5M-8

ICAによる音源分離とミッシングフィーチャマスクの自動生成による同時発話認識

武田 龍 † 山本 俊一 ‡ 駒谷 和範 ‡ 尾形 哲也 ‡ 奥乃 博 ‡

† 京都大学 工学部情報学科 ‡ 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

将来, 様々な面で人間をサポートするようなヒューマノイドロボットは人間と同等の認識能力を有する必要がある. 通常人は音声を使ってコミュニケーションを行うので, 実環境で音声認識を行う上で複数の話者が同時に話しても聞き分ける機能は不可欠である.

音源分離, 混合音認識といったロボット聴覚機能を実現する上で必要不可欠な条件は, できるだけ特定の環境に特化しない汎用な処理を実現することである. 分離手法の多くはマイクロホンの位置や話者の情報を必要とする. 混合音認識においても, 例えば, マルチコンディション学習が有効であるが, 汎用的に利用できるとはいえない.

本稿では音声の独立性のみを仮定するICAによる音源分離を行う. さらにクリーン音声での学習のみで, 分離による歪みに対応できるミッシングフィーチャ理論 (MFT) を応用した音声認識 [1] を用いる. これにより汎用的な聴覚システムの構築を行う. さらに構築したシステムを2話者の同時発話認識率によって評価する.

2. 基礎となる手法

2.1 ICA を用いた音源分離

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.

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

ここで, $s(t) = [s_1(t),...,s_I(t)]^T$ 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. ここで, $[\cdot]_{ji}$ 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.

周波数領域で ICA を適用するので, 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扱う. これより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. 次に, 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.

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

また最適な分離行列を求めるのに, Kullback-Leibler divergence を最小化するアルゴリズムがしばしば用いられる. 本稿では Choi[2] らが提案した, 音声に対して有効である non-horonomic 拘束適用による以下の反復学習則を用いる.

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

Recogonition of Simultaneous Speech Signals Based on Sound Separation by ICA and Automatic Genaration of Missing Feature Mask : *Ryu Takeda, Shunichi Yamamoto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

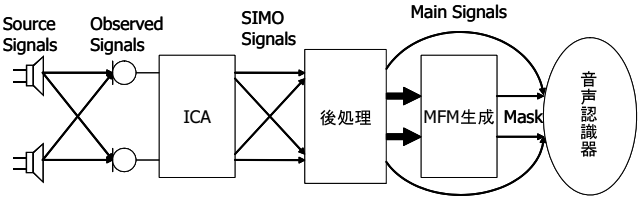

図 1: システムの概要

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.

周波数領域 ICA では, 各周波数ビン毎に分離行列を求めるため, ICA 特有のパワー (振幅) 及び出力信号の順番を決定できないことが知られている. 前者のスケーリング問題は分離行列の逆行列を掛ける手法 [3] で解決した. これにより Single Input Multiple Output (SIMO) モデルの信号が得られることになる. 後者のパーミュテーション問題はスペクトルのエンベロープを利用した手法により対処した [3].

2.2 ミッシングフィーチャ理論を用いた音声認識 [1]

一般に分離音声や入力音声にはクリーン音声に比べて歪みが生じ, 音声認識の特徴量に影響を及ぼし認識率が低下する. この問題に対して MFT による音声認識で対処する. すなわち, 歪んだ特徴のみをマスクすることにより, その影響を排除する.

MFT ベースの音声認識システムでは音声認識の特徴量として, MFCC (Mel Frequency Cepstrum Coefficient) ではなく, スペクトル特徴量 [4] を用いる. MFCC は入力スペクトルに歪みがあると, 特定の周波数領域での歪みであっても, MFCC の全係数に影響を与えロバスト性が低下するためである. MFCC の計算過程のケプストラム平均除去後, 逆コサイン変換を行いスペクトル領域に戻して, 1 次微分値と合わせて 48 次元の特徴量とする.

MFT を用いた音声認識を用いると, スペクトル特徴量における歪みを検出し, いかに適切なマスクを自動的に生成するかが重要となる.

3. 同時発話認識システムの構築

3.1 ICA の出力を音声認識する際の課題

システム構成を図 1 に示す. ICA による分離の出力は, 逆行列を用いてスケーリング問題を解決しているため, 信号 i をマイクロホン j で観測した信号 $f_{ij}(t)$ が出力される. そのため, まず認識対象とする信号を決定しなければならない. また, ICA の分離は完全ではなく出力には他信号やノイズが混入する. そのため, SIMO 信号を利用し, いかに分離やノイズによる歪みを検出し, ミッシングフィーチャーマスク (MFM) を生成するかが課題となる.

3.2 認識音声の選択と MFM の自動生成

2 話者の混合音声の分離では, ICA から 4 つの音声信号が出力される. その中から歪みの少ない信号を選択する方法を述べる. 各話者について得られる 2 つの分離音声に対して, 信号強度差によって話者の相対方向を求めた後, その話者に近い方の分離音声を優勢信号 $F_i^1(k,t)$, そうでない方を劣勢信号 $F_i^2(k,t)$ と判定する. ここで k は特徴量次元, t はフレーム番号を表す. 得られた $F_i^1(k,t)$ を信号 i の特徴量として認識させる.

この時以下の式でこの特徴量に対するミッシングフィーチャマスク M_i を作成する. ただし, $i=1,2$ とする.

$$M_i^1(k,t) = \begin{cases} 1 & |F_m^1(k,t)| < \theta_1 \quad (m \neq i) \\ 0 & otherwise \end{cases} \tag{4}$$

$$M_i^2(k,t) = \begin{cases} 1 & |F_i^1(k,t) - F_i^2(k,t)| < \theta_2 \\ 0 & otherwise \end{cases} \tag{5}$$

$$M_i(k,t) = \begin{cases} 1 & M^1(k,t) = 1 \ or \ M^2(k,t) = 1 \\ 0 & otherwise \end{cases} \tag{6}$$

M^1 は他信号の影響が大きい特徴量は利用しないという考えに基づく. M^2 は F_i^1 と F_i^2 は本来同一の音声であるので, 分離が理想的に行われていればほぼ同じ値になるので, 2 つの差が大きい部分は歪みが生じているという判断に基づいている. 最終的に式 (6) により, この 2 つのマスクを統合する. また, 特徴量の 1 次微分 ($k=25,...48$) に対するマスク $M(k,t)$ は以下のように設定する. なお, θ_1 及び θ_2 は実験的に決定し 1.1, 0.4 とした.

$$M(k,t) = \prod_{j=-2, j\neq i}^{2} M(k-24, t-j) \tag{7}$$

4. 実験

システムの評価を行うためヒューマノイド SIG2 (図 2) の外耳道モデルに埋め込まれた 2 本の無指向性マイクロホンで 2 話者同時認識実験を行った. 信号選択の評価のため, 劣勢信号及び優勢信号を用いた場合の実験を行った. また, 優勢信号と自動生成したマスク, MFM の有効性評価のため, クリーン音声と分離音声を比較して作成した a priori マスクを用いた実験も行った.

4.1 録音条件

録音には上述した SIG2 に設置されたマイクロホンを利用した. 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. このような条件で 2 話者同時発話を録音した. データセットは男性・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. 正面に女性話者, 右側に男性話者とした.

4.2 音声認識

今回はマルチバンド版 Julian[5] を MFT に基づく音声認識システムとして利用した. 音響モデルにはトライフォン (3 状態 4 混合の HMM) を利用し, 言語モデルには有限状態文法を利用して 200 組の孤立単語認識を行った. 利用したトライフォンは, 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.

図 2: SIG2

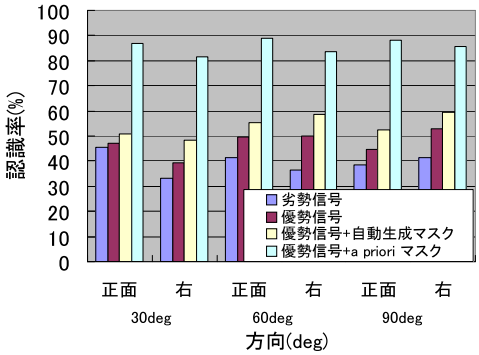

図 3: 同時発話認識結果

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 3 に示す. すべての場合において, 劣勢信号より優勢信号, さらに MFM を用いた方が孤立単語認識率が向上している. 自動生成マスクの方向ごとの平均認識率は 30 度, 60 度, 90 度の順に 50% , 57% , 56% となっている. また, a priori マスクを用いた場合, 飛躍的に認識率が向上していることより MFM の有効性が示された. 対して自動生成したマスクは a priori マスクの性能に及んでいない. また, マスクなしの場合の認識率が低いが, これは単語の組ごとに分離を行っているため, 発話が短い組 (同時発話区間が 2 秒未満) だと十分な分離が難しいなど ICA 自体の問題も関与している.

5. おわりに

汎用性のあるロボット聴覚機能を実現するため, 制約の少ない音源分離と分離音声を認識可能にすることを目指した. 音源分離に ICA を用い, その後段処理として MFM の自動生成を行い, MFT を応用した音声認識器を利用した. これにより認識器の学習にはクリーン音声のみが必要であり, 特定の環境に依存しない認識システムを構成した.

実験では 2 話者同時発話認識を行い, 結果として ICA による不完全な分離及び歪みを抑え, 認識率が向上したことを確認した. しかし, 今回の実験で得られた孤立単語認識率はおよそ 50% 程度であり高いとはいえない. 今後, ICA と MFM の改良を行うとともに, マスク生成と分離の間にフィルタ処理などを挟むことにより認識率の向上を目指していく. また, 孤立単語認識だけではなく連続音声認識への拡張についても検討する予定である.

謝辞 本研究の一部は, 科研費, 21 世紀 COE, SCAT, 電気通信普及財団の支援を受けた.

参考文献

[1] 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), 101-16, 2005.

[2] Seungjin CHOI 他: "Natural Gradient Learning with a Nonholonomic Constraint for Blind Deconvolution of Multiple Channels", *Proc. of International Workshop on ICA and BSS*, 371-376, Jan.1999.

[3] Murata 他: "An approach to blind source separation based on temporal structure of speech signals", *Neurocomputing* **41**(2001)1-24

[4] 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, 117-118, 2004.

[5] http://www.furui.cs.titech.ac.jp/mband_julius/