ロボット音声対話のための MFTとICAによるバージイン許容機能の評価

武田 龍† 中臺 一博‡ 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

本稿では，自然な人・ロボット対話の実現を目指し，ロボット自身の発話を抑圧して，ユーザの割り込み発話（バージイン）を認識する機能を報告する．

具体的には，独立成分分析 (ICA) に基づく適応フィルタ (ICA-AF)，ミッシングフィーチャ理論（MFT）に基づく音声認識（MFT-ASR）という2種類の手法を組み合わせてバージイン機能の実現を目指す．ICA-AFは，残響や雑音に対して頑健である [1]．しかし，音声認識性能を低下させる分離歪みを生じる．MFT-ASRは特徴量に信頼度を与えることでICA-AFの分離歪みに対応でき，音響モデルの環境適応が不要という利点がある反面，信頼度推定が複雑で困難という欠点がある [2]．

この信頼度推定問題については，ロボット自身の発話は既知であるというバージイン特有の問題設定を利用し，従来法よりも計算コストの低い手法を開発した．同時発話の孤立単語認識実験により，提案手法の評価を行う．

2. システム

本システムの構成を図1に示す．各機能について，以下に説明する．

2.1 ICA-AF: ICA に基づく適応フィルタ

計算コスト，分離性能の点で優れている時間-周波数領域処理を考える [3]．窓長 T，シフト長 U による短時間フーリエ解析（STFT）を行うことで時間-周波数領域でのスペクトル信号が得られる．フレーム f，周波数 ω における，ロボット発話の既知スペクトル及びマイク入力の観測スペクトルを $S(\omega,f)$ と $Y(\omega,f)$ で表現する．この時，観測スペクトルを次のようにモデル化する．

$$Y(\omega,f)=\sum_{m=0}^{M} W(\omega,m)S(\omega,f-m)+E(\omega,f), \quad (1)$$

$W(\omega,m)$ は遅延が m フレームの既知スペクトル $S(\omega,f-m)$ の重み係数，M は最大遅延フレーム数，$E(\omega,f)$ はユーザ発話のスペクトルである．ただし，重み係数 W は時不変であると仮定する．

Design and Evaluation of Barge-In enable Robot Audition System with ICA and MFT-based ASR *Ryu Takeda (Kyoto Univ.), Kazuhiro Nakadai (HRI Japan), Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

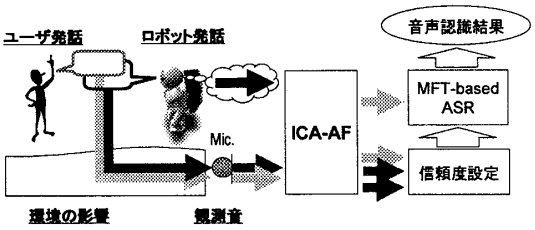

図1: バージイン許容機能の実装の構成

ロボット発話を別音源として扱い，ICAを用いて分離するため，分離過程を以下のように表現する．

$$\begin{pmatrix}\hat{E}(\omega,f)\\ \boldsymbol{S}(\omega,f)\end{pmatrix}=\begin{pmatrix}B(\omega) & -\hat{\boldsymbol{W}}^{T}(\omega)\\ \boldsymbol{0} & \boldsymbol{I}\end{pmatrix}\begin{pmatrix}Y(\omega,f)\\ \boldsymbol{S}(\omega,f)\end{pmatrix} \quad (2)$$

$$\boldsymbol{S}(\omega,f)=[S(\omega,f),S(\omega,f-1),\ldots,S(\omega,f-M)]^{T} \quad (3)$$

$$\hat{\boldsymbol{W}}(\omega)=[\hat{W}_0(\omega),\hat{W}_1(\omega),\ldots,\hat{W}_M(\omega)] \quad (4)$$

$\boldsymbol{S}$ は元信号ベクトル，$\hat{E}$ は雑音スペクトル，$\hat{\boldsymbol{W}}$ と B は分離フィルタ，$\boldsymbol{I}$ は $M+1$ 次の単位行列である．

Kullback-Leibler 情報量を自然勾配法と非ホロノミック拘束を用いて最小化することで，修正した確率勾配アルゴリズムが得られる [3, 4]．各フレームにおけるユーザ発話スペクトル及びフィルタ係数の推定値 $\hat{E}(f)$, $\hat{\boldsymbol{W}}(f+1)$ が次式で与えられる．表記の簡単化のため，インデックス ω は省略した．

$$\hat{E}(f)=Y(f)-\boldsymbol{S}(f)^{T}\hat{\boldsymbol{W}}(f) \quad (5)$$

$$\hat{\boldsymbol{W}}(f+1)=\hat{\boldsymbol{W}}(f)+\mu\phi(A(f)\hat{E}(f))\bar{\boldsymbol{S}}(f) \quad (6)$$

$$A(f+1)=A(f)+\mu[1-\phi(A(f)\hat{E}(f))\bar{A}(f)\bar{E}(f)]A(f) \quad (7)$$

ここで $\bar{x}$ は x の複素共役，μ は学習係数，A はスケーリング係数，非線形関数は $\phi(y)=\tanh(|y|)e^{j\theta(y_i)}$ である [4]．逐次的に推定を行うため，時変な W にもある程度の追従が可能である．

2.2 MFT-ASR: MFT に基づく音声認識

MFTは特徴量に信頼度を設け，出力確率の計算に反映することにより，特徴量歪みに対して頑健な音声認識を実現する手法である．本稿では2値信頼度に基づく手法を利用する [5]．

2値信頼度に基づく手法では，次元 d，フレーム f における音声特徴量 $F(d,f)$ に対して，その特徴量の信頼度 $M(d,f)$ を設定する必要がある．信頼度 $M(d,f)$ は特徴量が信頼できる場合に1を，できない場合は0の値を設定する．MFTに基づく音声認識の課題はいかに正しい信頼度を付与するかである．

表 1: 実験設定

Impulse Response	16kHz sampling
Room	4.2×7.0×3.2 m, RT20: 250 msec.
Distance	1.5 m
Input data	-1.0〜1.0 normalized
TestSet	1 male, 1 female (each 200 words)
TrainingSet	11 males, 12 females (each 216 words)
Acoustic Model	Triphone: 3-state 4-mix. HMM
Feature	MFCC, 25 dim. (12+Δ12+ΔPow)

3. 信頼度推定

ユーザ発話の特徴量の信頼度 $M(d,f)$ を推定しなければならない. MFT に関する研究では, 信頼できない特徴量を信頼できると誤る方が認識率への影響が大きいという知見が得られている [2]. この知見から, 少しでもロボット発話が影響を及ぼす特徴量は信頼しないという方針をとる.

今, ユーザ発話の推定スペクトルを $\hat{E}(\omega,f)$, マイク入力の観測スペクトルを $X(\omega,f)$ とする. 式 (5) から, この 2 つのスペクトルの差の要因は既知スペクトル $S(\omega,f)$ である. そのため, 特徴量において, これら 2 つの差が大きい部分はロボット発話 $S(\omega,f)$ の影響を強く受けており, 信頼できないとする. ここで, $F_e(d,f)$ 及び $F_x(d,f)$ により $\hat{E}(\omega,f)$ と $X(\omega,f)$ の フレーム f, 次元 d における特徴量を表す. この時, 特徴量 $F_e(d,f)$ の信頼度 $M(d,f)$ は次式に従って設定できる.

$$M(d,f) = \begin{cases} 1, & |F_e(d,f) - F_x(d,f)| < T_{th} \\ 0, & otherwise \end{cases} \quad (8)$$

ここで, T_{th} は閾値を意味する.

従来法では, 信頼度設定を行うため基本周波数推定, 調波構造抽出・追跡が必要であった [2, 5]. 一方, 本手法ではその問題設定により適応フィルタの出力を用いることで信頼度設定が可能である. 計算量は観測スペクトルの特徴量を抽出分のみである.

4. 同時発話認識による本手法の評価

4.1 実験設定

男女各 200 語の ATR 音素バランス単語に録音したインパルス応答を畳み込み, 同時発話の混合音を作成した. インパルス応答は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. MFT に基づく音声認識は Multi-band Julian [6] を使用した. その他の実験設定を表 1 にまとめる.

ステップサイズ μ は 0.004, 窓長 T は 1024 (64 msec.), シフト長 U は 128 (8 msec.) に設定, 最大遅延フレーム数 M を 4, 閾値 T_{th} は 40 種類の値, $n10^k, (n=1,2,...,9, k=-2,-1,1,2)$ を調べた. 分離処理は逐次的に行い, 事前に 3 秒ほど自発話のみのデータで学習した.

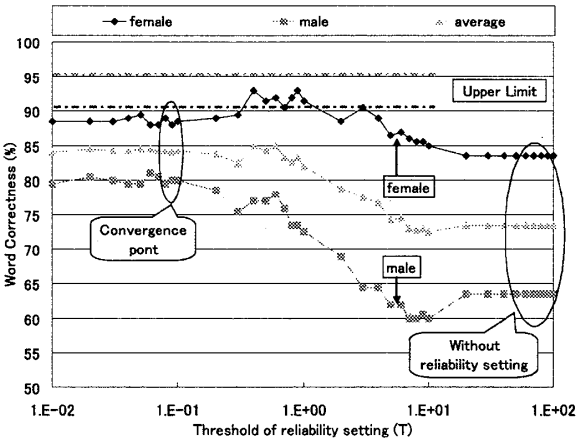

図 2: 単語正解率と信頼度閾値: ICA-AF なしの単語正解率は男性話者で 1%, 女性話者で 11.5% である.

4.2 実験結果及び考察

図 2 にマスク閾値 T_{th} と単語正解率との関係を示す. 閾値 $T_{th}=10^2$ の場合は信頼度設定を行わない場合と同値である. 単語正解率の上限値は単独発話の単語正解率であり, 男性話者で 95%, 女性話者で 91.5% である. また, ICA-AF による分離がない場合, 男性話者・女性話者の単語正解率はそれぞれ 1%, 11.5% である.

マスク閾値がある程度小さければ, MFT に基づく音声認識は一定の効果をもたらしている. $T_{th}=10^{-1}$ の場合, 信頼度設定なしと比較すると, 女性話者で約 5 ポイント, 男性話者で約 16.5 ポイント改善した.

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機能を, ICA-AF と MFT に基づく音声認識により実現した. 孤立単語認識率が平均 10 ポイント改善したことから, 開発した信頼度設定手法の有効性が確認できた. 今後は雑音抑圧処理とも統合し, システムの頑健性を高める予定である.

謝辞 科研費, グローバル COE の支援を受けた.

参考文献

[1] J. Yang, *et al.*": "A New Adaptive Filter Algorithm for System Identification Using Independent Component Analysis" *Proc. ICASSP 2007*, pp.1341–1344, 2007.

[2] Raj, *et al.*: "A Bayesian Framework for Spectrographic Mask Estimation for Missing Feature Speech Recognition", *Speech Communication*, pp.379–393, 2004

[3] 武田他: "独立成分分析に基づく適応フィルタのロボット聴覚への応用", 日本ロボット学会第 25 回大会, 1N6 , Sep. 2007.

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

[5] 山本他: "ミッシングフィーチャ理論に基づく音声認識を利用した複数話者同時発話認識," 計測自動制御学会, vol. 46, pp. 447–452, 2007.

[6] 西村他: "周波数毎の重みつき尤度を用いた音声認識の検討", 日本音響学会 2004 年春季研究講演論文集, pp.117-118, 2004.