6ZB-6

ICA と MFT に基づく音声認識における Soft Mask を用いた性能評価

武田 龍　　山本 俊一　　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

1. はじめに

我々は, 多様で動的な音響環境でも稼動するロボット聴覚機能を実現するために, 事前知識が最小となるシステムを研究してきた. 具体的には, 音源分離として音源方向情報等が不要な Blind Source Separation (BSS) である独立成分分析 (ICA) を, 音声認識には単一の音響モデルを使用したミッシングフィーチャ理論 (MFT) による音声認識 (ASR) を採用している.

システムは, 図 1 に示した 3 つのモジュール, (1) ICA による音源分離, (2) Mask 生成, (3) MFT に基づく音声認識, から構成されている. 本システムの特徴は, 目的音を定めず, すべての音源からの音声に対して認識を行っていることである. 従来のカクテルパーティ効果を狙った研究ではなく, 聖徳太子ロボットを目指した研究といえる. 本システムでは, ICA と MFT-based ASR という個別技術の性能向上に加えて, Mask 自動生成が極めて重要である. これまで, 2 値信頼度 (Hard Mask) を使用し, ICA の分離出力よりもさらに約 9% の性能向上を得ていた. しかし, MFT 研究分野では, より効果的と言われる 連続値特徴量 (Soft Mask) は未使用であった.

本稿では, Soft Mask 生成法を設計し, 2 話者の同時発話孤立単語認識実験により, その評価を行う.

2. 要素手法

2.1 ICA による音源分離

一般に, 複数の音源信号が線形不変な伝達系を経て混合された場合, その観測信号は次式で表される.

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

ただし, $s(t) = [s_1(t),...,s_I(t)]^T$ は音源信号ベクトル, $x(t) = [x_1(t),...,x_J(t)]^T$ はマイクロホンアレーにおける観測信号ベクトル, $a(n) = [a_{ji}(n)]_{ji}$ は伝達系のインパルス応答を表す J 行 I 列の混合行列である. なお, $[\cdot]_{ji}$ は j 行 i 列要素が・である行列を表す. 本稿では音源数 I とマイクロホンの数 J は等しく 2 であると仮定する.

ICA は収束の早い周波数領域で適用する. 短時間分析を用いてフレーム毎に離散フーリエ変換された信号を入力とする. すなわち, 観測信号ベクトルは $X(\omega,t) = [X_1(\omega,t),...,X_J(\omega,t)]$ と表現できる. 分離行列 W を用いて, 分離信号 $Y(\omega,t) = [Y_1(\omega,t),...,Y_I(\omega,t)]$ を周波数毎に独立に以下の式で求める.

$$Y(\omega,t) = W(\omega)X(\omega,t) \tag{2}$$

行列の学習には 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 [3].

$$W^{j+1}(\omega) = W^j(\omega) - \alpha\{\text{off-diag}\langle\phi(y)y^h\rangle\}W^j(\omega) \tag{3}$$

ここで, α は学習係数, $[j]$ は更新回数, $\langle\cdot\rangle$ は平均である. また, off-diag (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 $\phi(y)$ は $\phi(y_i) = \tanh(|y_i|)e^{j\theta(y_i)}$ である [3]. 周波数領域 ICA 特有の問題であるスケーリング問題に関しては, 歪み検出を容易にするため Projection Back を使用している [4]. また, パーミュテーション問題はマイクロホン間の強度差を用いて解決した. というのは, ロボットの体の伝達関数の影響を受けて, マイクロホン間の強度差が強調されるからである.

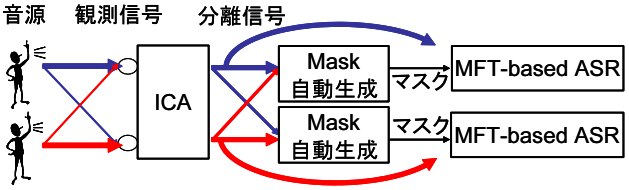

図 1: システムの概要

2.2 MFT に基づく音声認識

MFT-based ASR では, 認識率精度向上のために各音声特徴量に対し信頼度を定める. 出力確率計算時に信頼度を組み込む方法は Marginalization と呼ばれている [5].

Marginalization 法では, 2 値の信頼度を設定する Hard Mask と連続値信頼度を設定する Soft Mask がある. Soft Mask の場合, 状態 S の時, x である確率 $f(x|S)$ は以下のように定義される [6].

$$f(x|S) = \sum_{k=1}^{K} P(k|S) \sum_{d=1}^{D} (M(d)f(x(d)|k,S) + (1-M(d))\frac{1}{u(d)-l(d)}\int_{l(d)}^{u(d)} f(x(d)|k,S)dx) \tag{4}$$

ここで, $x(d), M(d)$ はそれぞれ次元 d における特徴量, とその信頼度を表し, K は混合数, D は次元数を意味する. 2 値信頼度の場合, 上式は第 1 項のみとなる. 連続値信頼度の場合, 2 値信頼度にはない積分範囲 (Upper/Lower Bound) $u(d), l(d)$ を設定しなければならない. この結果, Hard Mask の場合, 信頼できない特徴量 ($M(d) = 0$) は完全に尤度に反映されないのに対し, Soft Mask では尤度関数と Bound に従った値が出力されるようになる.

3. Soft Mask の生成

3.1 Soft Mask 生成におけるパラメータ設定

Soft Mask 生成に関する次の 2 つの課題が性能に大きく寄与する.

1) 連続値信頼度 $M(d,t)$ の設定
2) Upper/Lower Bound の設定

1) に関しては, 既開発の手法 [1] で自動検出した歪み $D(d,t)$ にシグモイド関数を掛けた値とする.

$$D(d,t) = |G(\hat{Y}_i,d,t) - G(\hat{Y}_i - \gamma(X_i - \hat{Y}_i),d,t)| \tag{5}$$

$$M(d,t) = \exp\left(\frac{1}{1+\alpha(D(d,t)-\beta)}\right) \tag{6}$$

Design and Evaluation of Soft Mask with ICA BSS and MFT-based ASR: *Ryu Takeda, Shunichi Yamamoto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

ここで, $\hat{Y}_i$ はスケーリング処理を施された目的信号, X_i は i 番目のマイクロホンで観測された観測信号である. また, G は特徴量への写像, $M(d,t)$ は次元 d, フレーム t における自動設定した信頼度, α, β, γ は適当な係数である.

2) に関しては, 特徴量の算出過程に依存するため, 特徴量に合わせた設定方法を考える必要がある. 一般的に MFT に基づく音声認識では, パワースペクトル等を用い, 歪みがほぼ単調増加とみなした上で,

$$u(d) = x(d),\ l(d) = -\infty \tag{7}$$

のように設定される. しかし, このような正規化・平滑化を行わない特徴量では, 伝達系の影響やノイズによる影響を受けやすく, 頑健性に欠ける. 本研究では, 正規化・平滑化されたスペクトル特徴量を用いるため [7], 従来の Bound 値が適切とはいえない.

3.2 歪みの方向に基づく Bound の設定

従来の Bound は, 特徴量における歪みがほぼ一方向に生じるため, 現在の特徴量値から真の特徴量値に向かう方向だけに設定すればよかった. しかし, 一般的に特徴量は負の方向にも正の方向にも歪みが生じる. 従って, 式 (4) を考慮して, 歪みがある特徴量から歪みが無い場合の特徴量に向かうように Bound を設定する.

次元 d, フレーム t における upper bound, u と lower bound, l を分離音の特徴量 $F(d,t)$ とそのクリーン音声の特徴量, つまり, 真の特徴量 $T(d,t)$ を用いて以下のように決定する.

$$u(d,t) = \begin{cases} \infty & F(d,t) \le T(d,t) \\ F(d,t) & F(d,t) > T(d,t) \end{cases} \tag{8}$$

$$l(d,t) = \begin{cases} F(d,t) & F(d,t) > T(d,t) \\ -\infty & F(d,t) \le T(d,t) \end{cases} \tag{9}$$

4. Soft Mask による効果の評価

ヒューマノイド SIG2 の外耳道モデル (図 2) に埋め込まれた 2 本の無指向性マイクロホンで, 2 話者同時発話孤立単語認識を行った.

(A) 分離音そのもの,
(B) Hard Mask (従来手法),
(C) 式 (7) に基づいた Bound による Soft Mask I,
(D) 式 (8, 9) に基づいた Bound による Soft Mask II,

の 4 種類を比較する. なお, パラメータは実験的に定め, Soft Mask I では $\alpha = 1.0$, Soft Mask II では $\alpha = 10.0$ を用い, 他は共通の $\beta = 0.05$, $\gamma = 0.01$ を用いている.

4.1 録音条件

録音を行った部屋は 4m × 5m の広さで, 残響時間 (RT20) が約 0.2 秒であった. 男性・女性話者で, マイクとスピーカの距離は約 1m, スピーカーの配置は 1 つが正面固定・もう一つが右側に 30 度, 60 度, 90 度間隔で配置した 3 パターンである (図 3). 正面に女性話者, 右側に男性話者とした.

4.2 MFT-based ASR

Casa Toolkit (CTK) を MFT に基づく音声認識として利用した [8]. 音響モデルには, クリーン音声 25 話者 (男性 13 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) を作成した. スペクトル特徴量は 24 + Δ24 次元の計 48 次元である. ただし, 認識に用いた音声の話者は学習データに含まれていない.

図 2: SIG2 と耳モデル

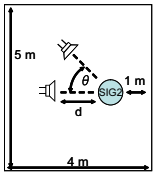

図 3: 話者配置

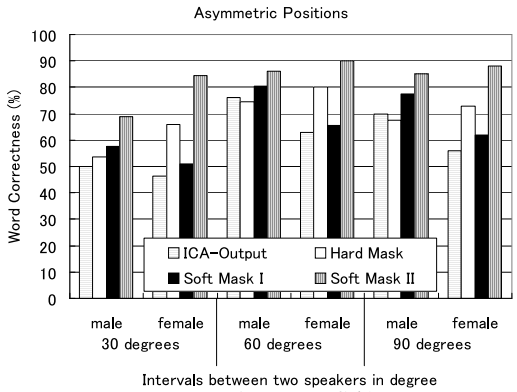

図 4: 同時発話認識結果

4.3 実験結果及び考察

2 話者同時発話認識結果の認識率を図 4 に示す. (C) による Bound 設定では一般に性能が悪い. 性能は, (D) > (B) > (A) となり, (D) は (B) より 平均 14 points 向上, (B) は (A) より 9 points 向上している.

以上から, 式 (8, 9) で示した Bound 設定による Soft Mask が Hard Mask よりも性能に大きく貢献することがわかった. この実験により, 歪んだ特徴量から真の特徴量への方向情報を付加するだけで, 性能が大幅に向上することが示された. このことから, 歪みに基づく Bound の設定が音源分離処理との親和性が高いことが考えられる. 音源分離処理によって特徴量は真の値に近づくため, 歪みの方向は分離結果を用いることで推定できる可能性がある.

5. おわりに

実験では 2 話者同時発話認識を行い, 自動設定した連続値信頼度と歪みの方向に基づいた Soft Mask により, 従来の Hard Mask よりも性能が向上することを確認した. 今後は, 音源分離結果から真の特徴量へのベクトルの判定を行い, Soft Mask の自動生成に取り組む予定である.

謝辞 科研費, 21 世紀 COE, HRI-JP の支援を受けた.

参考文献

[1] Takeda, *et al.*: "Improving Speech Recognition of Two Simultaneous Speech Signals by Integrating ICA BSS and Automatic Missing Feature Mask Generation", *Proc. of ICSLP*, pp.2302–2305 , Sep. 2006.
[2] 山本他.: "ミッシングフィーチャ理論を適用した同時発話認識システムの同時発話文による評価", AI チャレンジ研究会 (第 22 回), 101–106, 2005.
[3] 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.
[4] Murata, *et al.*: "An approach to blind source separation based on temporal structure of speech signals", *Neurocomputing*, 41, pp.1–24, 2001.
[5] Raj, *et al.*: "A Bayesian Framework for Spectrographic Mask Estimation for Missing Feature Speech Recognition", Speech Communication, pp.379–393, 2004
[6] Barker, *et al.*: "Soft Decision In Missing Data Techniques for Robust Automatic Speech Recognition", *Proc. of Interspeech*, 2000
[7] 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.
[8] CASA Toolkit: http://www.dcs.shef.ac.uk/ jon/ctk.html