周波数領域ブラインド音源分離のための極座標表示に基づく活性化関数*

◎澤田 宏, 向井 良, 荒木 章子, 牧野 昭二
(日本電信電話株式会社，NTT コミュニケーション科学基礎研究所)

1 はじめに

ブラインド音源分離 (BSS: Blind Source Separation) は，線形混合された複数の音を，混合された音だけから分離することである．瞬時混合の問題に対しては，独立成分分析 (ICA: Independent Component Analysis) [1, 2, 3] の技術を用いることで十分な分離性能が得られている．しかしながら，残響の影響が強い場合には，十分な性能が得られていないのが現状である．

残響に対処する一つの手法として，周波数領域での解法が盛んに研究されている [4, 5, 6, 7] ．そこでは，短時間フーリエ変換により，時間領域の信号が周波数 bin 毎の信号に変換される．すると，時間領域における畳み込み混合の問題は，周波数 bin 毎の瞬時混合の問題に変換されるため，従来の ICA アルゴリズムが適用できる．ただし，フーリエ変換の結果を扱うため，ICA では複素数を扱う必要があり，非線形の活性化関数も複素数に拡張する必要がある．これまでに，複素数を実部と虚部に分解して別々に非線形関数を適用する方法が提案されている [4]．しかしこの方法では，余分な制約が発生して収束を阻むことがある [6]．対応策としては，分離行列の更新式を別のものにして，余分な制約を発生させない方法が提案されている [5, 6]．

本稿では，複素数の極座標表示に基づく新たな活性化関数を提案する．これを用いると，本来の分離行列の更新式を用いても上記の余分な制約が発生しない．その結果として，これまでよりも改善された SNR (Signal-to-Noise Ratio) が得られたことを示し，考察を行う．

2 周波数領域 BSS

互いに独立な N 個の音信号 $s_p(t)$, $(1 \le p \le N)$ が室内で残響も含めて混合され，M 個のマイク $x_q(t) = \sum_{p=1}^{N} \boldsymbol{h}_{qp} * s_p(t)$, $(1 \le q \le M)$ で観測されたとする．ここで，$\boldsymbol{h}_{qp}$ は音源 p からマイク q へのインパルス応答，$*$ は畳み込みを示す．これを短時間フーリエ変換を用いて周波数領域に変換し，行列形式で表現すると，$\mathbf{X}(\omega, m) = \mathbf{H}(\omega)\mathbf{S}(\omega, m)$ となる．周波数領域 BSS では，各周波数 ω において，この瞬時混合問題をそれぞれ解く．すなわち，$\mathbf{H}(\omega)$ を知らずに，$\mathbf{X}(\omega, m)$ を互いに独立な N 個の信号 $\mathbf{Y}(\omega, m) = \mathbf{W}(\omega)\mathbf{X}(\omega, m)$ に分離する行列 $\mathbf{W}(\omega)$ を求める．

各周波数 bin において分離行列 $\mathbf{W}$ は，$\mathbf{Y}$ の各要素間の相互情報量の最小化を目指して，学習則 $\mathbf{W}_{i+1} = \mathbf{W}_i + \Delta\mathbf{W}_i$ により徐々に改良される [1, 2]．$\Delta\mathbf{W}$ の計算には自然勾配法 (natural gradient) [3], $\Delta\mathbf{W} = \mu\left[\mathbf{I} - \langle\varphi(\mathbf{Y})\mathbf{Y}^T\rangle\right]\mathbf{W}$, が広く用いられている．ここで，$\mu$ はステップサイズ，$\langle\cdot\rangle$ は時間平均を表す．$\varphi(\cdot)$ は活性化関数であり，$\varphi(\mathbf{Y}) = \tanh(\eta \cdot \mathbf{Y})$ が非線形の活性化関数として広く用いられている [1, 2]．η は φ の非線形性の強さを制御するパラメータである．

周波数領域 BSS では，短時間フーリエ変換の結果が複素数であるため，$\Delta\mathbf{W}$ と活性化関数 $\varphi(\cdot)$ を複素数に拡張する必要がある．文献 [4] では，

$$\Delta\mathbf{W} = \mu\left[\mathbf{I} - \langle\Phi(\mathbf{Y})\mathbf{Y}^H\rangle\right]\mathbf{W} \tag{1}$$

$$\Phi(\mathbf{Y}) = \varphi[re(\mathbf{Y})] + j \cdot \varphi[im(\mathbf{Y})] \tag{2}$$

という拡張が提案された．ここで，$\mathbf{I}$ は単位行列，$\mathbf{Y}^H$ は $\mathbf{Y}$ の共役転置，$re(\mathbf{Y})$ と $im(\mathbf{Y})$ はそれぞれ $\mathbf{Y}$ の実部と虚部である．式 (1) に従うと $\mathbf{W}$ は

$$\langle\Phi(Y_p)Y_q^*\rangle = 0 \quad (p \ne q) \tag{3}$$

$$\langle\Phi(Y_p)Y_q^*\rangle = 1 \quad (p = q) \tag{4}$$

を満たす点に収束する．ここで，Y_q^* は Y_q の複素共役である．式 (3) は Y_p と Y_q が互いに独立になるように働く．式 (4) は Y_p の振幅の平均値をある値に近づける．式 (4) を実部と虚部に分解すると，

$$\langle\varphi[re(Y_p)]re(Y_p) + \varphi[im(Y_p)]im(Y_p)\rangle = 1 \tag{5}$$

$$\langle\varphi[im(Y_p)]re(Y_p) - \varphi[re(Y_p)]im(Y_p)\rangle = 0 \tag{6}$$

となる．ここで式 (6) が余分な制約を課していることがわかる [6]．例えば $re(Y_p)$ と $im(Y_p)$ が互いに独立であればこの制約を満たすが，このような制約は望ましくない．そこで，$\Delta\mathbf{W}$ を計算する別の式が

$$\Delta\mathbf{W} = \mu\left[\mathrm{diag}(\langle\Phi(\mathbf{Y})\mathbf{Y}^H\rangle) - \langle\Phi(\mathbf{Y})\mathbf{Y}^H\rangle\right]\mathbf{W} \tag{7}$$

提案された [5]．これによると，$\mathbf{W}$ は式 (3) のみを満たす点に収束し，$\mathbf{Y}$ の振幅の平均はそれほど変化しない．

3 極座標表示に基づく活性化関数

余分な制約 (6) の問題を別の方法で解決するため，複素数の極座標表示に基づく新たな活性化関数を提案する．

$$\Phi(\mathbf{Y}) = \varphi[abs(\mathbf{Y})] \cdot e^{j \cdot angle(\mathbf{Y})}, \tag{8}$$

ここで，$abs(\mathbf{Y})$ と $angle(\mathbf{Y})$ はそれぞれ $\mathbf{Y}$ の各要素の絶対値と偏角を表す．この活性化関数は，元の $\mathbf{Y}$ の絶対値のみを変更し，その偏角は変化させない．これは，実数に対する活性化関数 $\varphi(\cdot)$ の自然な拡張であり，実数に対しては双方とも同じ値を出力する．

この活性化関数を使うことで，式 (6) の様な余分な制約は発生しない．なぜなら，$\theta = angle(Y_p)$ とすると，Y_p^* が Y_p の複素共役であることから，

$$\begin{aligned}\Phi(Y_p)Y_p^* &= \varphi[abs(Y_p)] \cdot e^{j\theta} \cdot abs(Y_p) \cdot e^{-j\theta} \\ &= \varphi[abs(Y_p)] \cdot abs(Y_p)\end{aligned}$$

となり，式 (4) の $\langle\Phi(Y_p)Y_p^*\rangle$ における虚部は 0 となる．

*A polar-coordinate based activation function for frequency domain blind source separation, by Hiroshi Sawada, Ryo Mukai, Shoko Araki, Shoji Makino (NTT Communication Science Laboratories, NTT Corporation)

表 1: 異なる Φ と $\Delta\mathbf{W}$ に対する SNR (dB)

	$T_R = 150$ ms		$T_R = 300$ ms	
	Ref	Imp	Ref	Imp
Polar-I	18.3	19.7	12.7	16.3
Cartesian-I	17.9	19.4	12.3	15.6
Cartesian-diag	17.8	18.0	11.9	14.6

$\mu = 0.1$, $\eta = 100$, #iteration $= 100$
Ref: 音声信号自身で計測した SNR
Imp: インパルスで計測した SNR [7]

4 実験結果および考察

提案した活性化関数の有効性を示すため，以下に示す 3 種類の組合せに関して，残響下で混合された音声を分離する実験を行った．

Polar-I Φ に式 (8)，$\Delta\mathbf{W}$ に式 (1) を用いる
Cartesian-I Φ に式 (2)，$\Delta\mathbf{W}$ に式 (1) を用いる
Cartesian-diag Φ に式 (2)，$\Delta\mathbf{W}$ に式 (7) を用いる

混合音声は，ASJ 研究用音声コーパス中から選んだ 8 秒の音声データと部屋のインパルス応答を計算機上で畳み込んで作成した．音源数，マイク数は共に 2 である．サンプリング周波数は 8kHz である．2 種類の残響時間 $T_R = 150$ms (1200 サンプル), 300ms (2400 サンプル) に対して行い，短時間フーリエ変換のフレーム長はそれぞれ 1024,2048，シフト量は共に 256 とした．表 1 に，2 つの出力 Y_1, Y_2 の SNR の平均を示す．総じて “Polar-I” の結果が他を上回った．

まず，“Polar-I” と “Cartesian-I” の比較に際して，特に “Cartesian-I” に課せられている余分な制約 (6) に関して考察する．図 1 は，活性化関数 (2) および (8) を用いた場合の $[\mathbf{I} - \langle\Phi(\mathbf{Y})\mathbf{Y}^H\rangle]$ の絶対値を示している．これらのデータは，$T_R = 150$ ms の場合の第 100 番目 (773.4 Hz) の周波数ビンのものである．“Cartesian” には収束を妨げている振動が見られるが，“Polar” では滑らかに収束しており，“Cartesian” の場合と比べて $\mathbf{Y}$ の相互情報量 (非対角成分 [1,2] と [2,1]) は明らかに小さくなっている．なお，“Cartesian” の場合の振動は，2 章での議論の通り，$[\mathbf{I} - \langle\Phi(\mathbf{Y})\mathbf{Y}^H\rangle]$ の対角成分の虚部が起因していることを確認した [8]．

次に，“Polar-I” と “Cartesian-diag” の比較のため，図 2 に $T_R = 300$ ms の場合の収束速度の違いを示す．“Cartesian-diag” においては，$\mathbf{W}$ の更新回数 (#iteration) が 100 回の段階では十分に収束していないことがわかる．理由は，式 (4) が働かないことから，周波数 bin 毎で $\mathbf{Y}$ の大きさが揃わず，全体で見ると収束速度に大きなばらつきがあるためである．収束を速くするためにステップサイズ μ を大きくする対応策も考えられるが，その場合は，元々 $\mathbf{Y}$ が大きい周波数ビンで $\mathbf{W}$ が発散する危険性がある．

5 おわりに

周波数領域 ICA のための新たな活性化関数を提案した．これを用いると，制約 (6) が影響しないため，式 (1) を使っても収束を妨げる振動は発生しない．従って，制約 (6) を避けるために，周波数毎で収束速度がばらつく式 (7) を使わなくても良くなる．残響下での音声の分離に関する実験により，新たな活性化関数の有効性を示した．

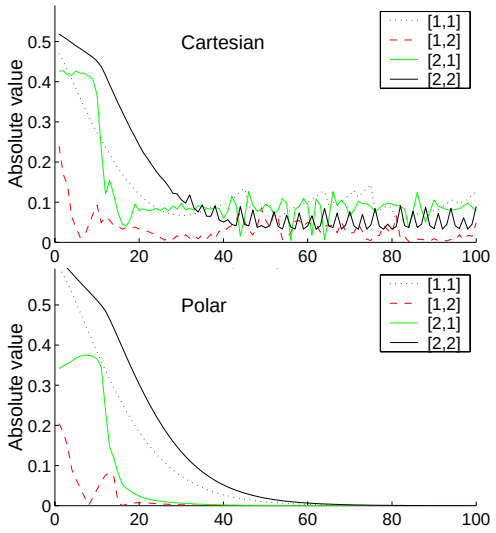

図 1: $[\mathbf{I} - \langle\Phi(\mathbf{Y})\mathbf{Y}^H\rangle]$ の絶対値

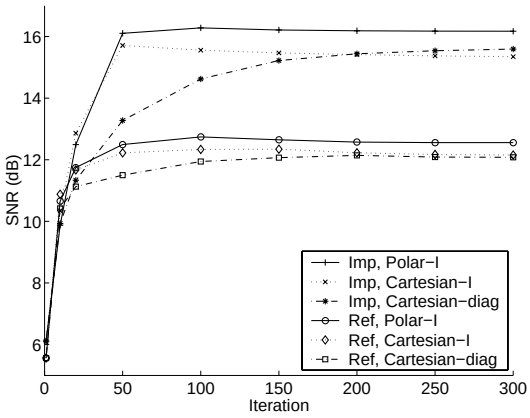

図 2: 収束速度

参考文献

[1] A. J. Bell and T. J. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[2] T. W. Lee, *Independent component analysis - Theory and applications*, Kluwer academic publishers, 1998.

[3] S. Amari, A. Cichocki, and H. Yang, “A new learning algorithm for blind signal separation,” in *Advances in Neural Information Processing Systems*. 1996, vol. 8, pp. 757–763, The MIT Press.

[4] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[5] N. Murata and S. Ikeda, “An on-line algorithm for blind source separation on speech signals,” in *Proc. NOLTA '98*, 1998, pp. 923–926.

[6] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, “Blind signal separation using directivity pattern,” in *Technical Report of Japanese Society for Artificial Intelligence*, Nov. 1999, pp. 21–26.

[7] R. Mukai, S. Araki, and S. Makino, “Separation and dereverberation performance of frequency domain blind source separation for speech in a reverberant environment,” in *Proc. Eurospeech2001*, Sept. 2001.

[8] H. Sawada, R. Mukai, S. Araki, and S. Makino, “A polar-coordinate based activation function for frequency domain blind source separation,” submitted to ICA2001.