実世界情報システムプロジェクト
～ネオサイバネティックス研究グループ～
マイクロフォンアレイ計測

眞溪　歩，高橋　稔
新領域創成科学研究科複雑理工学専攻

概要

複数話者が同時発話する環境において，混合音声を分離した後，特定話者の発話を音声認識することを考える．混合音声の分離は，カクテルパーティ問題としてよく知られている．一方，実環境における音声認識では，音源信号の残響による劣化が深刻な問題となる．

現在，マイクロフォンアレイ計測では，音源分離と残響除去が個別の問題として扱われている．ここでは，音源数に比してマイクロフォン個数が過多な場合について，音源分離と残響除去を同時に行う手法を提案する．

1 はじめに

複数話者が同時発話する環境において，混合音声を分離した後，特定話者の発話を音声認識することを考える．たとえば，会議での発言のテープ起こしを機械で行う場合，この問題設定が当てはまる．機械による音声認識では雑音と認識対象音声の残響が問題となる．音声認識率は，雑音が機械から生じる環境音である場合より認識対象以外の音声である場合，より著しく劣化する．また，音声認識率は，認識対象音声が無雑音であっても残響を含む場合，著しく劣化する．

一方，マイクロフォンアレイ計測では，音源分離と残響除去が個別の問題として扱われてきた．前者はカクテルパーティ問題として，後者は特別な音響環境でなくても音声を原音に近い状態で収録する問題として議論されている．どちらも，広義には，特定音源に対して高指向性を実現するビームフォーマと呼ぶことができる．

本研究では，音源数に比してマイクロフォン個数が過多な場合について，音源分離と残響除去を同時に行う手法を提案する．

2 手法

本提案手法では，2 音源の分離・残響除去を 16 個のマイクロフォンアレイを用いて行う．なお，各音源から各マイクロフォンへのインパルス応答は既知としている．図 1 に本提案手法の基本構成を示す．まず，8 個ずつ 2 組に分けたマイクロフォンアレイ収録音声信号に対し，遅延和ビームフォーマを用いた第 1 段階の残響除去を行い，4 組の混合音声を作成する．つぎに，混合行列の余因子行列を用いた音源分離を行い，2 組の残響状態は異なるが分離された音声を作成する．ここまではアレイごとに処理を行う．最後に，残響は異なるが分離された 2 個の音声に対し，アレイ間にまたがる MISO(Multiple Input Single Output)ウィナーフィルタを用いた第 2 段階の残響除去を行い，分離かつ残響除去音声を作成する．

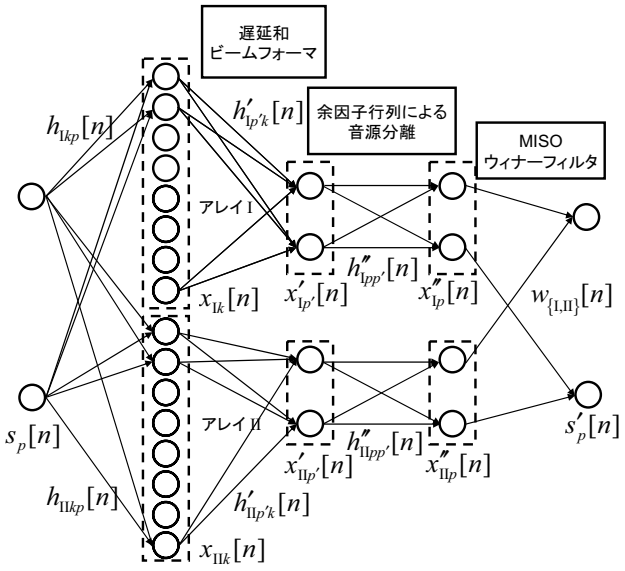

図 1 提案手法の基本構成

2.1 混合モデル

音源 $s_p[n]$, p=1,2 を，アレイ I, II のマイクロフォンによって測定した信号をそれぞれ $x_{\mathrm{I}k}[n]$, $x_{\mathrm{II}k}[n]$, k=1,…,8 とする．この間のインパルス応答を $h_{\{\mathrm{I,II}\}kp}[n]$とおくと，混合モデルは

$$x_{\{\mathrm{I,II}\}k}[n]=\sum_{p=1}^{2}h_{\{\mathrm{I,II}\}kp}[n]\otimes s_p[n] \tag{1}$$

と書ける．ここで，{·,·}は括弧内のいずれかの同順選択を，⊗はたたみ込みを表す演算子である．なお，個々のインパルス応答は室内環境による残響を含んでいる．

2.2 遅延和ビームフォーマ

遅延和ビームフォーマは，信号源センサ間の伝播遅延を，人工的な遅延手段によって同じにして加算する単純な信号強調手法である．このため，狭帯域信号に対しては効果があるが，音声のような広帯域信号に対しては効果を見込めない．これは，センサの設置間隔の持つ意味が各波長によって異なるために，各周波数一律の遅延では不都合な周波数帯域が生じるためである．このため，一般には，周波数帯域を分けて遅延和がとられる．一方，広帯域の信号であれば，その自己相関関数は比較的急に減衰すると考えられる．そこで，本研究では，各周波数一律の遅延和ビームフォーマを簡単な残響除去前処理として利用する．具体的には，

$$x'_{\{\mathrm{I,II}\}p'}[n]\Big|_{p'=p}=\sum_{k=1}^{8}\frac{x_{\{\mathrm{I,II}\}k}\left[n-n_{kp}\right]}{h_{\{\mathrm{I,II}\}kp}\left[n_{kp}\right]} \tag{2}$$

なる操作を行う．ここで，

$$n_{kp}=\arg\max_{n}\left\{\left|h_{\{\mathrm{I,II}\}kp}[n]\right|\right\} \tag{3}$$

である．式(2)を別の見方をすると，

$$x'_{\{\mathrm{I,II}\}p'}[n]\Big|_{p'=p}=\left(\sum_{k=1}^{8}h'_{\{\mathrm{I,II}\}p'k}[n]\right)\Bigg|_{p'=p}\otimes s_p[n] \tag{4}$$

なる測定を行っていることになる．ここで，

$$h'_{\{\mathrm{I,II}\}p'k}[n]\Big|_{p'=p}=\frac{h_{\{\mathrm{I,II}\}kp}\left[n-n_{kp}\right]}{h_{\{\mathrm{I,II}\}kp}\left[n_{kp}\right]} \tag{5}$$

である．

2.3 余因子行列による音源分離

式(2)は，システム設計の意味において p=p'であるが，$x'_{\{\mathrm{I,II}\}p'}[n]$は音源 $s_p[n]$の推定にはなっていない．式(2)は，p'=p なる対応のもと，周波数によらず一律の遅延 n_{kp} によってビームフォーミングを行っているに過ぎない．このため，$x'_{\{\mathrm{I,II}\}p'}[n]$には音源$\{s_p[n]\}$が混合されている．一方，$s_p[n]$から$x'_{\{\mathrm{I,II}\}p'}[n]$への伝達におけるインパルス応答は式(4)の右辺括弧内なので，これを $h'_{\{\mathrm{I,II}\}p'p}[n]$とおくと，

$$\begin{pmatrix}x'_{\{\mathrm{I,II}\}1}[n]\\x'_{\{\mathrm{I,II}\}2}[n]\end{pmatrix}=\begin{pmatrix}h'_{\{\mathrm{I,II}\}11}[n] & h'_{\{\mathrm{I,II}\}12}[n]\\h'_{\{\mathrm{I,II}\}21}[n] & h'_{\{\mathrm{I,II}\}22}[n]\end{pmatrix}\otimes\begin{pmatrix}s_1[n]\\s_2[n]\end{pmatrix}$$
$$\mathbf{x}'_{\{\mathrm{I,II}\}}[n]=\mathbf{h}'_{\{\mathrm{I,II}\}}[n]\otimes\mathbf{s}[n]$$
$$\mathbf{X}'_{\{\mathrm{I,II}\}}(z)=\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\mathbf{S}(z) \tag{6}$$

となる．なお第 2 行は第 1 行のベクトル・行列表記, 第 3 行は第 2 行の z 変換表記である. ここで，$\mathbf{H}'_{\{\mathrm{I,II}\}}(z)$の逆行列，すなわち逆システムは，

$$\mathbf{S}(z)=\frac{\mathrm{cof}\left\{\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\right\}}{\det\left\{\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\right\}}\mathbf{X}'_{\{\mathrm{I,II}\}}(z) \tag{7}$$

となる．式(7)のシステムにおいて，$\det\{\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\}$は IIR(Infinite Impulse Response)システムとなっている．このため，式(7)は，$\det\{\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\}$が最小位相, すなわち $\det\{\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\}$=0 の零点がすべて単位円内に存在しない限り安定とならない．一般的に，$\det\{\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\}$に最小位相特性は期待できない．

ここで，本研究の目的を「音源分離+残響除去」と分けて考えると，式(7)において音源分離は$\mathrm{cof}\{\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\}$が，残響除去は $\det\{\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\}$が担当していることがわかる．そこで，この段階では音源の分離のみを行うこととする．つまり，

$$\begin{pmatrix}x''_{\{\mathrm{I,II}\}1}[n]\\x''_{\{\mathrm{I,II}\}2}[n]\end{pmatrix}=\begin{pmatrix}h'_{\{\mathrm{I,II}\}22}[n] & -h'_{\{\mathrm{I,II}\}12}[n]\\-h'_{\{\mathrm{I,II}\}21}[n] & h'_{\{\mathrm{I,II}\}11}[n]\end{pmatrix}\otimes\begin{pmatrix}x'_{\{\mathrm{I,II}\}1}[n]\\x'_{\{\mathrm{I,II}\}2}[n]\end{pmatrix} \tag{8}$$

となる．

式(8)によって音源分離された信号は，

$$\mathbf{X}''_{\{\mathrm{I,II}\}}(z)=\det\left\{\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\right\}\mathbf{S}(z)$$
$$\mathbf{x}''_{\{\mathrm{I,II}\}}[n]=\mathrm{ZT}^{-1}\left\{\det\left\{\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\right\}\right\}\otimes\mathbf{s}[n] \tag{9}$$

となる．ここで，$\mathrm{ZT}^{-1}\{\cdot\}$は逆 z 変換を表す演算子である．すなわち，$x''_{\{\mathrm{I,II}\}p}[n]$は，$s_p[n]$に残響システム $\mathrm{ZT}^{-1}\{\det\{\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\}\}$がたたみ込まれた信号である．さて，

$$\det\left\{\mathbf{H}'_{\{\mathrm{I,II}\}}(z)\right\} = H'_{\{\mathrm{I,II}\}11}(z)H'_{\{\mathrm{I,II}\}22}(z) - H'_{\{\mathrm{I,II}\}12}(z)H'_{\{\mathrm{I,II}\}21}(z) \quad (10)$$

であることから，$x''_{\{\mathrm{I,II}\}p}[n]$は実観測信号 $x_{\{\mathrm{I,II}\}k}[n]$ よりさらに長い残響を有することとなる．

2.4 MISO ウィナーフィルタ

ここまでの処理，遅延和ビームフォーマ，余因子行列による音源分離は，最初に 2 組に分けたマイクロフォンアレイごとに行ってきた．このため，残響システムが $\det\{\mathbf{H}'_{\mathrm{I}}(z)\}$と $\det\{\mathbf{H}'_{\mathrm{II}}(z)\}$で異なる 2 組の分離音 $x''_{\mathrm{I}\{1,2\}}[n]$と $x''_{\mathrm{II}\{1,2\}}[n]$が得られている．そこで，分離音ごとにペアを組み直し，アレイごとに重み $w_{\{\mathrm{I,II}\}}[n]$を設定し，MISO システムによる残響除去のウィナーフィルタを設計する．この MISO システムの z 変換表記は，

$$\begin{aligned}\hat{\mathbf{S}}_{\{1,2\}}(z) &= \begin{pmatrix}\mathbf{W}_{\mathrm{I}}(z) & \mathbf{W}_{\mathrm{II}}(z)\end{pmatrix}\begin{pmatrix}\mathbf{X}''_{\mathrm{I}\{1,2\}}(z)\\ \mathbf{X}''_{\mathrm{II}\{1,2\}}(z)\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{W}_{\mathrm{I}}(z) & \mathbf{W}_{\mathrm{II}}(z)\end{pmatrix}\begin{pmatrix}\det\left\{\mathbf{H}'_{\mathrm{I}}(z)\right\}\\ \det\left\{\mathbf{H}'_{\mathrm{II}}(z)\right\}\end{pmatrix}\\ &\quad \mathbf{S}_{\{1,2\}}(z)\end{aligned} \quad (11)$$

となる．つまり，$\mathbf{W}_{\mathrm{I}}(z)\det\{\mathbf{H}'_{\mathrm{I}}(z)\}+\mathbf{W}_{\mathrm{II}}(z)\det\{\mathbf{H}'_{\mathrm{II}}(z)\}$ が適当な時間シフトを持つデルタ状になるように最小 2 乗規範にもとづきウィナーフィルタを設計すればよいこととなる．

3 実験と結果・考察

標準的な実験室において実環境音声を収録し，2 節で述べた手法の評価を行った．

3.1 実験環境

図 2 に実験環境を示す．この実験室のサイズは 6.3(L)m×3.5(W)m×2.8(H)m であった．スピーカは 2 個設置し，DA 変換器(16bit@48kHz)を用いて模擬音声信号・音声信号を流した．マイクロフォンは 16 個設置し，1 個おきの順に組み合わせアレイ I, II とした．マイクロフォンで収録された信号は，AD 変換器(16bit@48kHz)を用いてサンプリングした．なお，DA 変換器と AD 変換器のクロックは同期させ，音声信号の周波数帯域に合わせ 2 節で述べた処理は 12kHz にダウンサンプリングして行った．

スピーカからマイクロフォンへの全組合せのインパルス応答は M 系列信号を用いた相関法によって計測した．残響時間(–60dB)の平均は 330ms 程度であった．

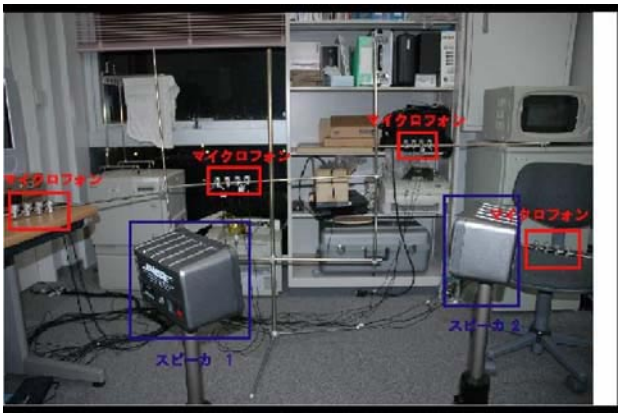

図 2 実験環境

3.2 音声認識エンジン評価

音声認識には，連続音声認識コンソーシアムによって開発された Julius2002 年度版を利用した．音声信号は Julius に付属していた音声信号 24 センテンス(16bit@16kHz，5.8s)をアップサンプリングして使用した．

表 1 に，残響が音声認識に与える影響を調べたシミュレーションの結果を示す．残響システムには，残響が異なる環境(120, 310, 380ms)において実計測されたインパルス応答を利用した[1]．このインパルス応答を PC 内で源音声信号に直接たたみ込むことによって，シミュレーションデータを作成した．なお，残響時間 0ms は，源音声そのものを意味している．

表 1 残響時間と音声認識率の関係

残響時間 [ms]	0	120	310	380
音声認識率 [%]	84.0	77.0	32.7	24.6

源音声では 84%であった音声認識率が，実験室とほぼ当程度の残響時間 310ms では 33%まで減少している．通常，この程度の残響時間であれば，人間による音声認識には全く問題とならない．しかし，このような残響時間であっても，機械による音声認識性能に与える影響は大きいことがわかる

3.3 音源分離性能評価

2 節において説明したように，中間出力信号に

おいても音源に関係する引数 p または p'が付記されている．そこで，遅延和ビームフォーマ後(処理 1 後)，余因子行列による音源分離後(処理 2 後)，MISO ウィナーフィルタ後(処理 3 後)の各段階において，音源分離性能を SIR(Signal to Interference Noise Ratio)によって評価した．

まず，p=1 のスピーカのみから M 系列信号を流した．p=1 または p'=1 の引数を持つ中間段階の出力を $a_1[n]$，p=2 または p'=2 の場合は $a_2[n]$と表すことにする．このとき，$a_1[n]$には M 系列信号が現れ，$a_2[n]$には分離できず漏れ出した信号が現れる．$a_{\{1,2\}}[n]$ともとの M 系列信号との循環相互相関関数 $r_{\{1,2\}}[n]$を求めると，この実験に本質的に関与した信号のみを抽出できる．つまり，M 系列のもつ自己相関特性によって，環境雑音などの影響は排除できる．そこで，SIR を

$$\mathrm{SIR} = 10\log\left(\frac{\sum_n r_1^2[n]}{\sum_n r_2^2[n]}\right) \tag{12}$$

と定義した．

処理 1, 2, 3 後の SIR を表 2 に示す．

表 2 処理の各段階と SIR との関係

処理	1	1-2	1-2-3
SIR [dB]	6.6	38.0	35.4

SIR は処理 2 後，すなわち余因子行列による音源分離後に大幅に向上していることがわかる．また，処理 3 後には 2.6dB 低下している．これらの原因は，積極的な音源分離は処理 2 のみが行っているためと考えられる．

3.4 残響除去性能評価

残響除去性能評価は，除去したい残響分を雑音と考え SNR(Signal to Noise Ratio)によって行う．ここでも，計測は 3.3 節と同じ方法で行い，SNR を

$$\mathrm{SNR} = 10\log\left(\frac{\sum_{n\in D} r_1^2[n]}{\sum_{n\notin D} r_1^2[n]}\right) \tag{13}$$

と定義した．なお，式(13)中の領域 D は，残響成分ではないと主観的に判断される時間区間であり，ここでは $r_1[n]$ピーク付近の約 8.3ms 間とした．

処理 1, 2, 3 後の SNR を表 3 に示す．

表 3 処理の各段階と SNR との関係

処理	1	1-2	1-2-3
SNR [dB]	6.0	–2.5	16.5

SNR は処理 3 後，すなわち MISO ウィナーフィルタ後に大幅に向上していることがわかる．また，処理 2 後には 8.5dB 低下している．これらの原因は，積極的な残響除去は処理 3 のみが行っているためと考えられる．

3.5 音声認識率による評価

最後に，先に述べた 24 センテンスを 2 組ずつ同時にスピーカ 1, 2 から流し，Julius による音声認識率を評価した．ここでは，3.3 節，3.4 節での処理 1, 2, 3 後に加え，処理 1 を省略し，処理 2, 3 のみを行った場合についても評価した．

これら各段階での音声認識率を表 4 に示す．

表 4 処理の各段階と音声認識率との関係

処理	1	1-2	1-2-3	2-3
音声認識率 [%]	12.4	34	75.8	66.1

音声認識率は，一連の処理をすべて行った場合に最大の 75.8%となっており，源音声信号直接の音声認識率 84.0%(表 1 参照)に非常に近い値となっている．また，遅延和ビームフォーマ処理を省略すると，音声認識率は 66.1%に減少するため，周波数一律の遅延和によるビームフォーマ処理は，前処理として効果があったことがわかる．

4 結論

実音響環境において混合された 2 音声の音源分離と残響除去を行い，機械認識に耐える音声信号を抽出する問題に取り組んだ．ここでは，マイクロフォンアレイの冗長さを利用して，遅延和ビームフォーマ，余因子行列よる音源分離，MISO ウィナーフィルタを段階的に利用する方法を提案した．個々の処理は，音源分離と残響除去に対してトレードオフの関係を持つが，組み合わせることで，理想音声認識率 84.0%に対して 75.8%の性能を達成した．

文献

[1] http://tosa.mri.co.jp/sounddb/micarray/index.htm