2—5—11

ブラインド音源分離後の残留スペクトルの推定と除去*

○向井 良　澤田 宏　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

実環境におけるブラインド音源分離 (BSS) の問題点の一つに，残響下での分離性能の低下がある．周波数領域での独立成分分析 (ICA) は，残響下での BSS に用いられる代表的な手法であるが，残響をカバーするために長いフレーム (フィルタ) を用いると，かえって分離性能が低下してしまう．これは，数秒程度のデータを用いる場合には，推定すべきパラメータの数が増える一方で，周波数あたりのデータサンプルが少なくなり，全体として推定誤差が増大するためである [1].

我々はこれまでに，まず短いフレーム長の周波数領域 BSS で分離を行い，その後に分離音中に含まれる残留雑音のスペクトルを推定して除去する手法 (Time-Delayed Spectral Subtraction, TDSS) を提案した [2]．これは，ICA による分離では，比較的短いフレームでも妨害音の直接音は十分除去されていること，および分離音中に含まれる残留雑音が主に妨害音の残響に起因することに着目した手法である．

TDSS では周波数領域における各チャンネル間の遅延と漏洩係数を推定パラメータとして残留雑音のモデル化を行ったが，今回はこれを一般化し，周波数領域における分離信号をフィルタに通すことでスペクトルを推定する手法を提案する [3].

2 アルゴリズム

2.1 残留雑音のモデル化

本システムの全体のブロック図を図 1 に示す．周波数領域において，音源信号 $\boldsymbol{S}(\omega,n) = [S_1(\omega,n),...,S_N(\omega,n)]^T$ ，観測信号 $\boldsymbol{X}(\omega,n) = [X_1(\omega,n),...,X_M(\omega,n)]^T$ ，ICA による分離信号 $\boldsymbol{Y}(\omega,n) = [Y_1(\omega,n),...,Y_N(\omega,n)]^T$ とする．ここで ω は周波数，n はフレーム番号 (時間)，N は音源数，M はマイクロホンの数である．

$Y_i(\omega,n)$ は，目的信号 $S_i(\omega,n)$ に起因する出力であるストレート成分 $Y_i^{(s)}(\omega,n)$ と，妨害音 $S_j(\omega,n)(j \neq i)$ に起因するクロス成分 $Y_i^{(c)}(\omega,n)$ の和 $Y_i(\omega,n) = Y_i^{(s)}(\omega,n) + Y_i^{(c)}(\omega,n)$ であり，$Y_i^{(c)}(\omega,n)$ が分離音中に含まれる残留雑音である．図 2 は 2 入力 2 出力の場合を示したものである．図 3 は，実際の音声信号による $|Y_1^{(c)}(\omega,n)|$，$|Y_2^{(s)}(\omega,n)|$ の例であるが，これらの間には明らかに関係がある．

TDSS ではこの関係を，各周波数 ω およびチャンネルの組 ij ごとの遅延パラメータ $\tau_{ij}(\omega)$，漏洩係数 $\alpha_{ij}(\omega)$ を用い，

$$|Y_i^{(c)}(\omega,n)|^\beta \approx \sum_{j\neq i} \alpha_{ij}(\omega)|Y_j^{(s)}(\omega,n-\tau_{ij}(\omega))|^\beta \quad (1)$$

*Estimation and elimination of residual crosstalk components in blind source separation, Ryo Mukai, Hiroshi Sawada, Shoko Araki, and Shoji Makino (NTT Communication Science Laboratories, NTT Corporation)

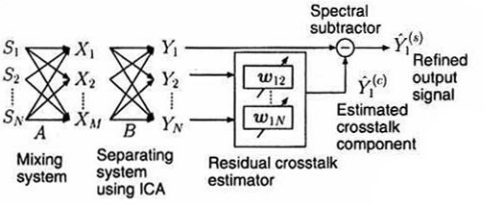


図 1: 提案手法のブロック図

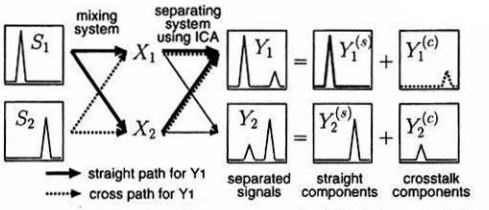


図 2: 周波数領域で見たストレート成分とクロス成分

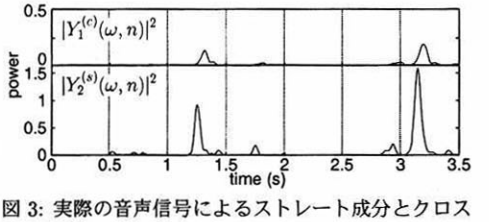


図 3: 実際の音声信号によるストレート成分とクロス成分の例

とモデル化した．$\beta = 1$ の場合は振幅スペクトル，$\beta = 2$ の場合はパワースペクトルによる処理になる．今回提案する手法では，τ_{ij}，α_{ij} の代わりに長さ L の FIR フィルタ $\boldsymbol{w}_{ij}' = [w_{ij0},...,w_{ijL-1}]^T$ を用い，

$$|Y_i^{(c)}(\omega,n)|^\beta \approx \sum_{j\neq i}\sum_{k=0}^{L-1} w_{ijk}(\omega,n)|Y_j^{(s)}(\omega,n-k)|^\beta \quad (2)$$

とモデル化する．今回のモデルは TDSS のモデルを包含しており，より一般化したものである．

2.2 フィルタの推定

$\boldsymbol{w}_{ij}$ は $|Y_j^{(s)}(\omega,n)|^\beta$ を入力としたときに $|Y_i^{(c)}(\omega,n)|^\beta$ を出力するようなフィルタである．これを Normalized LMS (NLMS) [4] を基にした適応アルゴリズムによって推定する．ここで，$|Y_j^{(s)}|$ と $|Y_i^{(c)}|$ は実際には未知であるため，$|Y_j|$，$|Y_i|$ で近似する必要がある．

音源信号として互いに独立な音声を想定した場合，狭帯域信号 $|Y_i^{(s)}(\omega,n)|$ と $|Y_j^{(s)}(\omega,n)|$ が同時刻に大きな値をもつことは少ない [5, 6]．この性質を利用

表 1: 実験条件

Common	Sampling rate = 8 kHz Window = hanning Reverberation time t_R=150 ms, 300 ms Length of source signal = 6 s
ICA part	Frame length T_{ICA}= 32, 512 points (4, 64 ms) Frame shift = frame length/4 $\mu = 0.1$, $g = \infty$ Number of iterations = 100
NLMS & Spectral subtraction part	Frame length T_{SS}= 1024 points (128 ms) Frame shift = 64 points (8 ms) Filter length L = 16 $\mu = 0.1$, $\delta = 0.01$, $\beta = 2$

し，チャンネル番号を要素とする集合 $\mathcal{I}_S(\omega,n) = \{i\colon |Y_i(\omega,n)| \approx |Y_i^{(s)}(\omega,n)|\}$ および $\mathcal{I}_C(\omega,n) = \{i\colon |Y_i(\omega,n)| \approx |Y_i^{(c)}(\omega,n)|\}$ を考える．これはチャンネル番号 $i \in \mathcal{I}_S(\omega,n)$ については $|Y_i^{(s)}(\omega,n)|$ は $|Y_i(\omega,n)|$ で近似でき，$i \in \mathcal{I}_C(\omega,n)$ については $|Y_i^{(c)}(\omega,n)|$ は $|Y_i(\omega,n)|$ で近似できることを意味している．これらの集合を用いて，選択的にフィルタ係数を更新していくことにより $\boldsymbol{w}_{ij}$ を推定する．

$\boldsymbol{u}_j(\omega,n) = [|Y_j(\omega,n)|^\beta, |Y_j(\omega,n-1)|^\beta, \ldots, |Y_j(\omega,n-L+1)|^\beta]^T$ をタップ入力ベクトル，$e_i(\omega,n) = |Y_i(\omega,n)|^\beta - \sum_{j\neq i} \hat{\boldsymbol{w}}_{ij}^T(\omega,n)\boldsymbol{u}_j(\omega,n)$ を推定誤差とおくと，更新式は以下のようになる．

$$\Delta\hat{\boldsymbol{w}}_{ij}(\omega,n+1) = \begin{cases} \dfrac{\mu}{\delta + \|\boldsymbol{u}_j(\omega,n)\|^2}\boldsymbol{u}_j(\omega,n)e_i(\omega,n) \\ \quad (\text{if } i \in \mathcal{I}_C(\omega,n), \text{ and } j \in \mathcal{I}_S(\omega,n)) \\ 0 \ (\text{otherwise}) \end{cases} \tag{3}$$

ここで，μ はステップサイズパラメータ，δ は $\boldsymbol{u}_j$ が小さいときに計算が不安定になるのを防ぐための正の定数である．

(3) で求めた $\hat{\boldsymbol{w}}_{ij}$ と (2) のモデルから，$Y_i^{(c)}$ のスペクトルを推定する．ここでも $Y_j^{(s)}$ は未知であるため，Y_j を近似値として用いる．

$$|\hat{Y}_i^{(c)}(\omega,n)|^\beta \approx \sum_{j\neq i} \hat{\boldsymbol{w}}_{ij}^T(\omega,n)\boldsymbol{u}_j(\omega,n). \tag{4}$$

これを Y_i から引くことにより，ストレート成分のスペクトルを求める．

$$\hat{Y}_i^{(s)}(\omega,n) = \begin{cases} (|Y_i(\omega,n)|^\beta - |\hat{Y}_i^{(c)}(\omega,n)|^\beta)^{1/\beta}\dfrac{Y_i(\omega,n)}{|Y_i(\omega,n)|} \\ \quad (\text{if } |Y_i(\omega,n)| > |\hat{Y}_i^{(c)}(\omega,n)|) \\ 0 \quad (\text{otherwise}) \end{cases} \tag{5}$$

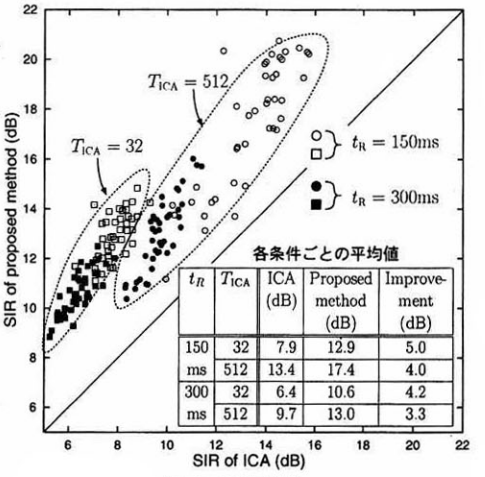
3 実験

3.1 実験方法

音源数 2 の場合について実験を行った．実験条件を表 1 に示す．$\boldsymbol{w}_{ij}(\omega,n)$ の選択的更新には，単純な大小比較による以下のルールを用いた．

if $|Y_1(\omega,n)| > |Y_2(\omega,n)|$
then $\mathcal{I}_S(\omega,n) = \{1\}$, $\mathcal{I}_C(\omega,n) = \{2\}$
else $\mathcal{I}_S(\omega,n) = \{2\}$, $\mathcal{I}_C(\omega,n) = \{1\}$

ASJ 研究用連続音声コーパス中の話者 4 名 (男声 2, 女声 2)，2 通りの文の音声を用い，計 42 通りの組み合わせについて ICA のみによる分離性能と提案手法による性能を評価した．学習データ，評価データとも 6 秒である．ICA 部分のアルゴリズムは文献 [2, 3] と同じものを用いた．ストレート成分 $y_i^{(s)}$ をリファレンス信号として，出力 SIR を $\mathrm{SIR}_{Oi} \equiv 10\log(|y_i^{(s)}|^2/|\hat{y}_i^{(s)} - y_i^{(s)}|^2)$ (dB) と定義し，SIR_{O1} と SIR_{O2} の平均値を評価尺度として用いた．



t_R	T_{ICA}	ICA (dB)	Proposed method (dB)	Improvement (dB)
150	32	7.9	12.9	5.0
ms	512	13.4	17.4	4.0
300	32	6.4	10.6	4.2
ms	512	9.7	13.0	3.3

図 4: 実験結果

3.2 実験結果

実験結果を図 4 に示す．横軸は ICA の出力の SIR，縦軸は提案手法による SIR であり，一つの点は音声の組合せに対応している．ICA 部のフィルタ長と残響時間を変えたときの SIR の平均値および改善量をグラフ中の表にまとめた．すべての組合せにおいて性能が向上しており，平均 3〜5 dB 程度の改善がみられた．

4 まとめ

ICA によって求めた分離音中に残留するクロス成分のスペクトルを，フィルタを用いてモデル化し，推定・除去する手法を提案した．フィルタは，音声を狭帯域信号で見たときのスパース性を利用し，選択的更新を行う NLMS アルゴリズムによって推定した．残響下での混合音声を用いた実験によって本手法の効果を確認した．

参考文献

[1] 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. of Eurospeech'01*, 2001, pp. 2599–2602.

[2] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Removal of residual cross-talk components in blind source separation using time-delayed spectral subtraction," in *Proc. of ICASSP'02*, 2002, pp. 1789–1792.

[3] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Removal of residual cross-talk components in blind source separation using LMS filters," in *Proc. of NNSP'02*, 2002, accepted.

[4] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 2002.

[5] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoust. Sci. & Tech.*, vol. 22, no. 2, pp. 149–157, Feb. 2001.

[6] S. Rickard and Ö. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. of ICASSP'02*, 2002, pp. 529–532.