音声のスパース性を用いたUnderdetermined音源分離

Sparseness based Underdetermined Blind Speech Separation

荒木章子　Shoko Araki
澤田宏　Hiroshi Sawada
牧野昭二　Shoji Makino

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
NTT Communication Science Laboratories, NTT Corporation

1　はじめに

ブラインド音源分離 (BSS) 手法として広く検討されている独立成分分析 (ICA) は、音源数 $N \leq$ マイク数 M の場合には実環境においても高い性能をあげている [1, 2] が、$N > M$ の場合には適用が難しい。これに対し、DUET[3] に代表される信号のスパース性に基づく方法は、各時間周波数における観測信号のマイク間振幅比・位相差を分類し、$N > M$ の場合にも BSS を実現している。しかしこの方法では、残響が長い場合や、空間的 aliasing が生じる場合などに、振幅比・位相差が周波数毎に異なるために分類が難しく、性能が低かった。本稿では、音声信号の BSS において、この問題を解決する方法を提案する。

2　スパース性を用いた音源分離の概要

本稿では、時間領域で観測した信号 $\mathrm{x}_j(t)$ $(j = 1, \ldots, M)$ に短時間フーリエ変換 (STFT) を適用し、時間周波数領域にて信号を取り扱う。時間周波数領域における観測信号 $x_j(f,t)$ は近似的に

$$\mathbf{x}(f,t) \approx \sum_{k=1}^{N} \mathbf{h}_k(f) s_k(f,t) \qquad (1)$$

と書ける。ここで $\mathbf{x} = [x_1, \ldots, x_M]^T$、$\mathbf{h}_k = [h_{1k}, \ldots, h_{Mk}]^T$ は信号源 k から各マイク $j = 1, \ldots, M$ への周波数応答、$s_k(f,t)$ $(k = 1, \ldots, N)$ は原信号の STFT 結果である。分離手法は、まず、信号 (図 1(a)) が時間周波数領域でスパースと仮定し、その仮定を用いて時間周波数マスク $M_k(f,t)$(図 1(c)) を推定する (推定法は後述)。次にこのマスクを任意の J 番目の観測信号 $x_J(f,t)$(図 1(b)) に乗じて分離信号 (図 1(d))

$$y_k(f,t) = M_k(f,t) x_J(f,t) \quad (k = 1, \cdots, N)$$

を推定し、最後に逆 STFT にて時間領域の分離信号を得る。

3　音声信号のスパース性

スパース性 (信号がほとんどの時間周波数で 0 に近く、大きな値を持つことは稀) に基づく音源分離では、各時間周波数 (f,t) において原信号のうちの 1 つ s_k のみが支配的であること、すなわち式 (1) が

$$\mathbf{x}(f,t) \approx \mathbf{h}_k(f) s_k(f,t) \qquad (2)$$

と近似されることを仮定する。これは音声信号においても STFT フレーム長を適切に選ぶことで近似的に成り立つことが確認されている (e.g., [1, 3])。

図 2 は、8 秒間の 3 音声の混合信号について、0 個〜3 個の信号がアクティブであるフレーム数の、全体フレーム数に対する割合 (l_ϵ^0-norm[4]) を周波数毎に調べたものである。ここでは各時間周波数における信号 $|s_k(f,t)|$ が閾値 $\epsilon(f) = \frac{1}{10}\max_k \max_t |s_k(f,t)|$ (最大振幅の 1/10) 以上の振幅を持つ時、その信号をアクティブとした。図 2 より、2 個以上の信号が同一時間周波数に存在することは比較的少ないことが分かる。

4　時間周波数マスク推定方法：従来法

マスクを推定するため、スパース性の仮定 (2) を用いて、各時間周波数にて、マイク間の振幅比と位相差

$$\mathbf{\Theta}(f,t) = [\mathbf{\Theta}^L(f,t), \mathbf{\Theta}^P(f,t)]^T \qquad (3)$$

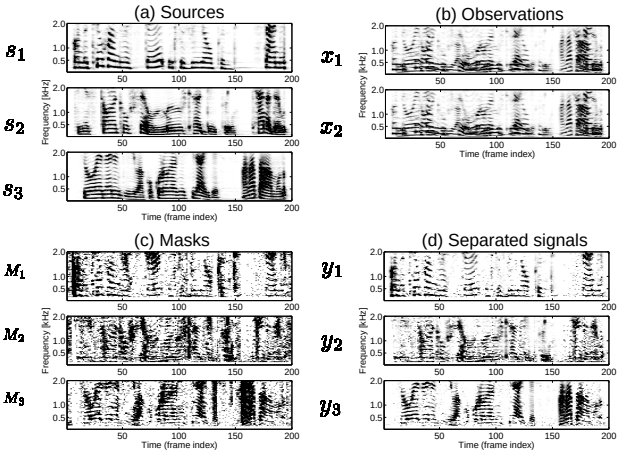

図 1　スペクトログラムの例。(a) 音源音声、 (b) 観測信号、 (c) マスク、 (d) 分離信号 ($N = 3$, $M = 2$)。

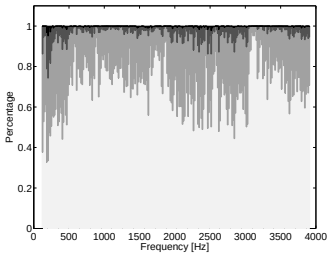

図 2　音源数 $N = 3$(無響) の場合の l_ϵ^0-norm。各色はそれぞれ l_ϵ^0-norm が、薄灰：0、 灰色：1、 濃灰：2、 黒：3。

を計算する [3, 5]。但しここで、

$$\mathbf{\Theta}^L(f,t) = \left[\frac{|x_1(f,t)|}{|x_J(f,t)|}, \cdots, \frac{|x_M(f,t)|}{|x_J(f,t)|}\right]$$

$$\mathbf{\Theta}^P(f,t) = \left[\frac{1}{\alpha_1 f} \arg\left[\frac{x_1(f,t)}{x_J(f,t)}\right], \cdots, \frac{1}{\alpha_M f} \arg\left[\frac{x_M(f,t)}{x_J(f,t)}\right]\right],$$

J は任意のマイク番号、$\alpha_j = 2\pi c^{-1} d_{jJ}$、$d_{jJ}$ はマイク j とマイク J の間隔 (不明な場合は d_{jJ} の最大値より大きな値)、c は音速である。これらマイク間振幅比と位相差は、理想的には、周波数非依存で、音源とマイクの位置によって決まり、音源 k 毎に固有の値を取る。よって、全時間周波数における $\mathbf{\Theta}(f,t)$ を N 個のクラスタ C_k $(k = 1, \cdots, N)$ に分類する (図 3(a)) と、各クラスタ C_k が各音源に対応する。よって、時間周波数マスクは、それぞれのクラスタメンバを抜き出す

$$M_k(f,t) = \begin{cases} 1 & (f,t) \in C_k \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

として推定できる。

図 4 は 2 音源 2 マイクの場合の $\mathbf{\Theta}(f,t)$ のヒストグラムである。2 つのクラスタができており、各クラスタが各音源に対応する。よって図 4 の例のように、残響時間が短く空間的 aliasing の問題も無い場合には、上述の従

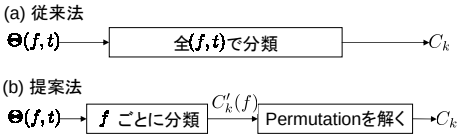

図 3　処理フロー。双方ともマスクは (4) にて得る。

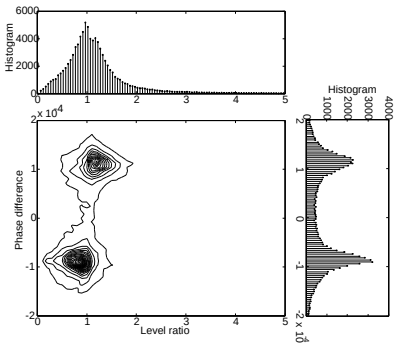

図 4　ヒストグラム例。$N = 2$, $M = 2$, $d_{12} = 4$ cm, 残響時間 $\mathrm{RT}_{60} = 128$ ms, 8kHz サンプリング。

来法にて音源分離が可能であった。

しかし、部屋の残響時間が長い場合には、マイク間振幅比や位相差に周波数依存性が出てくる。また、位相差 $\boldsymbol{\Theta}^P$ の arg 操作が $\pm 2\pi n$(n は整数) の任意性を持つため、$f > \frac{c}{2d_{jJ}}$ の帯域で空間的 aliasing を生じ、マイク間位相差に周波数依存性が出る。これらは、全周波数における分類を困難とし、音源分離性能を下げる要因となる。

5　提案法

そこで、上記のように $\boldsymbol{\Theta}(f,t)$ が周波数依存性を持つ場合にも適用可能な手法を提案する。

提案法では (図 3(b))、まず $\boldsymbol{\Theta}(f,t)$ の分類を周波数毎に行い、各周波数でクラスタ $C'_k(f)$ を得る。次に、全周波数でのクラスタを、同じ音源に起因するクラスタが全て同じ k となるよう再度分類し、最終的なクラスタ C_k を得る。これは、周波数領域 ICA における permutation の問題と酷似している。

ここでは、同じ音声信号では、異なる周波数における時系列が高い相関を持つ (図 1(a) 参照) ことを用いて permutation を解く。これを用いた方法として、各周波数における分離信号の振幅 $|y_k(f,t)|$ の、周波数間相関を最大にする方法が知られている [6]。しかし、図 5 に例示されるように、振幅系列 $|y_k(f,t)|$ は、同じ音声に関する成分であっても、周波数ペアによっては必ずしも高い相関を持たず、permutation を正しく解けない。

そこで提案法では、各クラスタ $C'_k(f)$ を、平均 $\mathbf{a}_k$、分散 σ_k の正規分布 p でモデル化し、各周波数で、$\boldsymbol{\Theta}(f,t)$ がクラスタ $C'_k(f)$ に属する事後確率 (Posterior) の時系列

$$v_k(f,t) = P(C'_k(f)|\boldsymbol{\Theta}(f,t), \mathbf{a}_k(f), \sigma_k(f)) \quad (5)$$

$$= \frac{\alpha_k(f) p(\boldsymbol{\Theta}(f,t)|\mathbf{a}_k(f), \sigma_k(f))}{\sum_k \alpha_k(f) p(\boldsymbol{\Theta}(f,t)|\mathbf{a}_k(f), \sigma_k(f))} \quad (6)$$

($\alpha_k(f)$ は重み係数) を得、この $v_k(f,t)$ の周波数間相関を最大にする [7]。図 6 より、提案している $v_k(f,t)$ は、振幅系列 $|y_k(f,t)|$ より高い周波数間相関を持つことが分かる。これは、音声が存在する (f,t) において、振幅系列 $|y_k(f,t)|$ は広いダイナミクスレンジを持つのに対し、$v_k(f,t)$ はほぼ 1 の値を取ることに起因する。

6　実験と結果

音源数 $N = 4$ マイク数 $M = 3$(間隔 4cm) とし、音源に音声信号を用い実験を行った。混合信号は、6 秒間の英語音声に、可変残響室 (残響時間 130〜450 ms)[7] にて計測したインパルス応答を畳み込んで作成した。

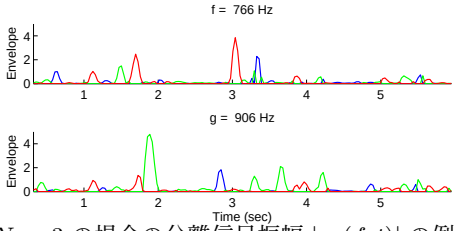

図 5　$N = 3$ の場合の分離信号振幅 $|y_k(f,t)|$ の例。相関係数 ρ は $\rho(|y_1(f,t)|, |y_1(g,t)|) = 0.01$, $\rho(|y_2(f,t)|, |y_2(g,t)|) = 0.10$, $\rho(|y_3(f,t)|, |y_3(g,t)|) = 0.44$。

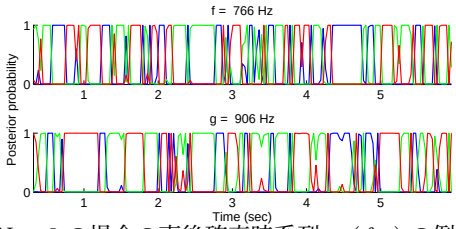

図 6　$N = 3$ の場合の事後確率時系列 $v_k(f,t)$ の例。相関係数 ρ は $\rho(v_1(f,t), v_1(g,t)) = 0.51$, $\rho(v_2(f,t), v_2(g,t)) = 0.46$, $\rho(v_3(f,t), v_3(g,t)) = 0.55$。

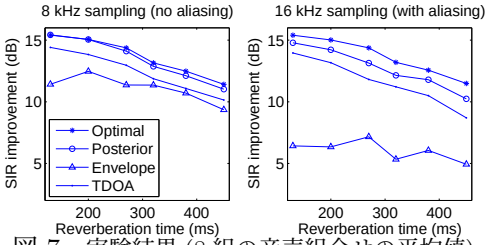

図 7　実験結果 (8 組の音声組合せの平均値)。

STFT フレーム長は 128 ms, フレームシフトは 32 ms であり、音声帯域が 0〜4kHz(8kHz サンプリング、aliasing 無) の場合と 0〜8kHz(16kHz サンプリング、aliasing 有) の場合について実験した。

図 7 に結果を示す。提案法 (**Posterior**, ○) は、従来の $\boldsymbol{\Theta}(f,t)$ を全 (f,t) で分類する方法 (**TDOA**, ·) や、permutation を分離信号の振幅 $|y_k(f,t)|$ で解く方法 (**Envelope**, △) よりも高い性能を示した。尚、図 7 における **Optimal**(●) は、原信号を用いて permutation を解いた場合 (non-blind) の結果である。提案法を用いることで、長い残響や、aliasing が生じる条件でも、**Optimal** に近い性能が得られた。以上より、まず各周波数にてマイク間振幅比/位相差を分類し、次に $v_k(f,t)$ の相関を用いて permutation を解く提案法の有効性が示された。

参考文献

[1] S. Makino, T.-W. Lee, and H. Sawada, Ed., *Blind Speech Separation*, Springer, 2007.

[2] 澤田, 向井, 荒木, 牧野, "多音源に対する周波数領域ブラインド音源分離", 人工知能学会 A I チャレンジ研究会, 2005.

[3] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on SP*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] S. Rickard, "Sparse sources are separated sources," in *Proc. EUSIPCO2006*, Sept. 2006.

[5] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, pp. 1833-1847, 2007.

[6] N. Murata, S. Ikeda, A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[7] H. Sawada, S. Araki, S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," WASPAA 2007, pp. 139-142, 2007.