混じりあった声を解く

—遠隔発話の認識を目指して—

Separation of Speech Signal

— To Realize Multiple Talker Speech Recognition —

牧野　昭二† Shoji MAKINO　向井　良† Ryo MUKAI　荒木　章子† Shoko ARAKI　片桐　滋† Shigeru KATAGIRI

あらまし

たくさんの音の中から聞きたい音を聞き分ける音源分離技術として，近年，独立成分分析（ICA：Independent Component Analysis）に基づく手法が脚光を浴びている．この手法は，音源位置の知識や目的音（妨害音）区間の切り出しを原理的に必要とせず，完全なブラインド分離が可能である．統計的処理であるICAは，物理的，音響的にはある種のブラックボックスであり，その中で何が行われているのか，何がどこまで分離できるのかがあまり分かっていなかった．NTTコミュニケーション科学基礎研究所ではこれまでの研究により，統計的手法であるICAを音響信号処理的な観点から分析して物理的意味付けを与え，従来の音響信号処理技術との関係を解明した．そして，ICAに基づくブラインド音源分離が，適応ビームフォーマと呼ばれるマイクロホンアレイと同じ動作原理を実現しており，2乗誤差最小の意味で等価であることを明らかにした．2マイクの適応ビームフォーマ（ABF）の支配的な動作は妨害音に1つの死角を向ける動作である．これより，様々な方向からの残響音を消せないことがICAが残響に弱い理由の1つであること，十分長いフィルタでの分離がICAでは難しくなること，ABFがICAの性能の上限を与えることなどを明らかにした．

Abstract

The rapid advances in automated speech recognition technology has enabled computers to recognize human speech with a high accuracy, if the speaker speaks politely into a microphone close to the mouth. However, the recognition rate decreases considerably when there are obstructive sounds such as another person's voice, background music, ambient noise, or reverberation. In such cases, computers are unable to recognize what was said. Recently, a statistical method called Independent Component Analysis (ICA) has attracted the attention of researchers as a technique for sound source separation. On the assumption that the sound sources, that is, one's voice, another person's voice, background music, and so on, are mutually independent, this method can restore the original signals if the observed signals are separated to statistically independent signals. This is the principle of source separation using ICA. Our approach and current results are shown in this paper.

1　まえがき

コンピュータによる音声認識技術は年々進歩しており，1人がマイクに向かって丁寧に話した言葉であれば，かなり高い精度で認識できるようになっている．しかしその一方で，目的の人以外の声，背景に流れる音楽，周囲騒音，残響のような邪魔な音があると認識率は急激に低下し，そのような状況ではコンピュータに私たちの話した声を認識させることはできない．私たちが普段それほど意識せずに行っている「聞きたい音を聞き分ける」という能力がコンピュータには欠けているのである．

NTTコミュニケーション科学基礎研究所では，「コミュニケーション環境理解」というテーマを掲げ，人にやさしい対話

† NTTコミュニケーション科学基礎研究所　NTT Communication Science Laboratories

型コンピュータを実現するための研究開発を進めている．この中で我々は，知的対話コンピュータの適用領域を，これまでの接話・単一話者の場合から遠隔発話・複数人話者での対話へ広げるための研究を推進している．

具体的には，会議室などの実環境において，複数話者が同時に発話することが想定される状況下での音源分離技術の実現を目標としている．この音源分離技術は，多様な音が存在する中で音声認識システムへ適切な入力を与えるための重要な要素技術である．

コンピュータによって聞きたい音を聞き分ける，すなわち，聞きたい音源を分離抽出するために，以下に述べる様々な研究が行われてきた．

混合系の逆フィルタを計算して分離をする手法としてMINT法がある[1]．これは混合系のインパルス応答を計測し，その逆フィルタを混合系が非最小位相系の場合にも安定に求める手法である．混合系を正しく計測できれば，原理的に分離音源を得ることが可能である．MINT法は混合系のインパルス応答を計測する必要があるが，近年ブラインドMINTが提案された[2]．これは白色音源を仮定することで，インパルス応答を計測することなく逆フィルタを求めるものである．MINT法は，大きな行列を扱うため，多大な計算コストを要し，時々刻々変化するインパルス応答hへの追随には適さない．

SAFIA法[3]は，2本の指向性マイクロホンを利用し，マイクロホンに近いほうの音源は，マイクロホンから遠いほうの音源よりも大きく早く収音されることを利用して，各周波数で2本のマイクロホン出力のレベルの大きいほうを出力することで分離を行う．比較的ロバストに動作するが，マイクロホンと音源の位置を知る必要がある．

スペクトル領域での分離手法としてはスペクトルサブトラクションがある．これは，ノイズのスペクトルを推定し，混合音から引き去る手法である．ノイズのみが存在する時間の検出，定常ノイズを仮定，など制約が多い手法である．

目的音が音声のように調波構造を持つと仮定し，基本周波数F_0を時々刻々推定して調波構造を抜きだす手法がある．この手法は妨害音の数によらずに分離が可能であるが，子音区間には適用できない，音声どうしの混合音の場合F_0抽出が難しくなる，F_0抽出の性能に分離性能が左右される，という問題点がある．

ノイズキャンセラは，妨害音の重畳する目的音をマイクロホンで受音（x_1）し，さらにreferenceマイクロホンを置いて妨害音をいつも監視（x_2）しながら，x_1とx_2を用いて妨害音を抑圧するフィルタを推定する手法である．referenceマイクロホンに目的音がリークする時間を避けて学習する必要があるほか，そもそもノイズのモニタリングが必要なため適用できる状況が限られる．

適応ビームフォーマ（ABF：Adaptive BeamFormer）は複数のマイクロホンへの音波の到達時間差を利用することで妨害音の方向に死角を向け，目的音のみを収音する技術である．目的音の方向の情報と妨害音のみが鳴っている時間の検出が必要であることが問題点である．

統計的手法として，独立成分分析（ICA：Independent Component Analysis）に基づくブラインド音源分離（BSS：Blind Source Separation）がある[4]~[6]．これは，複数音源が統計的に互いに独立であるという仮定のみを用い，出力が互いに独立となるようなフィルタを求める手法である．この手法は演算量が若干多いが，音源の調波構造などの仮定も，音源とマイクロホンの配置情報なども用いない分離処理が可能である．

これらの手法の中からNTTコミュニケーション科学基礎研究所では，ICAに基づく手法を用いている．この手法は，ある人が話している声と別の人の声，背景に流れる音楽，周囲騒音など，それぞれの音は互いに統計的に独立であるという仮定により，複数のマイクで観測した信号を互いに独立な信号に分離すれば，それぞれの元の音を復元できる，という原理に基づいている．

統計的処理であるICAは，物理的，音響的にはある種のブラックボックスであり，その中で何が行われているのか，何がどこまで分離できるのかがあまり分かっていなかった．我々はこれまでの研究により，統計的手法であるICAを音響信号処理的な観点から分析して物理的意味付けを与え，従来の音響信号処理技術との関係を解明した．

我々は統計的な手法と，音響信号処理的手法との長所をうまく関連付けることで，新しい分離技術を得ることを目標としている．

ここでは，コンピュータによって聞きたい音を聞き分ける，すなわち，聞きたい音源を分離抽出するために，NTTコミュニケーション科学基礎研究所が行っている取り組みについて述べる．

2　ブラインド音源分離

近年盛んに研究が進められている分離手法としてICAに基づくBSSがある[4]~[6]．これは，音源信号どうしが統計的に独立であることを仮定し，観測された混合音声のみを用いて，分離を行う手法である．そのため，音源位置の知識や目的音（妨害音）区間の切り出しを原理的に必要とせず，完全なブラインド分離が可能である．

2.1　混合信号モデル

音源iからの信号s_iに部屋の応答h_{ji}が畳み込まれた信号が混合したものが，マイクロホンjで信号x_jとして観測されるとすると，x_jは，

$$x_j(n)=\sum_{i=1}^{N}\sum_{p=1}^{P}h_{ji}(p)\,s_i(n-p+1)\qquad(j=1,\cdots,M)\qquad(1)$$

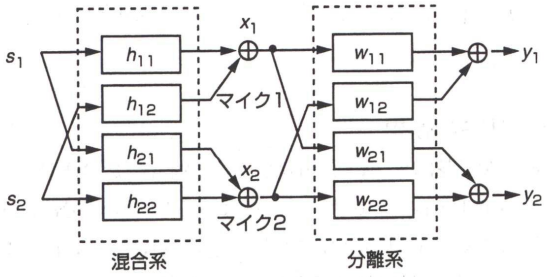


図1　音源分離モデル

で表される．ここでN個の音源は統計的に互いに独立であることを仮定する．h_{ji}は音源iからマイクロホンjへのPタップのインパルス応答であり，線型システムである．本論文では，簡単のため，$N=M=2$とする（**図1**）．

2.2　ブラインド音源分離の原理

ブラインド音源分離では，式(1)の形で得られる信号と，Qタップの分離フィルタ群w_{ij}からなる分離系を用いて分離する．分離して得られる信号y_iは，

$$y_i(n)=\sum_{j=1}^{M}\sum_{q=1}^{Q}w_{ij}(q)\,x_j(n-q+1)\qquad(i=1,\cdots,N)\tag{2}$$

で表される．

式(1)のような畳み込み混合信号（Convolutive Mixture）の時間領域での分離は，演算量が多い，ローカルミニマムに陥りやすい，という難しさがある．この問題は，我々が扱うタップ長Pが大きい場合に特に顕著である．これを回避するため，周波数領域での分離アルゴリズムがよく用いられる（**図2**）．

周波数領域BSSでは，式(1)を，Tポイントの離散フーリエ変換（DFT：Discrete Fourier Transformation）を用いて周波数領域の表現に変換する．

$$\boldsymbol{X}(\omega,m)=\boldsymbol{H}(\omega)\,\boldsymbol{S}(\omega,m)\tag{3}$$

ここで，$\boldsymbol{S}(\omega,m)=[S_1(\omega,m),S_2(\omega,m)]^T$は音源信号である．これにより，式(1)の畳み込み信号の混合を，各周波数での瞬時混合として表現でき，問題を簡単化できる[7],[8]．

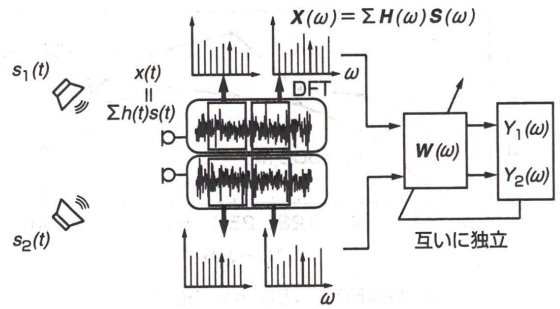


図2　周波数領域BSS

次に，各周波数において出力信号$Y_1(\omega,m)$と$Y_2(\omega,m)$が互いに独立となるよう，分離行列$\boldsymbol{W}(\omega)$を推定する．

$$\boldsymbol{Y}(\omega,m)=\boldsymbol{W}(\omega)\,\boldsymbol{X}(\omega,m)\tag{4}$$

ここで，$\boldsymbol{X}(\omega)=[X_1(\omega),X_2(\omega)]^T$は観測信号，$\boldsymbol{Y}(\omega)=[Y_1(\omega),Y_2(\omega)]^T$は分離信号，$\boldsymbol{W}(\omega)$は，各周波数における分離行列を表す．

分離行列を求めるにはいくつかの手法があるが，ここでは代表的な2つを簡単に説明する．

2.2.1　Higher Order Statistics (HOS)

推定された分離信号を$\hat{\boldsymbol{Y}}=[\hat{Y}_1(\omega,m),\hat{Y}_2(\omega,m)]^T$とする．$\hat{\boldsymbol{Y}}$の同時分布と，$\hat{Y}_i$の周辺分布の積とのKullback-Leibler divergenceは，Y_1とY_2が独立のときのみ0となる．よって，Kullback-Leibler divergenceを最小にする$\boldsymbol{W}$を式(5)により求める[7],[9]．

$$\boldsymbol{W}_{i+1}=\boldsymbol{W}_i+\eta\left[\mathrm{diag}\left(\left\langle\boldsymbol{\Phi}(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\right\rangle\right)-\left\langle\boldsymbol{\Phi}(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\right\rangle\right]\boldsymbol{W}_i\tag{5}$$

ここで$\langle\cdot\rangle$は期待値演算，iは更新回数，ηはステップサイズ，また$\boldsymbol{\Phi}(\boldsymbol{Y})$は以下の非線型関数である．

$$\boldsymbol{\Phi}(\boldsymbol{Y})=\frac{1}{1+\exp(-\boldsymbol{Y}^{(\mathrm{R})})}+j\frac{1}{1+\exp(-\boldsymbol{Y}^{(\mathrm{I})})}\tag{6}$$

2.2.2　Second Order Statistics (SOS)

BSSの問題は無相関化だけでは解けない，ということはよく知られている．しかし，音声のような非定常信号の場合には，定常を仮定できる異なったブロックの無相関化を同時に行うことで，分離が可能になる[10]~[12]．

非定常な音源信号$S_1(\omega,m)$と$S_2(\omega,m)$が，平均0，無相関であると仮定する．すなわち，

$$\begin{aligned}\boldsymbol{R}_S(\omega,k)&=\frac{1}{M}\sum_{m=0}^{M-1}\boldsymbol{S}(\omega,Mk+m)\,\boldsymbol{S}^*(\omega,Mk+m)\\&=\Lambda_s(\omega,k)\end{aligned}\tag{7}$$

ここで*は共役転置を表し，$\Lambda_s(\omega,k)$は，kによって異なる対角行列である．

分離行列$\boldsymbol{W}(\omega)$は，出力の共分散行列$\boldsymbol{R}_Y(\omega,k)$の非対角要素がすべての$k$で同時に0となるよう決定される．

$$\begin{aligned}\boldsymbol{R}_Y(\omega,k)&=\boldsymbol{W}(\omega)\,\boldsymbol{R}_X(\omega,k)\,\boldsymbol{W}^*(\omega)\\&\rightarrow\Lambda_c(\omega,k)\end{aligned}\tag{8}$$

ここで$\boldsymbol{R}_X$は$\boldsymbol{X}(\omega)$の共分散行列，$\Lambda_c(\omega,k)$は任意の対角行列である．

3 周波数領域BSSの性能評価実験

今，周波数領域BSSにおいて，混合系$\boldsymbol{H}$のインパルス応答長をPとし，分析窓長TとDFTサイズT，および分離系のフィルタ長Qは等しいとする．このとき分離系のフィルタ長について，例えば，あるシステムH（Pタップ）の同定を行う場合は，$Q \geq P$のFIR（Finite Impulse Response）フィルタ$\boldsymbol{W}$を用いるべきである[13]．また，linear convolutionである式(1)を正しく周波数領域（式(3)）の表現に変換するには，$2P < T$が必要である．さらに，あるシステムH（Pタップ）の逆フィルタを求める場合には，$P \ll Q$タップのFIRフィルタ$\boldsymbol{W}$を用いるべきである．

これらを基に考えると，周波数領域処理を用いて分離系$\boldsymbol{W}$を求める場合，部屋のインパルス応答長Pより長いフレームTを用いて分析を行う必要があると考えられ，これまで長いフレームを用いた分離が広く行われてきた[11],[12]．

しかし予備実験において，長いフレームを用いた際，良い分離性能が得られないことが分かった．そこで，ICAに基づくBSSにおけるフレーム長と分離性能の関係を実験で確かめた[14]．

3.1 実験方法

使用した信号は，**図3**に示す環境で実測したインパルス応答を音声信号に計算機上で畳み込んだものである．部屋のインパルス応答長T_Rは，0 ms, 150 ms (P=1200), 300 ms (P=2400)の3種類である．音声信号は，男女それぞれ2名ずつが発話したASJ研究用連続音声コーパスの中の2文であり，長さは約8秒である．この冒頭3秒を用いて2.2.1項の式(5)における学習を行い，8秒全体の信号を分離した．

実験において，DFTのフレーム長Tを32から2048に変化させた．サンプリング周波数は8 kHz，フレームシフトは$T/2$，分析窓はHamming窓である．周波数領域BSSにおいて問題となるpermutationの問題については，栗田らの手法で解決した[9]．

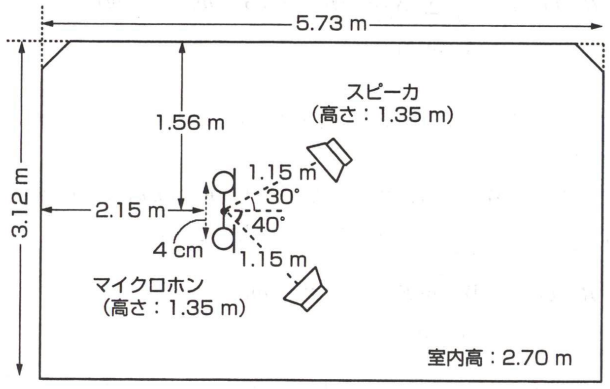


図3 室内レイアウト

3.2 実験結果

実験結果を**図4**に示す．分離性能の評価には，NRR（Noise Reduction Rate）を用いた．これは，出力信号のSN比から入力信号のSN比を引いたものとして定義される．図4では便宜上，y_1, y_2それぞれに対するNRRの平均値を示した．

図4(b)より，残響時間T_R＝150 msのときはDFTフレーム長T＝256において，T_R＝300 msのときはT＝512において，最も良い分離性能が得られていることが分かる．

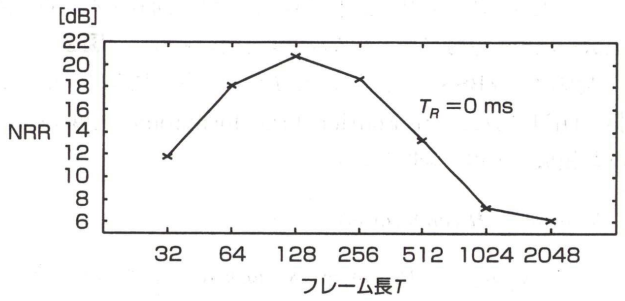
以上により，部屋のインパルス応答長が長い場合においても，より短いフレーム長での分析において最も良い分離性能が得られた．

3.3 実験に対する考察

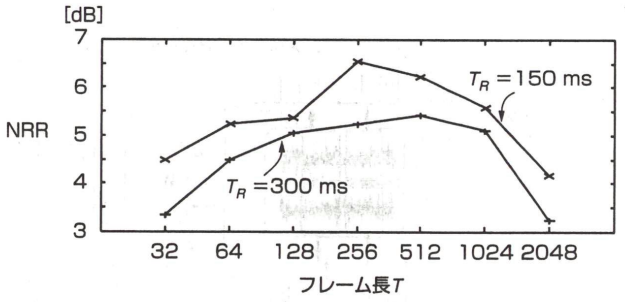
まず特記すべきは，残響環境下での分離性能が6 dB程度と非常に低いことである．残響下での分離性能を上げていくことが大きな課題であることが分かる．

次に，本実験の目的であるフレーム長と分離性能の関係であるが，インパルス応答長より短いフレームを用いた場合のほうがインパルス応答長程度の長さのフレームを用いた場合より高い性能が得られた．

フレーム長が長い場合，今回はフレームシフトを$T/2$に固定



(a) 残響時間＝0 ms



(b) 残響時間＝150 ms, 300 ms

図4 分離性能の評価結果

しているので，各周波数において学習に用いることのできるデータ数が少なくなり，ゼロ平均，独立性などICAによる分離の際にデータに要求される仮定が満たされないことが，性能が低い原因の1つと考えられる．また，各周波数での誤差がフレーム長にわたって広がるためプリエコーが発生し，分離性能が下がるだけでなく聴感上も劣化が大きくなる．

短いフレームでは長い残響に対処するだけのタップ長を用意できていないことが性能の低い原因である．

よって，①各周波数でのデータの性質，②残響への対処，のトレードオフで，最も性能の高いフレーム長が決まることが分かった．

我々はさらに，長い時間の学習データを用いたり，オーバサンプリングを行ったりすることで，より長いフレーム長でより良い性能が得られることを確認している[15]．よって，長い分離フィルタを効率良く精度良く求めることで，長い残響下でも分離性能を上げることができると考えられる．

4 BSSとABFのフレームワークの統一的理解

筆者らは，BSSのフレームワークを音響システムとして理解するために，ABFのフレームワークとの比較を行い，2乗誤差最小の意味で両者が等価であることを示した[16]．BSSとABFの等価性を明示的に示したのは，これが初めてである．以下，ICAによるBSSとABFの等価性を示すが，同様の考察で，referenceマイクロホンに信号漏れのある場合のノイズキャンセラのフレームワークとの等価性も示すことができる[10],[17]．

4.1 ABF

ABFは，複数のマイクロホンへの音波の到達時間差を利用することで，妨害音の方向に指向特性の死角を向け，目的音のみを収音する技術である．目的音の方向の情報と妨害音のみが鳴っている時間の検出が必要である．

まず，周波数領域でのABFについて考える．

ABFでは，目的音方向既知を仮定する．目的音S_1がない時間に，妨害音S_2の出力を最小にするようWを推定する(**図5**(a))．

$$Y_1(\omega, m) = \boldsymbol{W}(\omega)\,\boldsymbol{X}(\omega, m) \tag{9}$$

ここで$\boldsymbol{W}(\omega) = [W_{11}(\omega), W_{12}(\omega)]$, $\boldsymbol{X}(\omega, m) = [X_1(\omega, m), X_2(\omega, m)]^T$である．2乗誤差最小の規範によりエラー関数を次のように定義する．

$$\begin{aligned} J(\omega) &= E[Y_1^2(\omega, m)] \\ &= \boldsymbol{W}(\omega)\,E[\boldsymbol{X}(\omega, m)\,\boldsymbol{X}^*(\omega, m)]\,\boldsymbol{W}^*(\omega) \\ &= \boldsymbol{W}(\omega)\,\boldsymbol{R}(\omega)\,\boldsymbol{W}^*(\omega) \end{aligned} \tag{10}$$

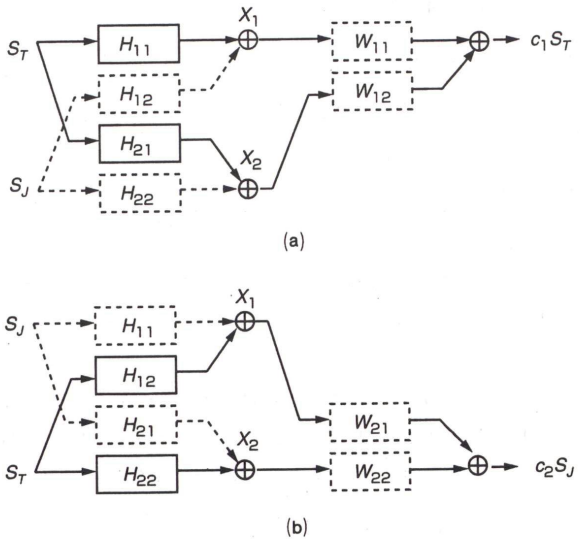


図5　2組の適応ビームフォーマ（ABF）

ここでEは期待値を表し，

$$\boldsymbol{R}(\omega) = E\begin{bmatrix} X_1(\omega,m)\,X_1^*(\omega,m) & X_1(\omega,m)\,X_2^*(\omega,m) \\ X_2(\omega,m)\,X_1^*(\omega,m) & X_2(\omega,m)\,X_2^*(\omega,m) \end{bmatrix} \tag{11}$$

である．$J(\omega)$を目的音方向へのゲインが0にならないという拘束条件のもとで最小化する問題を考えると，

$$W_{11}H_{12} + W_{12}H_{22} = 0 \tag{12}$$

$$W_{11}H_{11} + W_{12}H_{21} = c_1 \tag{13}$$

の連立方程式が得られ，この解の$\boldsymbol{W}$を用いて1組のABFが得られる．

同様に目的音がS_2，妨害音がS_1の場合のABF（図5(b)）も考え，2組のABFを，フィルタの係数を用いてまとめて表すと，各周波数で

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix} \tag{14}$$

の関係を得る．

4.2 BSS

ここでは簡単のため，Second OrderのBSSを考える．非定常信号に対してSOSに基づくBSSのフレームワークでは，システムの出力$\boldsymbol{Y}$の共分散行列

$$E\begin{bmatrix} Y_1Y_1^* & Y_1Y_2^* \\ Y_2Y_1^* & Y_2Y_2^* \end{bmatrix} \tag{15}$$

を，すべてのブロックにおいて対角化する$\boldsymbol{W}$が分離行列となる．

ここで，$\boldsymbol{H}$と$\boldsymbol{W}$のカスケード系を

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \tag{16}$$

と置き，式(15)の非対角要素の2乗を書き下すと，

$$\begin{aligned} &(E[Y_1\ Y_2^*])^2 \\ &= \{ad^*E[S_1 S_2^*] + bc^*E[S_2 S_1^*] + (ac^*E[S_1^2] + bd^*E[S_2^2])\}^2 \\ &= 0 \end{aligned} \tag{17}$$

となる．式(17)において，S_1とS_2は無相関を仮定しているので，第1項と第2項は0である．よって，SOSに基づくBSSは式(17)の第3項を0にするように進む．すなわち，

$$ac^* = bd^* = 0 \tag{18}$$

であり，次の2とおりの解が得られる．

CASE 1：$a = c_1,\ c = 0,\ b = 0,\ d = c_2$

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix} \tag{19}$$

これはABFの式(14)と同じである．

CASE 2：$a = 0,\ c = c_1,\ b = c_2,\ d = 0$

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} 0 & c_2 \\ c_1 & 0 \end{bmatrix} \tag{20}$$

この式はpermutation解（$Y_1 = c_2 S_2,\ Y_2 = c_1 S_1$と推定される）に相当する．

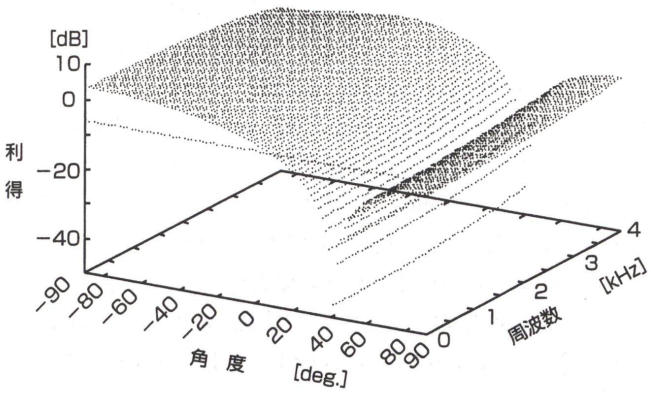
よって，BSSのエラー関数（非対角要素）の最小化はABFの2乗誤差最小化と同じであり，ABFとの等価性が示された．

すなわち，BSSでは妨害音，目的音ともに存在する時間に適応できるABFを形成するということが分かる．またさらに，信号間の無相関の仮定が成り立たない場合は，式(17)の第1，2項が値を持ち，バイアスノイズとなることも分かる．

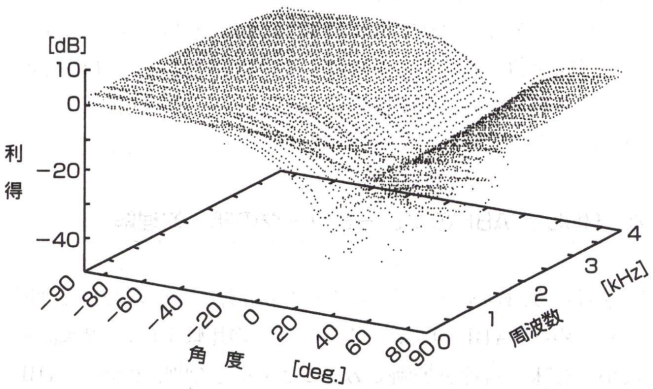
4.3 考　　察

2マイクのABFの支配的な動作は妨害音に1つの指向特性の死角を向ける動作である．BSSにより得られたWによる指向特性を**図6**に示す．これより，様々な方向からの残響音を消せないことがBSSが残響に弱い理由の1つであると考える．

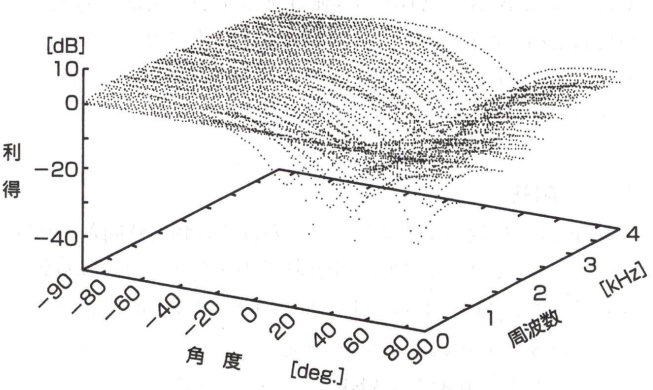
また3章より，十分長いフィルタでの分離がBSSでは難しくなるので，ABFがBSSの性能の上限を与えるものと考えられる．よってBSSによる分離性能をABFに近づけることは意義があると考える．



(a) 死角型ビームフォーマによる指向特性



(b) BSSによる指向特性（無響）



(c) BSSによる指向特性（残響時間150 ms）

図6　BSSとABFの指向特性

5 分離抽出された音声の認識に向けて

様々な音信号や画像信号が混在するコミュニケーションシーンにおいて，「聞きたい音を聞き分ける」あるいは「聞きたい人の声を理解する」ためには，本論文で述べてきた音源分離の技術に加えて，分離抽出された音信号を正確に認識する技術が必要となる．

こうした音信号の認識技術は，特に音声信号を対象とする音声認識技術としてこれまでにも盛んに研究されてきている．認識システムの構造として，隠れマルコフモデルと呼ばれる確率モデルが広く用いられ，計算機の性能向上ともあいまって，語彙数や発話態度などに制約はあるものの，様々な音声認識課題において利用され得る状況に至っている．しかし，大語彙課題，あるいは語彙を特定しないオープン語彙課題における認識性能はまだ不十分である．さらに，人どうしの対話のような，自由な連続会話音声を高精度に認識するにも至っていない．

コミュニケーションシーン理解における認識技術の側面に関しても，我々は様々な研究開発の取り組みを行っている．

まず，認識性能を効率的に向上させる設計手法の実現を目指し，一般化確率的降下法と呼ばれる最近の設計手法の改良を行っている(18)．ここでは，特にオープン語彙の認識などでその必要性が大きい，認識結果の信頼度測度に関する研究に焦点を置く．

人間と機械との対話モードのコミュニケーションにおいて，処理の実時間性は重要である．しかし，対象語彙が大きくなるにつれ，認識候補単語などを探索する処理は膨大となり，認識処理の高速化・実時間化は困難となる．こうした問題を克服することを目指し，認識処理で用いる大規模な言語モデルネットワークの必要部分のみをオンデマンドモードでオンメモリとする手法の研究を行っている(19)．

聞き分けられない外国語を話すのは難しい．話せない外国語を聞き分けることも難しい．人間の音声知覚能力と発話能力との間には密接な関係があることが指摘されている．しかしながら，例えば隠れマルコフモデルのような現今の音声認識システムは，必ずしも，こうしたモダリティ間の相互作用あるいはその関連情報を十分には利用していない．発話機構の情報を活用することによって，音声認識性能を向上させることが可能になる可能性がある．我々は，こうした観点に立ち，音声発話に伴う拘束条件を音声認識用の特徴パタン表現に取り入れることよって，認識性能を向上させる試みを行っている(20)．この手法は，もともと音声認識用のシステムであった隠れマルコフモデルによって音声合成をも行う，認識と合成との2つの技術を1つのモデル枠組で統合的に実現するものであり，聞き真似学習をしながら成長するTalking Babyの動作原理をも担うものとも考えている．

6 あとがき

コンピュータによって聞きたい音を聞き分ける，すなわち，聞きたい音源を分離抽出するための，NTTコミュニケーション科学基礎研究所が行っている，独立成分分析に基づく取り組みについて述べた．

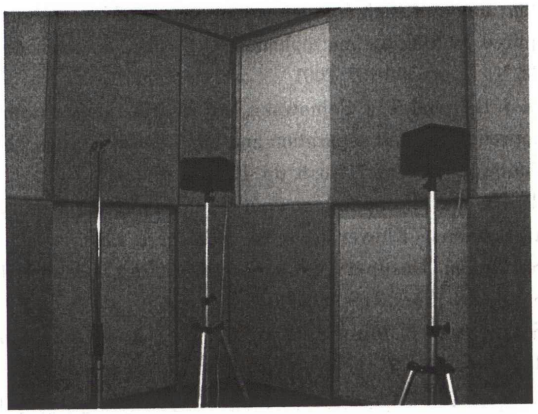


図7　2つのスピーカを用いた実環境での分離実験

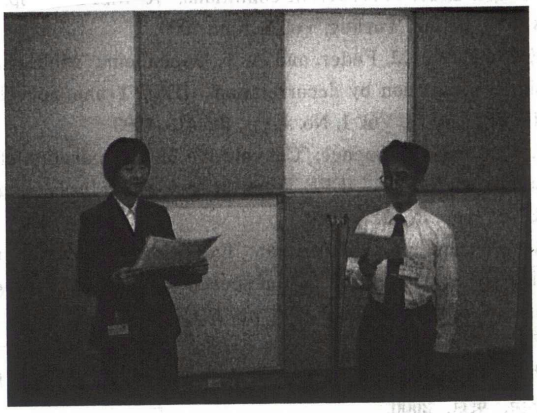


図8　2人の話者の声の実環境での分離実験

我々は，会議室内に音源分離実験システムを構築し，実環境においてどの程度の性能が得られるかの実験を行った（図7，8）．その結果，スピーカを用いた実験はもとより，実際の部屋で2人の話者が同時に発声した声をその場で分離することにも成功した．

今後は，複数人の声が聞き分けられるように改良を進め，これまで接話・単一話者に限定されていた音声認識システムの適用領域を，遠隔発話・複数人話者での対話へ広げ，会議の議事録を自動編集するシステムやロボットの耳などへの応用を目指す．

文　献

(1) M. Miyoshi and Y. Kaneda: "Inverse filtering of room acoustics," IEEE Trans. ASSP, Vol. 36, No. 2, pp. 145-152, 1988.
(2) K. Furuya and Y. Kaneda: "Two-channel blind deconvolution for non-minimum phase impulse responses," ICASSP'97, Vol. 2, pp. 1315-1318, 1997.
(3) M. Aoki, M. Okamoto, S. Aoki, H. Matsui, I. Sakurai, and Y. Kaneda: "Sound source segregation based on estimating inci-

dent angle of each frequency component of input signals acquired by multiple microphones," J. Acoust. Soc. Jpn. (E), Vol. 22, No. 2, pp. 149-157, 2001.
(4) 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.
(5) T. W. Lee: "Independent component analysis -Theory and applications," Kluwer academic publishers, 1998.
(6) S. Haykin: "Unsupervised adaptive filtering," John Wiley & Sons, New York, U.S.A., 2000.
(7) S. Ikeda and N. Murata: "A method of ICA in time-frequency domain," ICA'99, pp. 365-370, Aussois, France, Jan. 1999.
(8) P. Smaragdis: "Blind separation of convolved mixtures in the frequency domain," Neurocomputing, Vol. 22, pp. 21-34, 1998.
(9) S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura: "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," ICASSP2000, pp. 3140-3143, Istanbul, Turkey, Vol. 5, June 2000.
(10) E. Weinstein, M. Feder, and A. V. Oppenheim: "Multi-channel signal separation by decorrelation," IEEE Trans. Speech Audio Processing, Vol. 1, No. 4, pp. 405-413, 1993.
(11) L. Parra and C. Spence: "Convolutive blind separation of non-stationary sources," IEEE Trans. Speech Audio Processing, Vol. 8, No. 3, pp. 320-327, 2000.
(12) M. Z. Ikram and D. R. Morgan: "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," ICASSP2000, pp. 1041-1044, Istanbul, Turkey, Vol. 2, June 2000.
(13) 大賀・山崎・金田："音響システムとディジタル処理," 電子情報通信学会, 東京, 2000.
(14) S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari: "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," ICASSP2001, Salt Lake City, Utah, U.S.A., Vol. 5, pp. 2737-2740, May 2001.
(15) R. Mukai, S. Araki, and S. Makino: "Separation and dereverberation performance of frequency domain blind source separation for speech in a reverberant environment," Eurospeech2001, Älborg, Denmark, Vol. 4, pp. 2599-2602, Sept. 2001.
(16) S. Araki, S. Makino, R. Mukai, and H. Saruwatari: "Equivalence between frequency domain blind source separation and frequency domain adaptive null beamformers," Eurospeech2001, Älborg, Denmark, Vol. 4, pp. 2595-2598, Sept. 2001.
(17) S. V. Gerven and D. V. Compernolle: "Signal separation by symmetoric adaptive decorrelation: stability, convergence, and uniqueness," IEEE Trans. Speech Audio Processing, Vol. 43, No. 7, pp. 1602-1612, 1995.
(18) E. McDermott, A. Biem, S. Tenpaku, and S. Katagiri: "Discriminative Training for Large Vocabulary Telephone-Based Name Recognition," ICASSP2000, Istanbul, Turkey, Vol. 6, pp. 3739-3742, June 2000.
(19) D. Willett, E. McDermott, Y. Minami, and S. Katagiri: "Time and Memory Efficient Viterbi Decoding for LVCSR Using a Precompiled Search Network," EUROSPEECH 2001, Älborg, Denmark, Vol. 2, pp. 847-850, Sept. 2001.
(20) Y. Minami, E. McDermott, A. Nakamura, and S. Katagiri: "A Recognition Method Using Synthesis-Based Scoring That Incorporates Direct Relations Between Static and Dynamic Feature Vector Time Series," CRAC 2001, Älborg, Denmark, Sept. 2001.

著者紹介

牧野　昭二
NTT コミュニケーション科学基礎研究所主幹研究員（グループリーダ）
昭和56年入社．主に，電気音響変換器，拡声電話機，音響エコーキャンセラ，および，ブラインド音源分離などの音響信号処理の研究に従事．
昭和54年東北大学工学部機械工学科卒業．56年同大学院機械工学専攻修士課程修了．平成5年工学博士（同大学）．
IEEE Senior Member・IEEE Associate Editor・日本音響学会・電子情報通信学会会員．
平成7年日本電信電話株式会社社長表彰受賞，同年日本音響学会技術開発賞受賞．9年電子情報通信学会第34回業績賞受賞．

向井　良
NTT コミュニケーション科学基礎研究所主任研究員
平成4年入社．主に，交換ノード用プロセッサ，分散処理システムの研究開発に従事．現在，音響信号処理，音源分離の研究に従事．
平成2年東京大学理学部情報科学科卒業．4年同大学院理学系研究科修士課程修了．
日本音響学会・電子情報通信学会・情報処理学会・ACM・IEEE 会員．

荒木　章子
NTT コミュニケーション科学基礎研究所社員
平成12年入社．主に，ブラインド音源分離の研究に従事．
平成10年東京大学工学部計数工学科卒業．12年同大学院工学系研究科計数工学専攻修士課程修了．
日本音響学会・IEEE 会員．

片桐　滋
NTT コミュニケーション科学基礎研究所知能情報研究部長
昭和57年入社．主に，音声認識の研究に従事．現在，学習理論や信号処理，テキスト処理等の知能情報学の研究に従事．
昭和52年東北大学工学部電気工学科卒業．54年同大学院工学研究科情報工学専攻修士課程修了．57年工学博士（同大学）．
IEEE・日本音響学会・電子情報通信学会・米国音響学会・人工知能学会会員．
昭和57年・62年日本音響学会論文賞受賞．平成6年IEEE Signal Processing Society Senior Award 受賞．13年IEEE Fellow.