解説

音源分離技術の最新動向

Recent Advances in Audio Source Separation Techniques

澤田　宏　荒木章子　牧野昭二

Abstract

音源分離技術は，実環境におけるハンズフリー音声認識やコンピュータによる音環境理解のために必要不可欠な技術である．音源の位置や話者の特徴など，事前知識を必要としない，いわゆるブラインド処理に関する技術がこの10年で大きく進展した．本稿では，独立成分分析やスパース性など，ブラインド音源分離に必要な基本技術を分かりやすく解説し，研究動向や現状での到達点を述べる．

キーワード：ブラインド信号分離，独立成分分析，スパース性，畳込み混合，短時間フーリエ変換

1. はじめに

人間は，少々騒がしい中でも目的の音を聞き取ることができる．2人，3人ぐらいならば，同時に話されても，何をいったかを聞き取ることができる．そのような能力をコンピュータで実現すること（図1）を目標として，様々な研究が長年行われてきた．中でも，マイクロホンを複数個用いる方法が効果的である．人間も二つの耳により上記の聞き取り能力が向上する．音源の空間的な位置の違いを利用できるからである．

マイクロホンを複数個用いるものとして，マイクロホンアレーの技術(1)が古くから研究されてきた．複数のマイクロホンでの受音信号に異なる時間遅れを与えて（位相をそろえて）足したり引いたりすることで指向性を形成する技術である．これらの技術では，目的とする音源の方向やその音源が無音である時間区間などの情報が事前知識として必要である．そのため，事前にこれらの情報が得られない，あるいは推定できたとしてもその精度が低い場合には，十分な性能が得られにくい．

1990年代に入り，ブラインド信号分離，特に独立成分分析という技術(2),(3)が研究され始めた．これは，音などの信号がどのように混ざったかという情報を知らなくても，源信号の独立性に基づいて分離ができるという技術である．ある種の不思議さ，理論的深み，適用分野の広さ，などから多くの研究者を魅了し，ある程度成熟した今でも活発に研究されている．

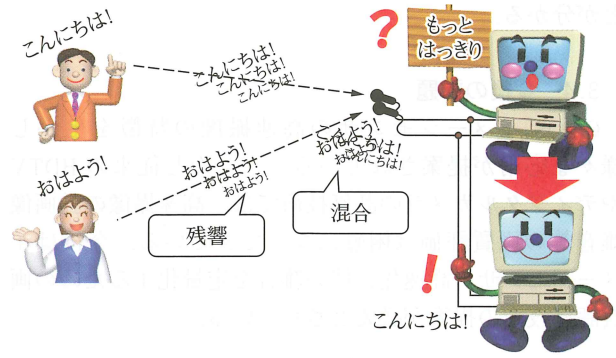


図1　音源分離技術により同時発話を聞き取るコンピュータ　実環境では，単なる混合ではなく，残響も伴う畳込み混合となる．

実環境で混ざり合った音を独立成分分析により分離する研究も，1990年代終りごろから盛んに行われている．そのブラインド性により，従来のマイクロホンアレー技術に必要であった事前知識が不要となる．ただし，実環境での混合に対して，独立成分分析をそのまま直接適用するだけではうまく分離できない．実環境での混合は，図1に示すように，時間遅れや壁などからの反射・残響を伴う，いわゆる畳込み混合となるからである．

畳込み混合への対処には様々なアプローチがあるが，

澤田　宏　荒木章子　正員　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
牧野昭二　正員：フェロー　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
Hiroshi SAWADA, Shoko ARAKI, Members, and Shoji MAKINO, Fellow (NTT Communication Science Laboratories, NIPPON TELEGRAPH AND TELEPHONE CORPORATION, Kyoto-fu, 619-0237 Japan).
電子情報通信学会誌　Vol.91 No.4 pp.292-296 2008年4月

短時間フーリエ変換を施して周波数領域で分離を行う方法が，分離性能及び計算コストの面でよい．本稿では，この周波数領域の方法をできるだけ分かりやすく説明し，最後に現状での到達点と今後の課題を述べる．

2. 畳込み混合と周波数領域での分離

説明のため，多少の定式化を行う．N 個の音源 $s_1, \cdots, s_N$ が空間を通じて混ざり合い，j 番目のマイクロホンで観測された信号 x_j は，畳込み混合として，

$$x_j(t) = \sum_{k=1}^{N} \sum_{l=0}^{P-1} h_{jk}(l)\, s_k(t-l) \qquad (1)$$

と記述される．t は時間である．h_{jk} は，音源 s_k からマイクロホン x_j までのインパルス応答（伝達経路にかかわるもの）であり，サンプリング周波数を 8kHz と低めに設定しても，数千にも及ぶ P 点のサンプルが関与する．

短時間フーリエ変換を施して周波数領域に持ち込むと，畳込み混合が周波数ごとの瞬時混合に近似できる[(4),(5)]．結果として，上記の数千もの係数が絡む複雑な式（1）が以下のように簡略化される．

$$x_j(f, \tau) = \sum_{k=1}^{N} h_{jk}(f)\, s_k(f, \tau) \qquad (2)$$

f は周波数，τ は時間フレーム番号を示す．各周波数 f ごとに見れば，各音源 s_k に係数 h_{jk} が掛けられて足し合わされるという，独立成分分析で一般的に扱われているシンプルな瞬時混合モデルで記述される．これが，周波数領域で分離を行う利点である．

図 2 に，周波数領域で分離を行う音源分離手法の処理の流れを示す．まず，それぞれのマイクロホンでの観測信号に短時間フーリエ変換を施して，周波数ごとの時系列に変換する．図 2 では，奥行き方向が周波数を示す．次に，独立成分分析などを適用して，周波数ごとに分離処理を行う（3.）．しかし，この分離処理は，周波数ごとに別々に行われるため，どの音源の周波数成分が何番目の出力に出てくるかが不定となっている．そのため，後段の処理として，分離信号を適切に並べ替える必要がある（4.）．その処理をした後，分離信号に逆フーリエ変換を施すことで，時間領域での分離信号が得られる．

3. 独立成分分析とスパース性

周波数ごとに観測信号を分離するために，独立成分分析，あるいは，スパース性に基づく分離手法を用いる．本稿では，マイクロホンを複数個用いる方法を考えているが，その数 M 個の要素から成るベクトル $\mathbf{x} = [x_1, \cdots, x_M]^T$ を考えると，式（2）の混合モデルは，

$$\mathbf{x}(f, \tau) = \sum_{k=1}^{N} \mathbf{h}_k(f)\, s_k(f, \tau) \qquad (3)$$

となる．$\mathbf{h}_k$ は，k 番目の音源 s_k からすべてのマイクロホン $\mathbf{x}$ への伝達特性を並べた M 次元ベクトルである．

3.1 独立成分分析

独立成分分析は，M 次元のベクトル $\mathbf{x}$（例えば図 2 の場合はマイクロホンが 3 本なので M=三次元）に，$N \times M$ の行列 $\mathbf{W}$ で表現される線形変換

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

を施す．その結果，ベクトル $\mathbf{y} = [y_1, \cdots, y_N]^T$ の要素が互いに（例えば y_1 と y_2 が）独立になればよい．独立性を評価する基準や方法には様々なものがある[(2),(3)]が，混合する前の源信号の性質をある程度知っていれば，その統計的性質をあらかじめモデル化しておき，分離信号 y_i の性質をそれに近付けていく方法が簡便でよい[(6)]．

音声や多くの楽器の音は，0 に近い振幅の頻度が高いことが知られており，ラプラス分布に類するものでモデル化できる．一方，中心極限定理として知られていることとして，様々な信号を混ぜていくと，信号数が増えるに従って，その分布はガウス分布に近付く．図 3 に，そのことを直観的に示した．幾つかの音が混ざり合った観

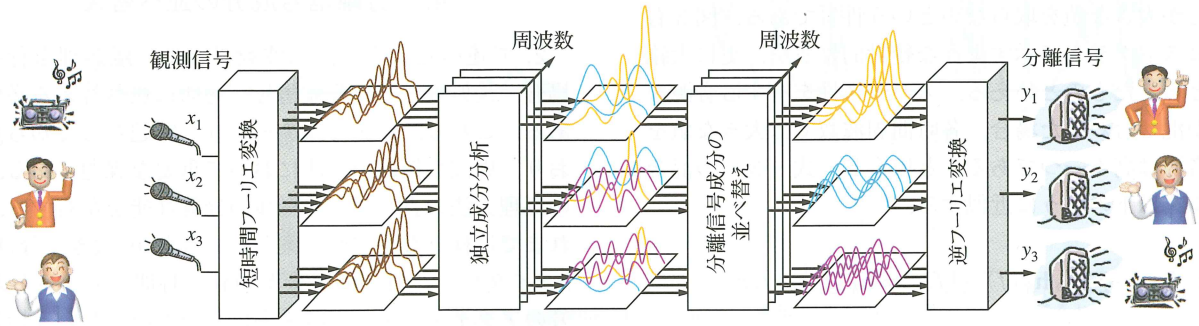


図 2 周波数領域での音源分離　マイクロホンでの観測信号に短時間フーリエ変換を適用して，周波数ごとに分離処理を行う．その後，分離信号成分を適切に並べ替えてから，逆フーリエ変換で時間領域の信号に戻す．

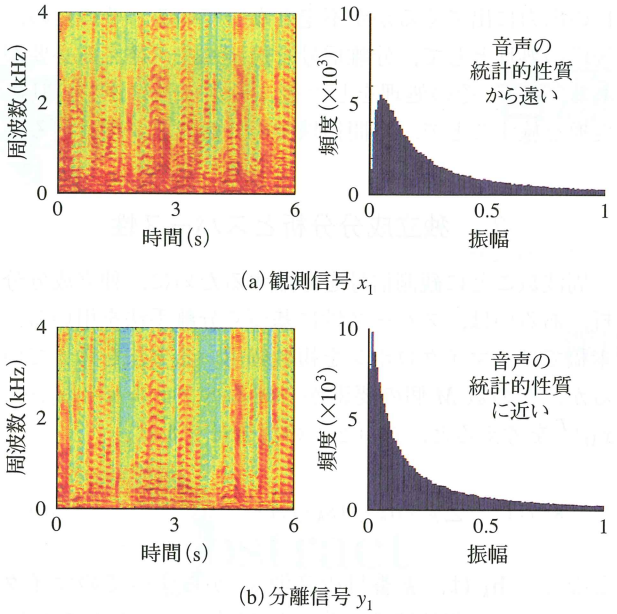


図3　観測信号と分離信号をスペクトログラムで表示したもの（左）と，その振幅のヒストグラム（右）　観測信号は，種々の音が混ざっているため，複素ガウス分布（その振幅は Rayleigh 分布）に近い．分離信号は，独立成分分析により，0に近い振幅の頻度が高くなるような線形変換の結果として得られた．

測信号 x_1 の振幅ヒストグラム（右上）を見ると，0に近い振幅の頻度が低くなっている．音が混ざり合うことにより，振幅が0に近い各音源の時間周波数スロットを互いに埋め合っている．一方，独立成分分析の解としての分離信号 y_1 を見ると，その振幅ヒストグラム（右下）はモデル化された音声の統計的性質に近くなっており，スペクトログラム（左下）では音声の調波構造や無音部分がはっきりと識別できるようになっている．

3.2　スパース性

独立成分分析は線形変換であるため，これを用いた場合，分離できる音の数は，マイクロホンの数と同等かそれ以下に限られてしまう．一方，スパース性に基づく音源分離手法では，そのような制限はなく，マイクロホンより多い数の音に分離することも可能である．

スパース性とは，信号の振幅が多くの場合は0で，たまにしか大きな値を取らないという性質である．図3右下のヒストグラムで示すような統計的性質を，更に大雑把にとらえたものといえる．スパース性を有する音源を幾つか混ぜ合わせた場合，各時間周波数では大きな値を取る音源は高々一つである，と仮定し，式（3）の混合モデルを以下のように近似する．

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

N 個の音が混ざっているはずなのに，各時間周波数スロット (f,τ) で見ると，ある k 番目の音しかマイクロホンで観測されないと近似するわけである．この少々乱暴とも思える近似は，図4や文献（7），（8）などで示されているとおり，意外にも多くの状況で成り立つ．

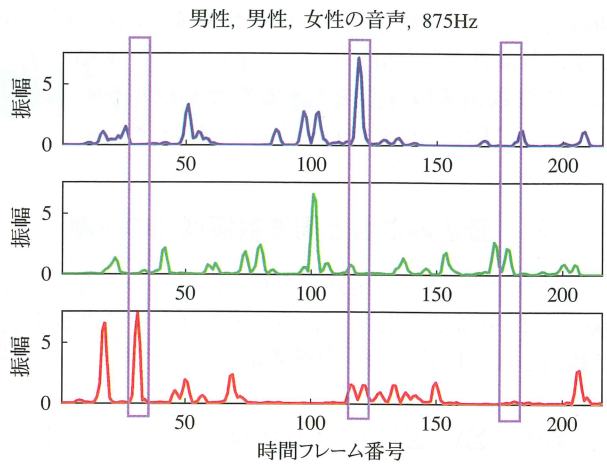


図4　スパース性を持つ信号の例として，同時に発話された三人の音声から，ある周波数成分だけを時系列として示したもの　各時間フレームに着目すると，大きな振幅を持つのは高々一人の音声だけである場合が多い．

スパース性に基づく場合，時間周波数マスキングにより分離の処理が行われることが多い．i 番目の音源を取り出す分離信号 y_i は，その音源が大きな値を取る時間周波数スロット (f,τ) を推定した後，

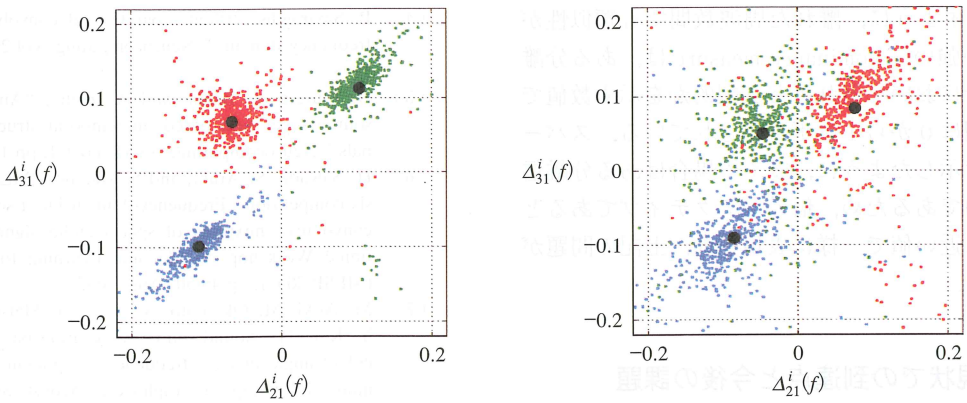
$$y_i(f,\tau)=\begin{cases} x_1(f,\tau) & i\text{番目の音源が}(f,\tau)\text{で大きな値} \\ 0 & \text{その他の場合}\end{cases}$$

として構成される．そのために，各音源が大きな値を取る時間周波数スロット (f,τ) の集合を推定しなければならないが，これは，k-means や EM アルゴリズムを用いて，観測信号のベクトル $\mathbf{x}$ をクラスタリングすることで達成される．振幅と位相を正規化して $s_k(f,\tau)$ の影響を取り除く[9]と，例えば音源が3個の場合は，$\mathbf{h}_1(f)$，$\mathbf{h}_2(f)$，$\mathbf{h}_3(f)$ というベクトルの周辺にそれぞれのクラスタが形成される．

4.　分離信号成分の並べ替え

2.で述べたように，周波数ごとに分離処理を行った場合，後処理として分離信号を適切に並べ替える必要がある．これは，パーミュテーション問題として知られており，周波数領域の手法において重要な課題である．その処理方法の良し悪しが全体の分離性能を左右する．これまでに様々な手法が提案されてきたが，大きく分けて，「マイクロホン対への各音源の到達時間差」と「分離信号のアクティビティを時系列にしたもの」という情報が有用である[10]．

まず，「各音源の到達時間差」は，独立成分分析の結果，



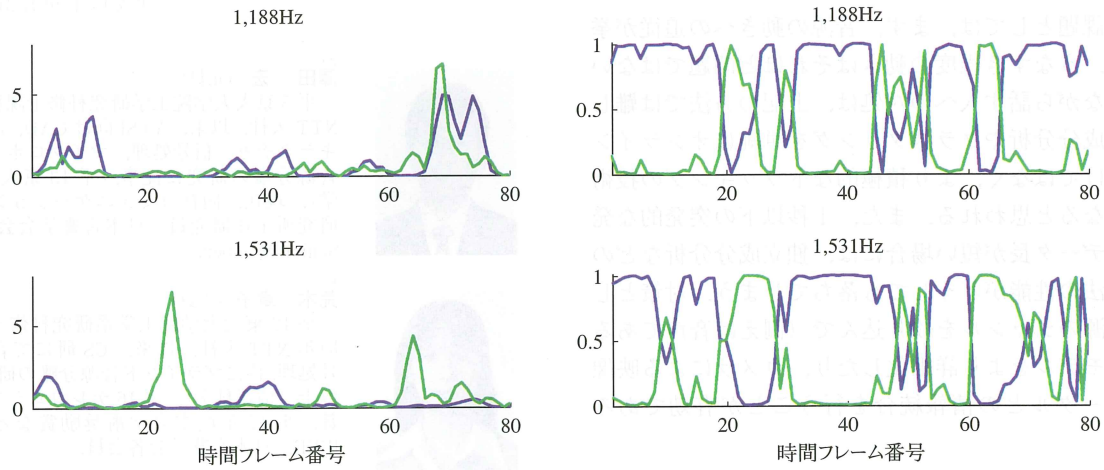
(a) 残響の影響が弱い場合　　(b) 残響の影響が強い場合

図5　各音源からマイクロホンへの到達時間差 Δ^i_{21}, Δ^i_{31} を周波数 f ごとに推定し，クラスタリングを行ったもの　横軸は，マイクロホン2と1の間に生じた到達時間差，縦軸は3と1の間に生じたもの．各クラスタを各音源に対応させることで，パーミュテーション問題を解決できる．残響の影響が弱い場合は，クラスタがはっきりとしておりパーミュテーションの間違いも少ないが，残響の影響が強くなると，到達時間差の推定値が周波数によって大きく異なり，パーミュテーションの間違いも増える．

あるいはスパース性に基づく手法ではクラスタリングの結果から推定することができる．そして，図5に示すように，推定された時間差情報を全周波数にわたってクラスタリングすることで，パーミュテーション問題を解決できる．この方法は，マイクロホンの位置情報と組み合わせて音源方向が推定できるという点で興味深いが，反射や残響の影響が強い場合には，正確な時間差が推定できずに精度が下がる．

一方，「分離信号のアクティビティ」に基づく方法は，音源の性質に基づいており，残響の影響を比較的受けにくい．同じ音源に属する分離信号は，そのアクティビティを時系列として見た場合，周波数が異なっても類似性（相関係数で計算）が高いという性質を利用している．アクティビティを表現するものとして，これまで，分離信号の envelope（包絡線，周波数領域では振幅そのもの）が広く用いられてきた．同じ音源に起因する分離信号の envelope は，近接する周波数間では類似性が高くなるという性質を利用して，パーミュテーション問題を解決する．

しかしながら，図6に示すような離れた周波数間では，envelope の類似性は必ずしも高くならない．信号がアクティブであるという事実が，様々な振幅値で表現されているからである．その問題を解決するものとして，最近筆者らは dominance measure を提案しているが[11]，それ



$$= \begin{bmatrix} 0.10 & -0.14 \\ -0.11 & 0.06 \end{bmatrix}$$

(a) envelope

$$= \begin{bmatrix} 0.54 & -0.54 \\ -0.54 & 0.54 \end{bmatrix}$$

(b) dominance measure

図6　分離信号のアクティビティを envelope と dominance measure で表現したもの　青と緑の線がそれぞれ分離信号 y_1 と y_2 に対応する．異なる周波数間でアクティビティ時系列の類似性が高ければ，同じ音源に属する分離信号であると判定できる．周波数1,188Hz と1,531Hz は比較的離れているため，envelope を用いた場合，同じ音源でも類似性が低くなっている．一方，dominance measure では，十分に類似性が高くなる．

は図6(b)に示すように，離れた周波数間でも類似性が高くなる傾向が強い．dominance measureは，ある分離信号が観測信号においてどれほど支配的かを示す数値であり，その範囲は0から1に正規化されている．スパース性のところで示したように，多くの場合はある分離信号のみが支配的であるため，信号がアクティブであるという事実が1に近い値で一様に表現され，上記の問題が解決されている．

5. 現状での到達点と今後の課題

これまでに述べたような音源分離技術により，音源の数が2～4程度で，3秒以上の観測信号が得られれば，残響の影響がある程度あっても，比較的良好な分離信号が得られるようになってきた．本稿で述べたものとは少し異なる手法として，周波数領域での音源信号を多次元ベクトルでモデル化するもの[(12),(13)]があるが，そのような手法でも同等の性能が得られる．

マイクロホンの数が音源数より少ない場合は，不良設定問題になるため難しいものとされてきたが，本稿で述べたように，スパース性に基づく分離とその後の分離信号成分の並べ替えにより，残響下でも同様に良好な分離が達成される．特に，マイクロホンが2個で済めば，一般に普及しているステレオICレコーダを活用でき，手軽に音源分離技術を適用できる．筆者らのWebページでは，3人の同時発話音声をステレオICレコーダで録音した際の収録音と分離結果を聞くことができる[(14)]．また，音声や楽器音をステレオ録音して混合したデータを集めたものと，様々な研究者による分離結果が文献(15)で公開されている．

今後の課題としては，まず，音源の動きへの追従が挙げられる．うなずき程度の動きはそれほど問題ではないが，歩きながら話す人への対処は，上記の方法では難しい．独立成分分析やクラスタリングを単純にオンライン化するだけではなく，より積極的なトラッキングの技術が必要になると思われる．また，1秒以下の突発的な発話など，データ長が短い場合には，独立成分分析などの統計的手法の性能がどうしても落ちてしまう．対策としては，音源のジャンルを絞り込んで（例えば音声であるとして）モデルをより詳細化したり，カメラによる映像など別モーダルとの情報統合を行うことが有効であろう．

文　献

(1) 大賀寿郎, 山崎芳男, 金田　豊, 音響システムとディジタル処理, 電子情報通信学会, 1995.
(2) A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.
(3) A. Cichocki and S. Amari, Adaptive Blind Signal and Image Processing, John Wiley & Sons, 2002.
(4) P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomputing, vol.22, no.1-3, pp.21-34, 1998.
(5) N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," Neurocomputing, vol.41, no.1-4, pp.1-24, Oct. 2001.
(6) H. Sawada, S. Araki, and S. Makino, "MLSP 2007 data analysis competition: Frequency-domain blind source separation for convolutive mixtures of speech/audio signals," Proc. International Workshop on Machine Learning for Signal Processing (MLSP 2007), pp.45-50, Aug. 2007.
(7) 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. Technol., vol.22, no.2, pp.149-157, 2001.
(8) O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans, Signal Process., vol.52, no.7, pp.1830-1847, July 2004.
(9) S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," Signal Process., vol.87, no.8, pp.1833-1847, 2007.
(10) H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," IEEE Trans. Speech Audio Process., vol.12, no.5, pp.530-538, Sept. 2004.
(11) H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," Proc. ISCAS 2007, pp.3247-3250, May 2007.
(12) A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," Proc. ICA 2006 (LNCS 3889), pp.601-608, March 2006.
(13) T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," IEEE Trans. Audio, Speech and Language Processing, vol.15, no.1, pp.70-79, Jan. 2007.
(14) http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/ubssconv/
(15) Stereo audio source separation evaluation campaign. http://sassec.gforge.inria.fr/

(平成19年10月31日受付)

澤田　宏（正員）
平5京大大学院工学研究科修士課程了．同年NTT入社．以来，VLSI向けCAD，計算機アーキテクチャ，信号処理，マイクロホンアレー，音源分離の研究に従事．平13京大博士(情報学)．現在，同社コミュニケーション科学基礎研究所主任研究員．日本音響学会会員，IEEE Senior Member．

荒木　章子（正員）
平12東大大学院工学系研究科修士課程了．同年NTT入社．以来，CS研にて音声音響信号処理，特にブラインド音源分離の研究に従事．博士（情報科学）．平16テレコムシステム技術賞，平17年度本会学術奨励賞など各受賞．IEEE，日本音響学会各会員．

牧野　昭二（正員：フェロー）
昭56東北大大学院修士課程了．同年日本電信電話公社（現NTT）入社．以来，NTT研究所において，音響エコーキャンセラ，ブラインド音源分離などの音響信号処理の研究に従事．工博．北大客員教授．日本音響学会技術開発賞，本会業績賞，ICA Unsupervised Learning Pioneer Award各受賞．IEEE Fellow．