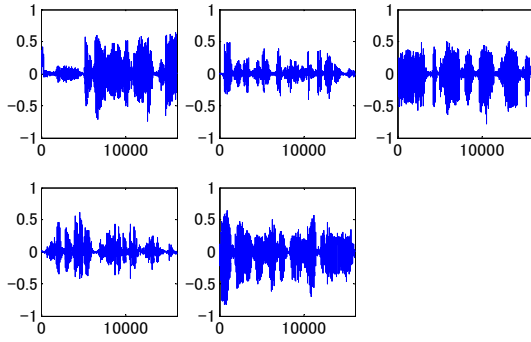


図 9: 雑音軽減マスクを適用前の推定原信号 (ノイズなし)

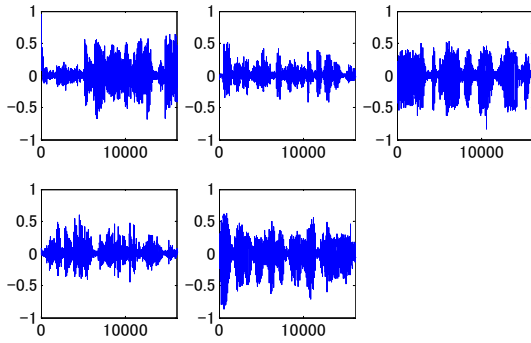


図 10: 雑音軽減マスクを適用後の推定原信号 (ノイズなし)

また、SDR では原信号が 0 の点について評価ができないため、原信号で 0 である点のうち推定後も 0 と判断された点の割合 (0 判定率) を図 12 に示す。各計算式は、

$$SDR_i = 10 \log \left[\frac{|s_i(t) - \bar{s}_i(t)|^2}{|s_i(t)|^2} \right] \quad (17)$$

$$Mean SDR_i = \frac{1}{T} 10 \sum_t \log \left[\frac{|s_i(t) - \bar{s}_i(t)|^2}{|s_i(t)|^2} \right] \quad (18)$$

$$Zero Determined Ratio = \frac{N_R}{N_S} \quad (19)$$

である。ただし、 s_i は S の i 番目の信号であり、 N_S は原信号 S のうち i 番目の信号のうち値が 0 であった点の数、 N_R は N_S に含まれる点のうち、音源分離後に 0 と判断された点の数である。

ϵ を変化した際の MeanSDR と 0 判定割合の平均を表 1 に示す。図 11、12 より、特に 0 の判定割合が上がり全体の SDR もほとんどの信号で向上したことがわかる。また聴覚上の変化としては、マスク適用前では 1 つのメイン音声に加えて、ノイズとして別の音声が聞こえていたが、適用後にはわずかなノイズは残るものの主音声を害するほどの大きさではなくなった。

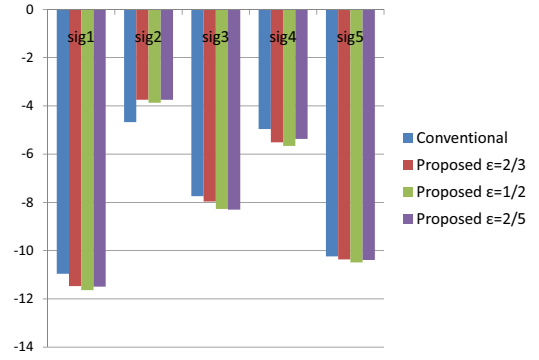


図 11: ϵ による MeanSDR の変化 (ノイズなし)

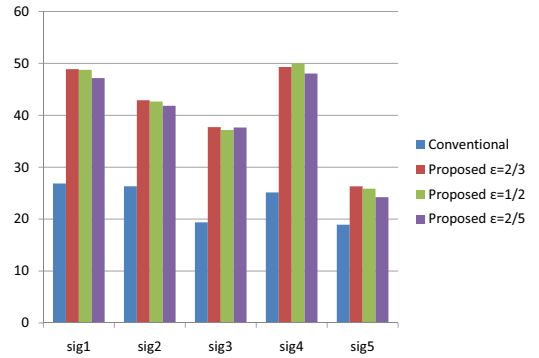


図 12: ϵ による 0 判定割合の変化 (ノイズなし)

5.2 ノイズがある場合

次に、ノイズを加えて同様の実験を行った。混合行列はノイズのない場合と同じものを用いて、SN 比 13dB のホワイトノイズを加えた。

推定混合行列は、

$$A = \begin{bmatrix} 0.7819 & -0.0011 & 0.0059 & -0.1152 & 0.9957 \\ -0.5519 & 0.8120 & -0.8152 & 0.2179 & 0.3214 \\ 0.1987 & 0.4687 & 0.5047 & 0.9228 & -0.4573 \end{bmatrix}$$

と推定された。SDR と 0 判定率は図 13、図 14 のようになった。

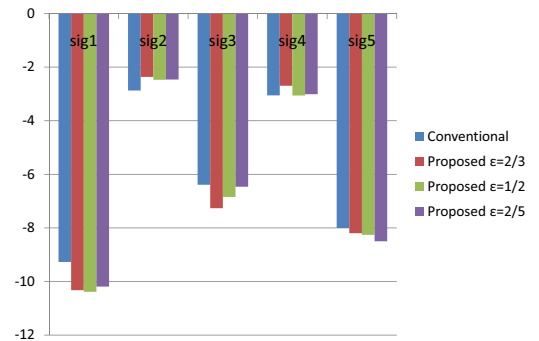


図 13: ϵ による MeanSDR の変化 (ノイズあり)

	平均 MeanSDR[dB]	平均 0 判定割合
従来法	-7.7147	23.3
$\epsilon=2/3$	-7.8104	41.0
$\epsilon=1/2$	-7.9852	40.1
$\epsilon=2/5$	-7.8601	39.8

表 1: ノイズのない場合の平均 MeanSDR と 0 判定割合

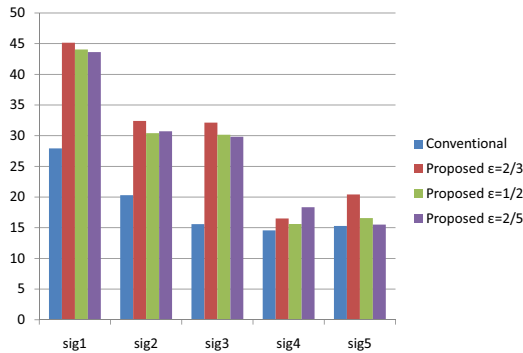


図 14: ϵ による 0 判定割合の変化 (ノイズあり)

同様に、ノイズのある場合の平均 MeanSDR と平均 0 判定割合を表 2 に示す。

	平均 MeanSDR[dB]	平均 0 判定割合
従来法	-5.9190	18.7
$\epsilon=2/3$	-6.1693	29.3
$\epsilon=1/2$	-6.2036	27.4
$\epsilon=2/5$	-6.1245	27.6

表 2: ノイズのある場合の平均 MeanSDR と 0 判定割合

ノイズのある場合も同様に、ノイズ軽減マスクによりノイズの軽減に成功した。聴覚上の変化としても、ノイズの軽減が感じられたが弱冠信号の劣化が見られた。

6 結論

従来の、混合行列推定法によるブラインド音源分離の欠点である線形計画問題によって出現した推定誤りを修正する雑音軽減マスクを提案した。実験により、完全にスパースではない混合信号の音源分離における、線形計画問題の推定誤りによるノイズの軽減に成功した。

参考文献

- [1] A.Hyvärinen and E.Oja, “Independent Component Analysis: algorithms and applications”, Neural

Networks, vol. 13, Issues 4-5, pp.411-430, June 2000.

- [2] Ö.Yilmaz and S.Rickard, “Blind separation of speech mixtures via time-frequency masking”, IEEE Trans, on SP, vol.52,no.7,pp.1830-1847, July 2004
- [3] Y.Z.Li, A.Cichocki, and S.Amari, “Underdetermined Blind Source Separation Based on Sparse Representaion,” IEEE Trans. Signal Processing, vol. 54, NO.2, pp.423-437, Feb. 2006.
- [4] 澤田, 荒木, 牧野, “音声のスパース性を用いた Underdetermined 音源分離”, 電子情報通信学会総合大会, AS-4-5, Mar. 2008.
- [5] Y.Luo, W.Wang, J.A.Chambers, “Exploitation of Source Nonstationarity in Underdetermined Blind Source Separation With Advanced Clustering Techniques”, IEEE Trans, on SP, vol.54,no.6,pp.2198-2212, June 2006.