G-007

ミッシングフィーチャー理論による音源分離と混合音声認識の統合型インターフェース

Missing Feature Theory Based Interface of Integrating Sound Source Separation and Automatic Speech Recognition

山本 俊一† 中臺 一博‡ 辻野 広司‡ 奥乃 博†
Shunichi Yamamoto Kazuhiro Nakadai Hiroshi Tsujino Hiroshi G. Okuno

1. はじめに

近い将来，ロボットは人間のパートナーとして活動するためにソーシャルインタラクションの能力を有することが期待されている．ソーシャルインタラクションを行うために最も重要な機能は音声言語によるコミュニケーションである．そのためには，ロボットは次のような状況に対応できる必要がある．

- 雑音環境において特定の音源に聞き耳をたてることができなければならない．人間のこの能力は『カクテルパーティー効果』として知られている．
- いくつかの同時発話を聞き取ることができなければならない．誰かの音声または何かの音によって会話が中断された場合にも対応できなければならない．これは，音声対話システムでは『バージイン』として知られている．

このように，人間が生活環境の中で聞く音は混合音なので，ロボットは混合音を扱う必要がある．一部のヒューマンロボットコミュニケーションシステムでは話者の口元に取り付けられたマイクロフォンを利用することで，混合音の問題に対処している．しかし，実環境ではロボット自身に設置されたマイクで聞く必要もあろう．

実環境において混合音を扱えるロボット聴覚を実現するためには，次のような課題がある．

1. 混合音の音源分離
2. 分離音声の認識
3. 音源分離と音声認識の間のインターフェース

これら3つの課題は独立に考えることはできない．混合音の音源分離は不良設定問題であるので，完全な音源分離を行うことはできず，分離された音声は歪む．その結果，分離音声が歪むことにより音声認識を行うことが困難となる．従って，音源分離との音声認識の間のインターフェースが重要となる．本稿では，3つの課題に対応するためにミッシングフィーチャー理論を応用してロボット聴覚システムを設計する．

ロボット聴覚システムは特定のロボット専用に設計されていることが多く，他のロボットに対する汎用性についての評価は行われていなかった．我々は3体のロボット，京都大学のSIG2とReplie，HondaのASIMOに提案するロボット聴覚システムを実装して，その汎用性についても評価した．

†京都大学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

2. ロボット聴覚システムにおける課題へのアプローチ

実環境において混合音を扱う場合には，混合音を直接認識することはできないので，混合音の音源分離，分離音声の認識，音源分離と音声認識の間のインターフェースという3つの課題があると述べた．音源分離と音声認識の間のインターフェースは，直列型と統合型の2つに分けることができる．後者は，音源分離と音声認識の両方に手を入れて情報統合により，性能向上を狙う．一方，前者は音源分離を音声認識のフロントエンドとして使い，音声認識側で分離音に対応する．

例えば，中臺らは音源分離と音声認識の間のインターフェースにマルチコンディション学習[1, 2]を適用し，三話者同時発話認識を行うロボット聴覚システムを開発した[3]．マルチコンディション学習は学習データとしてクリーンな音声だけでなく雑音を含んだ音声も利用して音響モデルを学習させる手法である．マルチコンディション学習によって得られた音響モデルは特定の状況において予測される雑音を反映したモデルになっているため，雑音が定常的である場合には効果的な手法である．この手法は現在，環境が限定された自動車や電話へのアプリケーションのための最も一般的な手法となっている．中臺らのインターフェースは方向別・話者別の51個の音響モデルを利用して音声認識を行うため，計算コストがかかる．

本研究では，音源分離と音声認識の間のインターフェースにミッシングフィーチャー理論を応用した統合型を設計した[4]．混合音の音源分離には中臺らのロボット聴覚システムのアクティブ方向通過型フィルタ（ADPF）を利用する．そして，音源分離において歪んだ特徴量をミッシングフィーチャーとして検出し，ミッシングフィーチャー理論に基づく音声認識によって分離音声を認識する．ミッシングフィーチャー理論を適用したことによってによって，分離音声の動的な特徴量の歪みに対応することができ，単一の音響モデルのみで認識することができるようになる．さらに，結果としてマルチコンディション学習に基づくシステムのように複数の音響モデルを必要とせず，単一の音響モデルだけでよいので計算コストが大幅に削減される．

3. ロボット聴覚システム

ミッシングフィーチャー理論に基づく音声認識はロバスト性を向上させるために有効な手法として研究されてきた[5, 6]．この手法では，入力音声から雑音によって歪んだサブバンドをミッシングフィーチャーとして検出する．検出されたミッシングフィーチャーによってシス

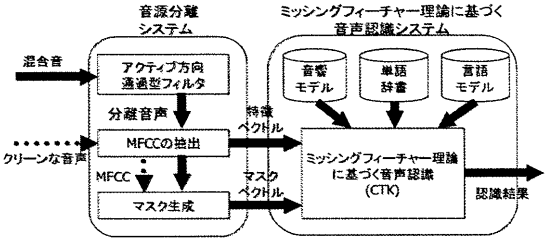


図 1: 複数同時発話認識を行うロボット聴覚システム

テムが影響を受けないように音声認識の際にマスクされる．このため，雑音が動的に大きく変化する場合にも柔軟に対応することができる．本稿では，音源分離と音声認識の間のインターフェースにミッシングフィーチャー理論を適用する．音源分離では，分離音声とターゲットとなる音声を比較することによって検出されたミッシングフィーチャーからミッシングフィーチャーマスクが生成される．音声認識では，生成されたマスクを利用してミッシングフィーチャー理論に基づく音声認識を行う．

3.1 ミッシングフィーチャー理論に基づく音声認識

本稿のミッシングフィーチャー理論に基づく音声認識では，通常の音声認識と同様に MFCC を特徴量とし利用し，隠れマルコフモデル（HMM）に基づく方法を利用する．ミッシングフィーチャー理論に基づく音声認識システムの概要を図 1 に示す．音声認識システムでは，状態遷移確率と出力確率から与えられた信号系列を最も高い確率で出力する状態遷移系列を求める．ミッシングフィーチャーマスクは入力音声から推定され，音声認識の際にその特徴がマスクされる．ミッシングフィーチャー理論に基づく音声認識では，通常の音声認識とは出力確率の計算方法が異なり，次のようになる．

特徴ベクトル $\boldsymbol{x}$，状態 S の時の出力確率 $f(\boldsymbol{x}|S)$ とすると，マスクされたときの出力確率は次の式で求める．

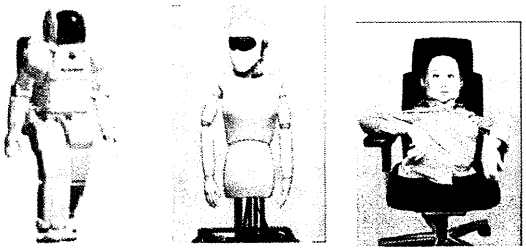
$$f(\boldsymbol{x}_r|S) = \sum_{k=1}^{M} P(k|S) f(\boldsymbol{x}_r|k, S)$$

ここで，M は混合正規分布の混合数，$P(k|S)$ は混合係数，$\boldsymbol{x}_r$ は $\boldsymbol{x}$ のうち信頼できる特徴である．つまり，信頼できる特徴だけを出力確率の計算に用いるので，信頼できない特徴による影響を除去することができる．

3.2 音源分離におけるミッシングフィーチャーマスクの生成

音源分離には ADPF を利用する．ADPF は 2 つのマイクロフォンを利用して特定方向から来る音を抽出するという手法である [7]．ADPF はまず左右のマイクロフォンで録音された音のパワースペクトルのそれぞれのサブバンドから両耳間位相差（IPD）と両耳間位相差（IID）を計算する．次に，散乱理論によってそれぞれの方向の IPD と IID を推定する．そして，通過幅の範囲に収まるような IPD と IID を持つサブバンドが集められ，逆 FFT によって音が再合成され，分離音声として扱われる．

ミッシングフィーチャーの推定手法はいくつか報告されており，そのうちの一つの方法に，分離音声の特徴量とそれに対応する元のクリーンな音声の特徴量を比較することでミッシングフィーチャーマスクを生成する手法がある．これは，認識する音声の元のクリーンな音声が利用されるので，『**演繹的なマスク**』と呼ばれる [8]．本稿では，ミッシングフィーチャー理論の有効性を検証するためにミッシングフィーチャーマスクについては一番容易で効果的な演繹的なマスクを使用する．



(a) ASIMO (Honda)　(b) SIG2 (京都大学)　(c) Replie (京都大学)

図 2: Humanoid Robots

ミッシングフィーチャーマスクは MFCC の特徴ベクトルと同じ次元数のベクトルで，フレーム毎に存在する．マスクベクトルのそれぞれの成分の値は対応する MFCC の特徴量の信頼度を表している．この信頼度は 0 から 1 までの連続量を用いることもできるが，本稿では信頼できる（1），信頼できない（0）という 2 値の信頼度を利用する．

ミッシングフィーチャーマスクの生成の詳細なアルゴリズムを以下に述べる．

1. X をロボットのマイクで録音した音声から分離された音声の特徴量，Y をそれに対応する元のクリーンな音声の特徴量とする．特徴量は，MFCC12 次元，ΔMFCC12 次元，ΔPower の合計 25 次元を用いる．

2. $M_k(i)$ を k 番目のフレームの i 番目の特徴量に対するマスクとすると，

$$M_k(i) = \begin{cases} 1 & if\ |X_k(i) - Y_k(i)| < T \\ 0 & otherwise \end{cases}$$

ここで，T は実験的に求めた閾値である．

3. 特徴量の一次微分に対するマスクは次式で求める．

$$\Delta M_k(i) = M_{k-2}(i) M_{k-1}(i) M_{k+1}(i) M_{k+2}(i)$$

4. 評価

音源分離と音声認識の間の新しいインターフェースの導入の効果を評価するために，三話者同時発話認識を行う．システムの汎用性を確認するために，ASIMO, SIG2, Replie という 3 体のヒューマノイドロボットを実験に用いた．ASIMO, SIG2, Replie をそれぞれ図 2(a) から (c) に示す．

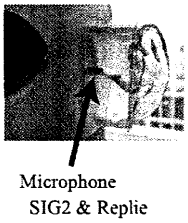


図 3: ヒューマノイドロボットの耳介

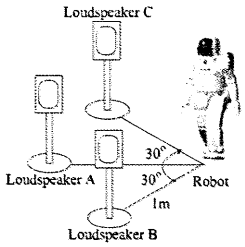


図 4: 実験の概要

4.1 実験に用いたヒューマノイドロボット

SIG2 と Replie は共に表面がシリコンで覆われており，音の反射をある程度防いでいる．マイクロフォンは人間から型をとったシリコン製の耳介の外耳道に取り付けられており，人間の耳介と同じ位置に設置されている (図 3)．この耳介は正面方向の音を 10 dB 強める効果がある．これらの 2 つのヒューマノイドロボットは外見が異なっており，SIG2 はプロのデザイナーが外見の美しさを考慮してデザインしたもので，Replie の外見は日本人の女性から型をとっている．このように，これらの 2 つのヒューマノイドロボットの音響的特性には相違点があるが，多くの部分は類似した特徴を持っている．

一方，ASIMO の表面は硬い材質で覆われ，頭部の形状は少し角ばっていて，2 つのマイクロフォンの位置も SIG2 と Replie とは異なった位置に設置されている．このため，ASIMO の音響特性は SIG2 と Replie とは大きく異なっている．

4.2 音響モデル

今回の実験では，分離音声を認識するための音響モデルは，方向や話者毎に用意するのではなく**単一の音響モデルのみ**を利用する．学習データは無響室で録音されたクリーンな音声で，合計 25 人の男女の音声で日本語の ATR 音素バランス単語 216 語である．特徴量は 25 次元で，MFCC12 次元，ΔMFCC12 次元，ΔPower である．音響モデルは，3 状態 8 混合の HMM でモノフォンとトライフォンである．

4.3 実験

三話者同時発話の孤立単語認識率によって本システムを評価した．実験の概略図を図 4 に示す．音源として 3 つのスピーカーをロボットから 1m の距離でそれぞれ 0°，±30° の方向に設置した．ASIMO の実験環境は 7.5 m×9 m の大きさで残響時間 (RT_{20}) は約 0.5 秒の部屋で，SIG2 と Replie の実験環境は 4 m×5 m の大きさで残響時間 (RT_{20}) は 0.3〜0.4 秒の部屋である．スピーカーから再生される音声は，学習データに用いた ATR 音素バランス単語 216 語のうちから互いに異なる 3 つの単語の組み合わせ 200 種類である．スピーカからのこれらの 3 つの単語の組み合わせの音声と部屋の雑音の混合音を各ヒューマノイドロボットの耳で録音した．

そして，このような三話者同時発話の孤立単語認識において，さまざまなパラメータを変えて実験を行った．変化させたパラメータは以下のようなものである．

- 語彙数：10，50，100，200 語彙
- 音響モデル：モノフォン，トライフォン
- ミッシングフィーチャーマスク：利用する，しない
- ロボット：ASIMO，SIG2，Replie

3 体のヒューマノイドロボットにおける実験結果を図 5，図 6 に示す．ASIMO，SIG2，Replie でミッシングフィーチャーマスクを利用しない場合の孤立単語認識率をそれぞれ図 5(a)，(b)，(c) に，ミッシングフィーチャーマスクを利用する場合の孤立単語認識率をそれぞれ図 6(a)，(b)，(c) に示す．ミッシングフィーチャーマスクを利用しない場合は，すべての特徴量が信頼するものとして認識が行われ，これは分離音声をそのまま通常の音声認識によって認識することと等価である．それぞれのグラフは，音源方向が左（30°），中央（0°），右（−30°）の方向で，モノフォンとトライフォンの音響モデルを利用した場合の認識結果を表している．グラフの横軸は語彙数，縦軸は単語認識率である．

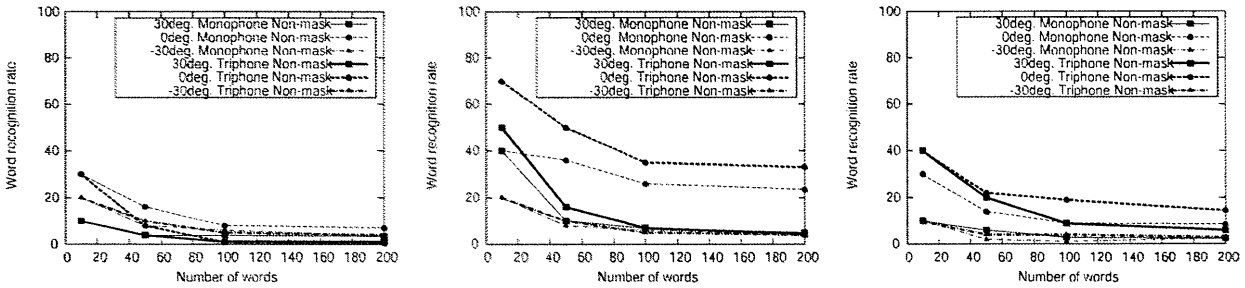
全体的に語彙数が増加するにつれて単語認識率は低下しており，ミッシングフィーチャーマスクを利用しない場合にこの傾向が特に顕著である．一方，ミッシングフィーチャーマスクを利用した場合の結果を見ると，語彙数の増加に対してロバストであることがわかる．ミッシングフィーチャーマスクがありの場合となしの場合を比較すると，マスクありの方が認識率が良く，左と中央の方向ではトライフォンで語彙数 200 の場合で約 80%にまで認識率が向上した．これは，提案手法がロボットにおける同時発話認識に有効であることを示している．しかし，右方向については全体的に認識率が低くなっている．これは，3 方向の同時発話が男性 2 人女性 1 人の音声の混合音であるので，女性の音声のピッチが男性の約 2 倍程度となることがあり，音源分離が困難となることが原因であると考えられる．

それぞれの結果において，モノフォンとトライフォンを利用した場合を比較すると，モノフォンよりもトライフォンの方が約 10%以上認識率が高くなっている．これは，ロボット聴覚のためのミッシングフィーチャー理論に基づく音声認識においてもトライフォンが有効であるあることを示している．これは，話者依存・方向依存の音響モデルによって得られる複数の認識結果を統合する手法 [7] による認識率を上回っている．このように，処理速度だけでなく認識精度においても提案するロボット聴覚システムの有効性が示された．

図 6 を見ると ASIMO，SIG2，Replie の 3 体のロボットすべてにおいてミッシングフィーチャー理論に基づく音声認識が有効に動作していることがわかる．聴覚処理は小さな環境の変化に対して敏感であることを考慮すると，これらのロボットの音響特性の違いにより認識精度も変化することが予想されるが，認識精度はそれぞれのロボットで同等の結果が得られている．これは，提案するインターフェースを適用したロボット聴覚システムが汎用的なシステムであることを示している．

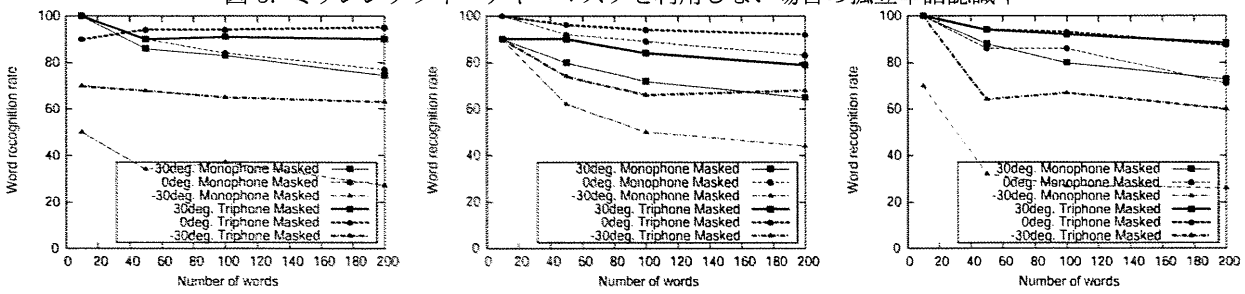
5. おわりに

ロボット聴覚は人間と同じ生活環境の中で活動する知的ロボットを実現するためには非常に重要な技術であ



(a) ASIMO (b) SIG2 (c) Replie

図 5: ミッシングフィーチャーマスクを利用しない場合の孤立単語認識率



(a) ASIMO (b) SIG2 (c) Replie

図 6: ミッシングフィーチャーマスクを利用した場合の孤立単語認識率

る．本稿では，そのようなロボット聴覚を実現するために音源分離と音声認識の間のミッシングフィーチャー理論に基づくインターフェースを提案した．さらに，本インターフェースを利用したロボット聴覚システムを実際にヒューマノイドロボットに実装して評価を行った．実験データには実験室において実際に録音された音声を利用し，三話者同時発話の認識によって提案するインターフェースに基づくロボット聴覚システムを評価した結果，高い認識精度が得られた．また，本システムはクリーンな音声で学習した単一の音響モデルしか使用していないので，可搬性が高くなっている．実際，3 体のヒューマノイドロボットを用いた実験によって本システムが汎用性のある手法であることを確認した．実験に利用した 3 体のヒューマノイドロボットは，そのうち 2 体が音響的特性が類似したものであるが，もう 1 体はこれらとは大きく異なる音響的特性をもったものである．今後の課題として，より一般的な状況で音声認識が行えるようにするために，動的なミッシングフィーチャーマスク生成手法の開発が挙げられる．

謝辞

本研究の一部は，科学研究費補助金（基盤研究（A），特定領域「情報学」），および，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」の支援を受けた．

参考文献

[1] R. P. Lippmann, E. A. Martin, and D. B. Paul. Multi-styletraining for robust isolated-word speech recognition. In *Proc. of ICASSP-87*, pp.705–708. IEEE, 1987.

[2] M. Blanchet, J. Boudy, and P. Lockwood. Environmentadaptation for speech recognition in noise. In *Proc. of EUSIPCO-92*, pp.391–394, 1992.

[3] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano. Applying scattering theory to robot audition system: Robust sound source localization and extraction. In *Proc. of IROS-2003*, pp.1157–1162. IEEE, 2003.

[4] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H. G. Okuno. Improvement of robot audition by interfacing sound source separation and automatic speech recognition with missing feature theory. In *Proc. of ICRA 2004)*, pp.1517–1523. IEEE, 2004.

[5] J. Barker, M. Cooke, and P. Green. Robust asr based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. In *Proc. of Eurospeech-2001*, pp.213–216. ESCA, 2001.

[6] P. Renevey, R. Vetter, and J. Kraus. Robust speech recognition using missing feature theory and vector quantization. In *Proc. of Eurospeech-2001*, pp.1107–1110. ESCA, 2001.

[7] K. Nakadai, H. G. Okuno, and H. Kitano. Robot recognizes three simultaneous speech by active audition. In *Proc. of ICRA-2003*, pp.398–403. IEEE, 2003.

[8] K. Palomaki, G.J. Brown, and J. Barker. Missing data speech recognition in reverberant conditions. In *Proc. of ICASSP-2002*, pp.65–68. IEEE, 2002.