

帯域に適した分離手法を用いるサブバンド領域ブラインド音源分離*

○荒木 章子¹, 牧野 昭二¹, △Robert Aichner^{1,2}, 西川 剛樹³, 猿渡 洋³¹ 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,² University of Erlangen-Nuremberg, ³ 奈良先端科学技術大学院大学

1. はじめに

ブラインド音源分離 (Blind Source Separation: BSS) は、観測された混合信号のみから音源信号を推定する手法であり、音源信号同士の独立性の仮定に基づく手法をはじめ近年多くの手法が提案されている [1]。

本稿では、実環境で混合された音声信号の分離問題を扱う。音声信号は低域で大きなパワーを持つため、低域での分離性能の向上は分離音声全体の分離性能の向上のため重要である。そこで本稿では、帯域毎の処理に向くサブバンド領域ブラインド音源分離 (サブバンド BSS) [2] を用い、低域での分離性能を高めることを考える。具体的には (1) 一般に低域では高域に比べて残響時間が長いこと、低域において長い分離フィルタを利用する、(2) 分離フィルタの推定精度を上げるため、低域においてデータを利用する際にオーバーラップシフトを用いてデータを有効利用する、の 2 点を提案し、これらの低域のみへの適用で効率的に高い分離性能が得られることを示す。

2. 混合信号のモデルと分離系

実環境では、音源信号に残響が畳み込まれ混合されて観測される。この観測信号を次のようにモデル化する。

音源 i からの信号 s_i に部屋のインパルス応答 h_{ji} が畳み込まれた信号が混合し、マイク j で信号 x_j として観測されるとする。

$$x_j(n) = \sum_{i=1}^{N_s} \sum_{k=1}^P h_{ji}(k) s_i(n-k+1) \quad (j=1, \dots, N_m). \quad (1)$$

N_s は音源数, N_m はマイク数である。本稿では $N_s = N_m = 2$ とする (図 1)。

推定する分離系は Q タップの分離フィルタ群 w_{ij} から成り、分離信号は次のように表される。

$$y_i(n) = \sum_{j=1}^{N_m} \sum_{k=1}^Q w_{ij}(k) x_j(n-k+1) \quad (i=1, \dots, N_s). \quad (2)$$

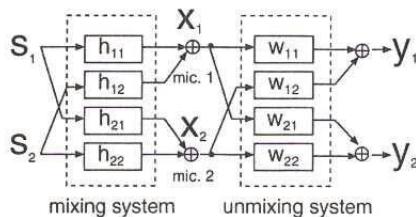


図 1 BSS system configuration.

*Subband based blind source separation with appropriate processing for each frequency band, by S. Araki, S. Makino (NTT Corporation), R. Aichner (Univ. of Erlangen-Nuremberg), T. Nishikawa, and H. Saruwatari (Nara Institute of Science and Technology).

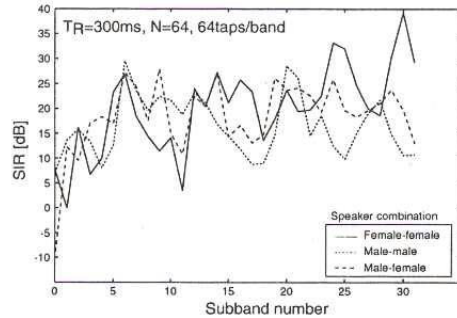


図 2 SIR in each subband.

3. 帯域に適した分離手法による BSS

一般に BSS では、低域における分離性能が低い。図 2 は、3 組の混合音声に対し、後述するサブバンド BSS を用いて分離した信号の各帯域における信号対妨害音比 (SIR) を示している。どの組についても、低域において分離性能が低いことが分かる。

低域で分離性能が低い理由の一つとして、一般に低域では高域に比べて残響時間が長いことが考えられる。そこでまず (1) 残響に対応するため低域で長い分離フィルタを用いること、が考えられる。また、音声信号は低域で大きなパワーを持つため、低域での分離性能の向上は音声分離の鍵である。そこで、(2) 低域に関して推定に用いるデータを有効利用して分離フィルタの推定精度を上げることが考えられる。

本稿では、帯域毎に分離の方法を変えることが容易なサブバンド BSS を用いて、低域のみに (1)(2) を適用し、分離性能を調べる。

なお、低域ではマイク間隔に比べて波長が長いこと位相差が小さく分離が難しいという理由も考えられるが、本稿では固定された 2 マイクでの BSS を考え、この点については議論しない。

4. サブバンド BSS

図 3 にサブバンド BSS の構成を示す。まず入力された信号を、ポリフェーズフィルタバンクを用いてサブバンド分析する [3]。このとき、各帯域で信号を実数で扱うために Single Side Band (SSB) サブバンド [4] を用いる。次に各帯域において、時間領域 BSS を用いて分離を行う。最後に各帯域で分離された信号を合成し、出力を得る。

各帯域での時間領域 BSS アルゴリズムは、任意のものを用いることができるが、本稿では、信号の非定常性に基づく時間領域 BSS アルゴリズム [5] を用いる。

帯域 k での分離フィルタ $w^k(m) = [w_{ij}^k(m)]$ (ij 要素が $w_{ij}^k(m)$ である行列) は、出力信号 $Y_i^{sb}(k, m)$ が、複数の分析区間全てにおいて、時間ずれの相関まで含めて無相関となるよう学習される。

$$w_{i+1}^k(m) = w_i^k(m) + \frac{\alpha}{B} \sum_{b=1}^B \{ (\text{diag} \mathbf{R}_y^b(0))^{-1} (\text{diag} \mathbf{R}_y^b(m)) - (\text{diag} \mathbf{R}_y^b(0))^{-1} \mathbf{R}_y^b(m) \} * w_i^k(m) \quad (3)$$

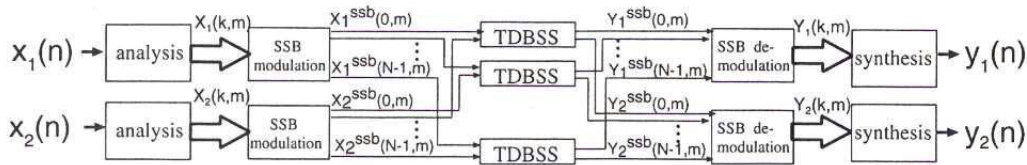


図3 System configuration of subband BSS. TDBSS: time-domain BSS.

ここで $R_y^b(\tau)$ は時間ブロック b ($b=0, \dots, B-1$) における出力信号の共分散行列 ($R_y^b(\tau) = \frac{1}{L} \sum_{m=1}^L y(b\frac{L}{S} + m)y^T(b\frac{L}{S} + m - \tau)$), $y(m) \equiv [Y_1^{ssb}(k, m), Y_2^{ssb}(k, m)]^T$, L はブロック長, S は時間ブロックに区切る際のシフト率, α はステップサイズである.

5. 実験

5.1 実験条件

実験に用いたセットアップはマイク間隔 4cm, 音源方位 -30° 及び 40° , 残響時間 $T_R = 300\text{ms}$ である. サンプル周波数は 8kHz であり, 300ms の残響時間は 2400tap に相当する. 混合音は音声に実測したインパルス応答を畳込んで作成した. 音声組合せは男男・男女・女女の 3 通りである. 学習データ, 分離データの長さはそれぞれ 3 秒, 約 8 秒である. 分離性能の評価尺度は, 信号対妨害音比 (SIR = 出力 SNR - 入力 SNR) である.

サブバンド BSS の帯域分割数 $N = 64$ とし, 各帯域での分離フィルタ長を 32 ~ 128tap と変化させた.

分離フィルタの初期値には $\pm 60^\circ$ 方向へ死角を作る死角型ビームフォーマを実現する $w(m)$ をサブバンド分析したものを用いた [6]. また時間領域 BSS アルゴリズムのステップサイズ α は 1.0×10^{-4} , ブロック数 B は 5.2 節で $B = 20$ を用いた. 5.3 節ではブロック長 L を 5.2 節でのブロック長と同じとした.

時間領域 BSS による信号の白色化歪に対しては, 各サブバンドにおける学習後の分離フィルタの利得が, 初期値における利得と同じになるよう正規化することで解決した.

5.2 実験 1: 低域のみ長い分離フィルタを使用

低域における長い残響をカバーするため, 低域のみ長い分離フィルタを使用した.

表 1 の (A)-(C) の no-overlap の列は, 高域 (帯域 No. 6 ~) に長さ 32tap 固定, 低域 (帯域 No. 0 ~ 5) に長さを変えた分離フィルタを用いた分離結果を示している. 低域のみで 32tap から 64tap と分離フィルタを長くすることで, 大幅に性能が上がる事が分かる.

5.3 実験 2: 低域のみオーバーラップシフトを使用

前節において, 低域で 128tap の分離フィルタを用いた時 ((C) の場合), 性能がかえって低下した. これは 128tap の学習に 3 秒の学習データでは足りなかったためと考えられる. そこで, 学習データ長を有効利用するため, 低域における時間領域 BSS でオーバーラップシフトを用いる.

表 1 の (B)-(F) は, 低域において 2 倍 (overlap $\times 2$. シフト率 $S = 2$) もしくは 4 倍 (overlap $\times 4$. シフト率 $S = 4$) のオーバーラップシフトを用いた結果を示している. オーバーラップシフトを用いれば, 128tap の分離フィルタの推定性能が向上し, 低域で 4 倍のオー

表 1 Separation performance of subband BSS. (A)-(F) the overlap-blockshift was executed only for bands 0-5, and (G) and (H) the overlap-blockshift was executed for all subbands

	band 0-5	band 6-32	no-overlap	overlap (x2)	overlap (x4)
(A)	32	32	5.82		
(B)	64	32	8.86	9.61	
(C)	128	32	8.71	9.75	10.09
(D)	64	64	10.31	10.80	10.79
(E)	128	64	10.28	11.21	12.01
(F)	128	128	10.28	11.22	12.00
(G)	64	64	10.31	10.88	10.99
(H)	128	128	10.28	11.22	12.14

(tap)

(SIR [dB])

バップシフトを用いて 128tap の分離フィルタを推定した時 ((E) の場合) に, 高い分離性能が得られた.

なお, 表 1 の (F) では全ての帯域で 128tap の分離フィルタを推定したが, 高域で半分の長さの分離フィルタを推定した (E) の場合と比べて性能の向上は見られなかった. 更に, (G) と (H) は全ての帯域にオーバーラップシフトを用いた結果を示しているが, (D) や (F) と比べて大きな性能の向上は得られなかった. 以上から, 低域のみ重点的に分離を行うことで, 全ての帯域で同じ処理を行う場合と同程度の性能が得られることが分かる.

6. まとめ

音声信号は低域で大きなパワーを持つため, 低域での分離性能の向上が重要である. これに対し本稿では, 帯域毎の分離処理に向くサブバンド領域ブラインド音源分離を用いて, 低域で長い分離フィルタ及びオーバーラップシフトを利用し, 効率的に高い分離性能を得た.

参考文献

- [1] S. Haykin, *Unsupervised Adaptive Filtering*, John Wiley & Sons, 2000.
- [2] 荒木 他, “サブバンド処理によるブラインド音源分離に関する検討,” 音講論, pp. 619-620, (2002.3).
- [3] M. R. Portnoff, “Implementation of the digital phase vocoder using the fast fourier transform,” *IEEE Trans. Speech Audio Processing*, Vol. 24, No. 3, pp. 243-248, June 1976.
- [4] R. Crochiere and L. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
- [5] T. Nishikawa et al., “Comparison of blind source separation methods based on time-domain ICA using nonstationarity and multistage ICA,” *IEICE Tech. Rep.*, pp. 45-52, Jan. 2002.
- [6] 荒木 他, “死角型ビームフォーマを初期値に用いる時間領域ブラインド音源分離,” 音講論, pp. 543-544, (2002.9).