ミッシングフィーチャー理論による三話者同時発話認識の向上

山本 俊一[†] 中臺 一博[‡] 辻野 広司[‡] 駒谷 和範[†] 尾形 哲也[†] 奥乃 博[†]
[†] 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

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．これを実現するためには，ロボット自身のマイクで雑音環境下で特定の音源を聞き分ける能力や，複数の音声を同時に聞き取る能力が必要とされる．しかし，従来の音声認識手法ではマイクは話者の口元にあり単一話者を仮定しているものが一般的である．

本稿では，対象以外の音声が雑音となる同時発話の認識に対してミッシングフィーチャー理論を適用し，ロボットのためのよりロバストな音声認識の手法を提案する．また，複数のロボットを利用した実験により提案するシステムの汎用性についても評価する．

2. ミッシングフィーチャー理論

ミッシングフィーチャー理論に基づく音声認識[1)2)] の処理の概要を**図 1** に示す．この手法では，特徴量のうち雑音によって歪んだ成分がミッシングフィーチャーとして入力音声から検出され，ミッシングフィーチャーをマスクすることで認識に悪い影響を及ぼさないようにする．この結果，マルチコンディション学習と比較してノイズが多様に変化する場合にも柔軟に対応することができると期待される．これまで，こうした研究では，AURORA に代表されるように予め録音した非音声雑音を計算機上で重畳させた音声を認識するものが多く，認識対象音声も英数字に限るなど語彙数が少なかった．本稿では，ロボットで実際に収音した日本語の孤立単語音声を対象とし，同時発話により音声が雑音となる場合を扱った．また，特徴量は MFCC（メル周波数ケプストラム係数）で音響モデルには音素 HMM （隠れマルコフモデル）を利用した．

2.1 ミッシングフィーチャー理論に基づく音声認識

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．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ようになる．

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

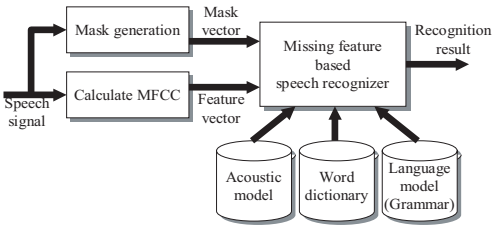


図 1 ミッシングフィーチャー理論に基づく音声認識

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．これは，信頼できる特徴だけが出力確率の計算に用いられるということであり，信頼できない特徴による影響を除去することができる．

2.2 ミッシングフィーチャーマスクの生成

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．本稿では信頼できる（1），信頼できない（0）という 2 値のマスクを利用する．

ロボットのマイクで録音された音声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によってミッシングフィーチャーマスクを生成する．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**[3)]』と呼ばれる．ミッシングフィーチャーマスク生成の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の特徴量，Y をそれに対応する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 番目の特徴量とすると，

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

ここで，T は実験的に求めた閾値である．

(3) マスクのデルタ成分については以下のように求める．

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

3. 三話者同時発話認識

2 節で述べたミッシングフィーチャー理論に基づく音声認識をロボット聴覚システムに適用して三話者同時発話認識を行う．

3.1 音源分離

音源分離には，2 つのマイクを利用して特定方向の音源を抽出する，アクティブ方向通過型フィルタ（ADPF）[4)] を利用する．これは，散乱理論によって推定した取得したい音源方向の両耳間位相差 (IPD) と両耳間強度差 (IID) 情報と，入力音声の各周波数サブバンドごとの IPD と IID を比較し，マッチしたサブバンドのみを通過させるフィルタである．ま

Improvement of Three Simultanious Speech Recognition by Missing Feature Theory
by Shun'ichi Yamamoto (Kyoto Univ.), Kazuhiro Nakadai, Hiroshi Tsujino (HRI-JP), Kazunori Komatani, Tetsuya Ogata and Hiroshi G. Okuno (Kyoto Univ.)

図 2 SIG2

図 3 Replie

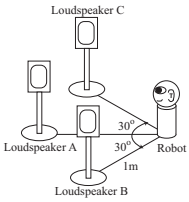


図 4 実験の概要

た，マッチしたサブバンドを集めて，逆 FFT を行えば，特定方向の音源情報を再合成することもできる．

3.2 音声認識

音声認識には，ミッシングフィーチャー理論を利用することができる CASA Tool Kit（CTK）[1] を利用した．ADPF によって抽出された音声を CTK によって認識させる．認識させる音声は，孤立単語の三話者同時発話である．

4. 評　　価

提案手法によって三話者同時発話認識が向上することを確認する実験を行った．実験ではシステムの汎用性を確認するために SIG2 と Replie という 2 つの異なるヒューマノイドロボットを実験に用いた．

4.1 実験に用いたヒューマノイドロボット

実験に用いたヒューマノイドロボット SIG2 と Replie を**図 2**，**図 3** に示す．これらのヒューマノイドロボット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．マイクは，人間の耳介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 1 つの音響モデルが共通に用いられる．音響モデルは HMM に基づくモデルで，無響室で録音されたクリーンな音声で学習させる．学習データは 25 人の男女の音声で日本語の音素バランス単語 216 語である．特徴量は 25 次元で，MFCC12 次元，ΔMFCC12 次元，ΔPower である．また，HMM は 3 状態 8 混合のモノフォンとトライフォンである．

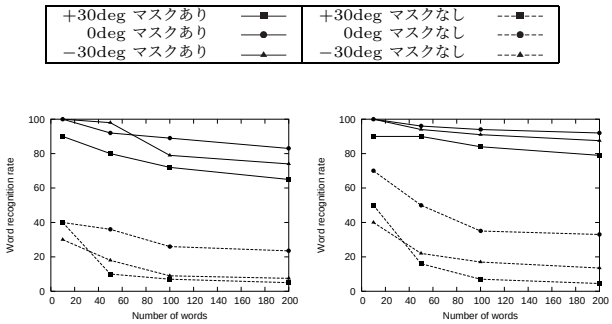
4.3 実　　験

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．実験を行った部屋は 4m×5m の大きさで，残響時間（RT_{20}）は 0.3〜0.4 秒である．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である．

このような三話者同時発話の孤立単語認識で，さまざまなパラメータを変えて実験を行った．

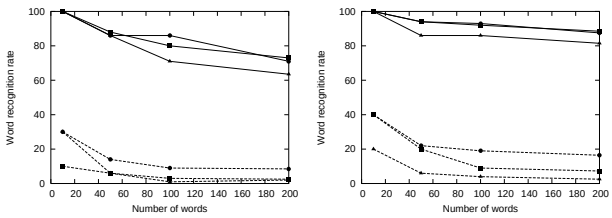
- 語彙数：10，50，100，200 語彙
- 音響モデル：モノフォン，トライフォン
- ミッシングフィーチャーマスク：利用する，しない
- ロボット：SIG2，Replie

SIG2，Replie での単語認識率を**図 5**，**図 6** に示す．それ



a) モノフォン　　b) トライフォン

図 5 SIG2 での単語認識率



a) モノフォン　　b) トライフォン

図 6 Replie での単語認識率

ぞれの図は，音源方向が左（30°），中央（0°），右（−30°）の方向の場合の認識結果を表しており，横軸が語彙数，縦軸が単語認識率である．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，トライフォンで語彙数 200 の場合で 80%以上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．

5. おわりに

本稿では，三話者同時発話の音源分離を行いミッシングフィーチャー理論に基づく音声認識を行う方法について述べた．実際にロボットで収音された音声を利用し，音源分離によって特徴が壊れた音声に対して提案手法が有効であることを確認した．さらに，複数の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．今後の課題として，演繹的なマスクではないミッシングフィーチャーマスクの生成手法の検討が挙げられる．

なお，本研究の一部は，科研費，21 世紀 COE の支援を受けた．

参 考 文 献

1) Barker, J., M.Cooke and P.Green: Robust ASR Based on Clean Speech Models: An Evaluation of Missing Data Techniques for Connected Digit Recognition in Noise, *Proc. of 7th EUROSPEECH-01*, Vol. 1, pp. 213–216.
2) Renevey, P., Vetter, R. and Kraus, J.: Robust Speech Recognition using Missing Feature Theory and Vector Quantization, *Proc. of 7th EUROSPEECH-2001*, Vol. 2, pp. 1107–1110.
3) Palomaki, K., Brown, G. and Barker, J.: Missing Data Speech Recognition In Reverberant Conditions, *Proc. of ICASSP-2002*, Vol. I, pp. 65–68.
4) Nakadai, K., Hidai, K., Okuno, H. G. and Kitano, H.: Real-Time Speaker Localization and Speech Separation by Audio-Visual Integration, *Proc. of IEEE ICRA2002*, pp. 1043–1049.