E-052

バージインを許容するロボット音声対話のための ICA を用いたセミブラインド音源分離

Semi-Blind Source Separation using ICA for Barge-In-Capable Robot Spoken Dialogue

武田 龍 † 中臺 一博 ‡ 駒谷 和範 † 尾形 哲也 † 奥乃 博 †
Ryu Takeda Kazuhiro Nakadai Kazunori Komatani Tetsuya Ogata Hiroshi G. Okuno

1. はじめに

人とロボットとの対話において，ロボットの発話中にユーザの発話を許容すること（バージイン）は，自然なインタラクションには不可欠である．ロボットに装着されたマイクでは自発話が入り込むため，バージインは相手の発話を認識する上で大きな障害になる．

我々は，バージインを許容するロボット音声対話を実現するため，独立成分分析（ICA）による自発話の直接入力信号を利用したセミブラインド分離を用いて，自発話のキャンセルを行う．ICA を使うメリットは（1）自発話区間の検出が不要，(2) ノイズが存在しても分離可能，という点がある．しかし，宮部ら [1] による手法では，反響のある環境において分離性能の低下を生じる．

本稿では，ICA によるセミブラインド音源分離において，時間周波数領域での畳み込みモデルを用いることで，反響による分離性能の劣化の低減化を狙う．同時発話孤立単語認識実験により，従来の ICA を用いた手法との性能比較を行う．

2. ICA を用いたセミブランド音源分離

2.1 音源信号の混合過程

問題を簡単にするため，2 音源の場合を扱い，$s_1(t)$ を未知音源，$s_2(t)$ を既知音源とする．2 音源モデルは一般性を失わず拡張できる．音が線形不変な伝達系を経て混合されると，その観測信号は次式で表される．

$$\mathbf{x}(t) = \sum_{n=0}^{N-1} \mathbf{a}(n)\mathbf{s}(t-n) \quad (1)$$

$$\mathbf{a}(n) = \begin{pmatrix} a_{11}(n) & a_{12}(n) \\ 0 & a_{22}(n) \end{pmatrix}. \quad (2)$$

$\mathbf{s}(t) = [s_1(t), s_2(t)]^T$ は音源信号ベクトル，$\mathbf{x}(t) = [x(t), s_2(t)]^T$ は観測信号ベクトル，$\mathbf{a}(n)$ は伝達系のインパルス応答を表す 2 行 2 列の混合行列である．ただし，既知音源に関する要素 $a_{22}(n)$ は瞬時混合であるため，$n=0$ のときのみ 1 であり，それ以外は 0 をとる．

2.2 周波数領域 ICA によるセミブラインド音源分離

周波数領域 ICA は時間領域 ICA よりも収束性が良いので，よく用いられる．窓長 T，シフト長 U による短時間フーリエ解析を行い，時間周波数領域での信号を得る．元信号 $s(t)$ 及び，観測信号 $x(t)$ は，フレーム f，周波数 ω をパラメータとした $S(\omega,f)$ と $X(\omega,f)$ で表現される．観測信号ベクトルは，$\mathbf{X}(\omega,f) = [X(\omega,f), S_2(\omega,f)]^T$. と記述できる．分離過程は以下のようになる．

$$\mathbf{Y}(\omega,f) = \mathbf{W}(\omega)\mathbf{X}(\omega,f), \quad (3)$$

$$\mathbf{W}(\omega) = \begin{pmatrix} W_{11}(\omega) & W_{12}(\omega) \\ 0 & 1 \end{pmatrix}. \quad (4)$$

ここで，$\mathbf{Y}(\omega,f) = [S_1(\omega,f), S_2(\omega,f)]^T$ は推定された元信号ベクトル，$\mathbf{W}(\omega) = [W_{ij}(\omega)]_{ij}$ は分離行列である．

† 京都大学大学院 情報学研究科 知能情報学専攻
‡(株) ホンダ・リサーチ・インスティチュート・ジャパン

分離行列の学習には，音声などの有色信号に有効な non-horonomic 拘束適用による KL 情報量最小化に基づく次の反復学習則を用いる [2].

$$\mathbf{W}^{[j+1]}(\omega) = \mathbf{W}^{[j]}(\omega) - \alpha\{\text{off-diag}\langle\phi(\mathbf{Y})\mathbf{Y}^H\rangle\}\mathbf{W}^{[j]}(\omega). \quad (5)$$

ここで，α は学習係数，$[j]$ は更新回数，$\langle\cdot\rangle$ は平均，off-diag $\mathbf{X}$ は対角要素を零に置き換える演算であり，非線形関数ベクトル $\phi(\mathbf{y})$ は $\phi(y_i) = \tanh(|y_i|)e^{j\theta(y_i)}$ である [2]. また，既知音源から既知音源への伝達特性は定数であるため，更新するのは $\mathbf{W}$ の 1 行目のみとなる．この手法は計算量は少ないが，(A) 反響音の遅延が窓長 T を超えると分離性能が低下する問題，及び，(B) 窓長を大きくとると ICA の分離性能自体が劣化する問題 [3] がある．

3. 本手法によるセミブラインド音源分離

3.1 時間周波数領域の畳み込みによる混合過程

以下，混合過程を時間周波数領域で考える．分離や音声認識特徴量抽出など後段の処理との整合性が良い．

本稿のアイデアは，次フレームに入り込んだ反響音を時間周波数領域における畳み込みで表現することで，(A)，(B) の問題に対処することにある．あるフレーム f の周波数成分が，M フレームにわたって観測信号の周波数成分に影響を及ぼすと仮定すると，次式のように記述できる．

$$X(\omega,f) = \sum_{m=0}^{M} A(\omega,m)S(\omega,f-m). \quad (6)$$

$A(\omega,m)$ は遅延 m である周波数成分 $S(\omega,f-m)$ の伝達関数である．これにより，同一の窓幅で残響へ対処が可能となる．模式図を図 1 に示す．本稿では自発話の抑制が目的であるため，既知音源である $S_2(\omega,f)$ に式 (6) を適用する．観測音 $X(\omega,f)$ は，畳み込まれた $S_2(\omega,f)$ と通常の伝達過程を経た $S_1(\omega,f)$ の混合とみなす．

3.2 周波数領域 ICA の適用

上述のモデルは線形混合過程とみなせるので，従来の周波数領域 ICA が適用できる．分離過程をベクトルで表現すると以下のようになる．

$$\mathbf{Y}(\omega) = \mathbf{W}(\omega)\mathbf{X}(\omega) \quad (7)$$

$$\mathbf{X}(\omega) = [X_1(\omega,f), S_2(\omega,f), \ldots, S_2(\omega,f-M)]^T \quad (8)$$

$$\mathbf{Y}(\omega) = [Y_1(\omega,f), S_2(\omega,f), \ldots, S_2(\omega,f-M)]^T \quad (9)$$

$$\mathbf{w}(\omega) = [w_0(\omega), w_1(\omega), \ldots, w_M(\omega)] \quad (10)$$

$$\mathbf{W}(\omega) = \begin{pmatrix} a(\omega) & \mathbf{w}(\omega) \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \quad (11)$$

$\mathbf{X}$ は観測信号ベクトル，$\mathbf{Y}$ は分離信号ベクトル，$\mathbf{W}$ は分離行列，$\mathbf{I}$ は $M+1$ 行 $M+1$ 列の単位行列である．

反射音をモデル化した $S_2(\omega,f),\cdots,S_2(\omega,f-M)$ 間の独立性は $S_1(\omega,f)$ の分離に影響しないことに注意する．式 (5)，(9)，(11) と $S_2(\omega,f),\cdots,S_2(\omega,f-M)$ が既知であることを考えると，$S_1(\omega,f)$ の分離には $S_1(\omega,f)$ と $S_2(\omega,f),\cdots,S_2(\omega,f-M)$ の独立性だけが評価され，$S_2(\omega,f)\cdots,S_2(\omega,f-M)$ 間の独立性は無関係である．

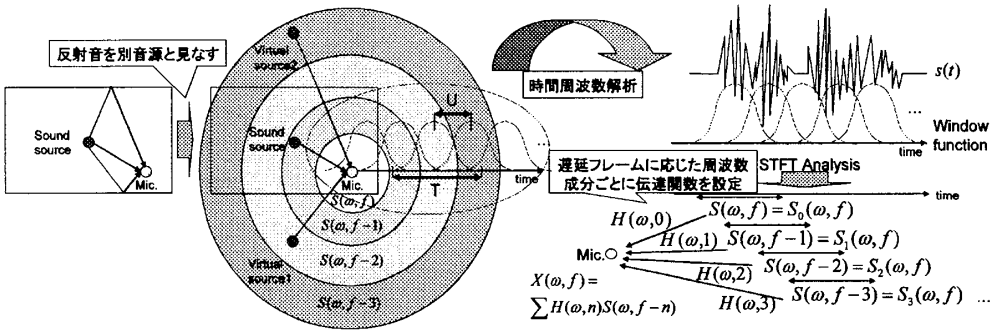

図 1: 時間周波数領域畳み込み：遅延フレームに応じた伝達関数を設定．それぞれ別音源と見なし，ICA を適用．

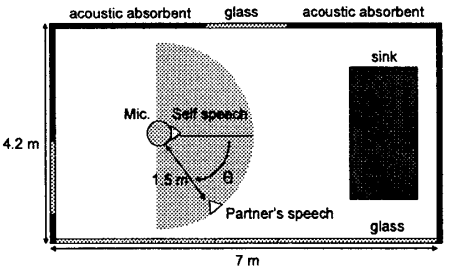

図 2: 配置関係：θ は正面からの角度（°）

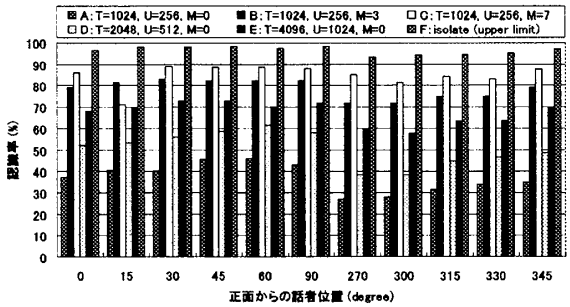

図 3: 認識結果：最大遅延フレーム数 M，角度 θ と認識率
A・B・C: それぞれ時間領域フィルタ長 $N=1024,2048,3072$ に相当，
D・E: 時間領域フィルタ長 $N=2048,4096$ に相当，F: 認識率の上限

周波数領域 ICA 特有の問題であるパーミュテーションは，未知音源が一つである場合には解く必要がない．もう 1 つのスケーリング問題は，歪みが最小となる Projection Back 法によって解決した [4].

4. 評価実験

ロボットと対話者からなるバージインを含む同時発話を想定し，インパルス応答を畳み込んだデータを用いて，同時発話 200 単語認識実験により評価を行う．

4.1 録音条件

インパルス応答は，4.2 m × 7 m の広さで録音した．残響時間（RT60）は約 0.3 秒である．自発話に対応するスピーカをマイク付近に設置し，この方向を正面とした．相手発話に対応するスピーカはマイクに向けて設置し，マイクとの距離は 1.5 m，位置は正面から右に向けて，0, 15, 30, 45, 60, 90, 270, 300, 315, 330, 345 度の 11 パターンとした．自発話，相手発話ともに男性話者を使用し，自発話の音量が対話者のそれよりも 10 dB ほど大きくなるように音量を設定した．

4.2 音声認識と分離パラメータ

音声認識エンジンは Julian[5] を使用した．音響モデルは，クリーン音声 23 話者 (男性 11 人, 女性 12 人) 分の ATR 音素バランス単語 216 語で学習したトライフォン (3 状態 4 混合の HMM) である．特徴量は，MFCC ($12+\Delta 12+\Delta Pow$) 25 次元である．ただし，認識に用いた話者の音声は学習データに含まれていない．

基本分離パラメータを示す．音声信号のサンプリングレートは 16kHz，窓長 T は 1,024 point（64 msec），シフト長 U は 256 point（16 msec），学習係数 α は 0.45 とした．分離性能の上限を比較するため，オフラインで処理を行い，十分に分離行列を学習させた．

4.3 実験結果及び考察

認識結果を図 3 に示す．図中の upper limit は単独発話の認識率である．最大遅延フレーム数 M を大きくすると認識率が大幅に向上していることが分かる．特に，A: $M=0$ の時（従来法）と C: $M=7$ の時（本手法）を比較すると，平均 47 points の認識率の向上が見られる．フィルタ長が実質同じである B と D を比較しても性能が向上している．これは，周波数領域での畳み込みモデルにより，効果的に自発話が抑制できることを示している．

逆に性能の向上の代価として，計算時間の増大と分離行列の更新回数増大がある．最大遅延フレーム数 M が直接影響を及ぼすため，M に関しては部屋の環境や処理に応じた設定が必要である．

5. おわりに

本稿では，バージインを許容するロボット音声対話を目指し，時間周波数領域畳み込みを用いた ICA により自発話の抑圧を行った．同時発話認識実験を行い，従来の ICA を用いた手法と比較し，性能向上を確認した．

今後は，計算量の軽減，最大遅延フレーム数 M に関して検討し，オンライン処理を目指す予定である．

謝辞 本研究の一部は科研費の支援を受けた．

参考文献

[1] S. Miyabe *et al.*: "Double-Talk Free Spoken Dialogue Interface Combining Sound Field Control with Semi-Blind Source Separation", *Proc. ICASSP 2006*, pp.809–812, 2006.

[2] Sawada *et al.*: "Polar Coordinate based Nonlinear Function for Frequency-Domain Blind Source Separation", *IEICE Trans. Fundamentals*, 3, E86-A, pp.505-510, 2003.

[3] S. Araki *et al.*: "The Fundamental Limitation of Frequency Domain Blind Source Separation for Convolutive Mixtures of Speech", *IEEE Trans. On Speech and Audio Proc.*, vol. 11, no.2, pp.109–116, 2003

[4] Murata *et al.*: "An approach to blind source separation based on temporal structure of speech signals", *Neurocomputing*, 41, pp.1-24, 2001.

[5] Julian: http://julius.sourceforge.jp/