6J-7

ロボット音声対話における Semi-blind ICA を用いた自己発話キャンセル

武田 龍† 中臺 一博‡ 高橋 徹† 駒谷 和範† 尾形 哲也† 奥乃 博†

†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

1. はじめに

人間同士の自然な対話では，相手の発話に割り込んで話すバージイン発話が良く笑われる．本稿では，人・ロボット音声対話をより自然にするために，バージイン発話を認識する手法として，ロボットが自分自身の発話をキャンセルし，相手の発話だけを聞き分けて音声認識する手法について報告する．

人・ロボット対話では，ロボットは自分自身に装着されたマイクを使って音を聞くので，相手発話だけでなく，次のような音の混入がある（図 1）．

1. 自分自身のロボット発話（既知）
2. ユーザ発話・ロボット発話の反響音

種として，上記 2 種の混入により，相手発話の音声認識性能は大きく低下する．

一方，バージイン発話は音声対話にとって有益な情報も有している．例えば，バージインの頻度やタイミングは，ユーザモデルの認識や発話内容の検証に使用できる [1, 2]．従って，バージイン発話を許容し，ユーザ発話だけを抽出することは音声対話システムを高機能化する上で不可欠な技術である．

我々は，独立成分分析 (ICA) という統計的信号処理の技術を応用することで，ロボット音声対話でのバージイン機能の実現を図ってきた．ICA は事前情報を必要としない音源分離手法であり，低演算量や他の技術との高い親和性などの理由から，周波数領域 ICA（FD-ICA）が実用上よく利用されている．しかし，従来の FD-ICA では，残響（反響音）を分離できない問題があった [3]．

我々はこの問題を，1) 短時間フーリエ変換（STFT）領域における音の混合モデル化，2) 観測音との独立性条件を用いた ICA の適用，により解決を行っている．これにより，残響時間に対して線形オーダの演算量で残響に対応することが可能となる．残響のある環境下での，連続音声認識実験により，本手法の有効性を評価する．

2. STFT 領域における音の混合モデル

本稿では，ユーザ発話とロボット発話を含む混合過程を STFT 領域で記述する [4]．そのため，基本的に STFT 後のスペクトルを扱い，用いる変数はすべて複素数である．以降，表記の簡略化のため，周波数インデックス ω を省略するが，実際は各周波数で同様のモデル化・処理を行っている（図 2）．

Self-speech cancellation with Semi-blind ICA for Robot speech interaction *Ryu Takeda (Kyoto Univ.), Kazuhiro Nakadai (HRI Japan), Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

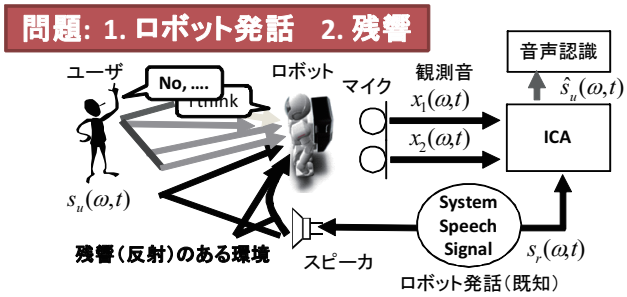

図 1: 本稿で想定する状況およびデータフロー

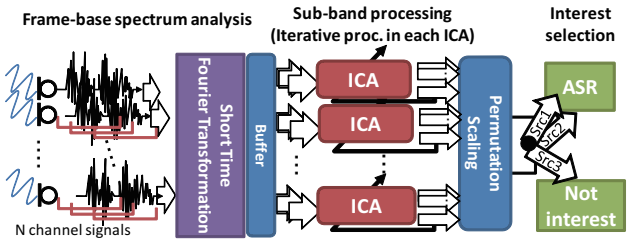

図 2: Semi-blind ICA の処理概要

フレーム t における，マイクロホン $1, \ldots, L$ の観測スペクトルをそれぞれ $x_1(t), \ldots, x_L(t)$ と表わす．また，ユーザと既知であるロボット発話のスペクトルをそれぞれ $s_u(t), s_r(t)$ と表す．本稿では以下の STFT 領域における線形混合モデルを仮定する．

$$x_j(t) = \sum_{n=0}^{K_u} h_j^u(n) s_u(t-n) + \sum_{m=0}^{K_r} h_j^r(m) s_r(t-m) \quad (j = 1, 2, ..., L) \qquad (1)$$

ここで，h_j^u，h_j^r はそれぞれマイク j に関するユーザ発話，ロボット発話の伝達係数であり，K_u，K_r はフィルタ長である．$K_u = K_r = 0$ の場合が従来の FD-ICA のモデルに相当する．そのため，音声認識技術，例えば，ケプストラム平均除去や音響モデル適応，で対処が困難な，フレーム外に渡る残響に対応できない．課題は，この $x_j(t), (j = 1, ..., L)$ のみを用いて，残響を含まないユーザ発話スペクトル $s_u(t)$ を抽出することである．

3. Semi-blind ICA を用いたロボット発話キャンセルと残響抑圧

音声は，異なる音源間及び異なる時間の要素間でも，ある程度統計的な独立性を有している．ユーザ発話は，ロボット発話とは独立した音源であり，また時間的な独立性を有している．ICA は出力信号が互いに独立となるように処理を行うため，出力信号がこれらと独立となるようにモデ

表 1: 実験設定

インパルス応答	16 kHz サンプリング
残響時間 (RT_{20})	240 [ms]，670 [ms]
話者位置	1.5 m，$\{0°, 45°, 90°, -45°, -90°\}$
マイク数	2 (embeded at ASIMO's head)
STFT	Hanning 窓: 32 [ms]，シフト長: 12 [ms]
入力データ	-1.0 – 1.0 正規化

表 2: 音声認識設定

テストセット	200 文章
学習セット	200 話者 (各 150 文章)
音響モデル	PTM-Triphone: 3-state, HMM
言語モデル	統計モデル，語彙 20k
音声解析	Hanning 窓: 32 [ms], シフト長: 10 [ms]
特徴量	MFCC 25 dim.(12+Δ12+ΔPow)

ルを設計すれば，ユーザ発話の直接音を抽出できると期待できる．まず，観測ベクトル $\boldsymbol{X}(t) = [\boldsymbol{x}(t), ..., \boldsymbol{x}(t-N_o)]$ と，ロボット発話ベクトル $\boldsymbol{S}_r(t) = [s_r(t), ..., s_r(t-N_r)]$ を定義する．N_o，N_r はタップ数である．分離音 $\hat{\boldsymbol{s}}(t)$ が $\boldsymbol{X}(t-d)$ と $\boldsymbol{S}_r(t)$ に独立となるような分離モデルを設定すれば，ユーザ発話の直接音が抽出される．ここで，$\boldsymbol{X}(t-d)$ と独立という条件は，分離音の時間的な独立性を間接的に評価していることになる [5]．また，d は，直接音と隣接する要素間の独立性が小さいことを考慮するための定数である．このような ICA 分離モデルは次式のようになる．

$$\begin{pmatrix} \hat{\boldsymbol{s}}(t) \\ \boldsymbol{X}(t-d) \\ \boldsymbol{S}_r(t) \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}_u & \boldsymbol{W}_d & \boldsymbol{W}_r \\ \boldsymbol{0} & \boldsymbol{I}_d & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_r \end{pmatrix} \begin{pmatrix} \boldsymbol{x}(t) \\ \boldsymbol{X}(t-d) \\ \boldsymbol{S}_r(t) \end{pmatrix} \quad (2)$$

ここで，$\boldsymbol{W}_u$ と $\boldsymbol{W}_d$ はそれぞれ，サイズが $L \times L$ と $L \times LN_o$ の分離行列，$\boldsymbol{W}_r$ は $L \times (N_r+1)$ サイズの分離行列，$\boldsymbol{I}_d$，$\boldsymbol{I}_r$ は対応する適切なサイズの単位行列である．通常の ICA と異なり，一部分は直接伝達するモデルであり，この ICA を Semi-blind ICA と呼ぶことにする．また，$N_r = 0$ かつ $\boldsymbol{W}_d$ を用いない時，従来の FD-ICA に相当する．

Kullback-Leibler Divergence を自然勾配法で最小化することで，$\boldsymbol{W}_u, \boldsymbol{W}_d$ と $\boldsymbol{W}_r$ の学習則を得る．

$$\boldsymbol{D} = \boldsymbol{\Lambda} - \mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\hat{\boldsymbol{s}}^H(t)] \quad (3)$$

$$\boldsymbol{W}_u^{[j+1]} = \boldsymbol{W}_u^{[j]} + \mu \boldsymbol{D}\boldsymbol{W}_u^{[j]} \quad (4)$$

$$\boldsymbol{W}_d^{[j+1]} = \boldsymbol{W}_d^{[j]} + \mu\Big(\boldsymbol{D}\boldsymbol{W}_d^{[j]} - \mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\boldsymbol{X}^H(t-d)]\Big) \quad (5)$$

$$\boldsymbol{W}_r^{[j+1]} = \boldsymbol{W}_r^{[j]} + \mu\Big(\boldsymbol{D}\boldsymbol{W}_r^{[j]} - \mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\boldsymbol{S}_r^H(t)]\Big) \quad (6)$$

ここで，j は反復回数番号，E は時間平均を表す演算子，μ は学習係数，$\boldsymbol{\phi}(\boldsymbol{x}) = [\phi(x_1), \cdots, \phi(x_L)]^T$ は非線形関数ベクトルであり，$\phi(x)$ として $e^{j\theta(x)}$ を用いる．$\boldsymbol{\Lambda}$ は収束性が良いとされる $\boldsymbol{\Lambda} = \mathrm{diag}(\mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\hat{\boldsymbol{s}}^H(t)])$ を用いる [6]．本手法の演算量は残響時間の線形オーダであるので，マイク数によっては実時間実行も可能である．

4. バージイン発話認識による本手法の評価

4.1 実験設定

JNAS 評価データセット 200 文章に録音したインパルス応答を畳み込み，バージイン発話の混合音（1 ユーザ発話+ロボット発話）を作成した．インパルス応答は 2 種類の環境で，ロボット頭部に設置されたマイクロホンを用いて測定したものである．その他の実験設定を表 1，表 2 にまとめる．

ICA の分離行列の初期値，スケーリング及びパーミュテーションは文献 [5] と同じである．学習係数は適応処理 [7] を行い，分離行列推定における反復回数は 20 を限度とした．また，3 秒毎のブロック処理として ICA を適用している．タップ数 N_o, N_r は同じ値を用い，RT_{20}: 240 [ms] 環境では 14，RT_{20}: 670 [ms] 環境では 35 とした．初期遅延パラメータは $d = 2$ とした．

表 3: 平均単語正解率 (%)

	Only spc.	No proc.	FD-ICA	Ours
RT_{20}: 240 [ms]	74.3	28.2	36.8	71.7
RT_{20}: 670 [ms]	26.1	11.0	10.5	43.3

4.2 実験結果及び考察

表 3 に単語正解率を示す．Only spc. は残響込のユーザ発話のみの認識率，No proc. は混合音の認識率，FD-ICA は従来法適用時の認識率，Ours は本手法適用時の認識率である．本手法は残響に対応できているため，様々な環境下で従来法の性能を上回っていることがわかる．また，RT_{20}: 670 [ms] 環境下では，ユーザ発話のみの認識率を上回っていることから，ユーザ発話残響除去の達成が確認できる．

5. おわりに

本稿では，ロボット音声対話におけるバージイン発話の認識を，ICA を応用することで実現した．単語正解率が最大 30 ポイント改善したことから，本手法の有効性が確認できた．今後は，ICA のオンライン実装，雑音抑圧処理との統合を行い，ロボット音声対話システム全体の評価を行う予定である．

謝辞 科研費，グローバル COE，および科研費奨励金の支援を受けた．

参考文献

[1] K. Matsuyama *et al.*: "Enabling a User to Specify an Item at Any Time During System Enumeration," in *Interspeech09*, 252-255.
[2] K. Komatani *et al*: "Predicting ASR Erros by Exploiting Barge-in Rate of Individual Users for Spoken Dialogue System," in *Interspeech08*, 183–186.
[3] S. Araki *et al.*: "The Fundamental Limitation of Frequency Domain Blind Source Separation for Convolutive Mixtures of Speech", *IEEE Trans.* , vol. 11, no.2, pp.109–116, 2003.
[4] T. Nakatani *et al.*: "Blind Speech Dereverberation with Multichannel Linear Prediction based on Short Time Fourier Transform Representation," in *ICASSP08*, 85–88.
[5] 武田 他: "残響下でのバージイン発話認識のための多入力独立成分分析を応用したロボット聴覚", 日本ロボット学会誌, Vol.27, No.7/8, pp.782-792, 2009.
[6] S. Choi, *et al.*:"Natural Gradient Learning With a Nonholonomic Constraint for Blind Deconvolution of Multiple Channels," *Proc. of International Workshop on ICA and BBS*, 371–376.
[7] Ryu Takeda *et al*: "Step-size Parameter Adaptation of Multi-channel Semi-blind ICA with Piecewise Linear Model for Barge-in-able Robot Audition," in *IROS09*, 2273–2282.