ソフトマスクと音響モデル適応を用いた3話者同時発話音声認識

高橋 徹† 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 ‡(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

近年, ロボットの役割は, 産業用から社会用へ変化している. 社会用ロボットは, 環境に応じた適切な対応を期待され, 産業用ロボットが, 環境に依存せず, 一定の動作を正確に繰り返す役割を担っている点と異っている. 例えば, HONDA の ASIMO [1] は, 人間の生活空間で活動することを想定して開発された社会用ロボットであり, SONY の AIBO [2] は, 人を楽ませる事に重点を置かれたロボットである.

ロボットが, 人を楽ませたり, 人の役に立つためには, 周囲の人を理解する必要があり, さらに人間とのコミュニケーション能力が不可欠である. 坊農ら [3] は, ロボットと人間の1対1対話に加え, 1対多のコミュニケーションや, 人間同士のコミュニケーションが円滑に進むよう支援するロボットの重要性を指摘している.

このようなコミュニケーション能力を実現するために, 複数話者の同時発話音声認識が必要である. 多人数でのコミュニケーションでは, 特定の話者のみを認識対象にすると, 会話の流れを把握できない. 一般に, 同時刻に発話する人を1名に限定できないため, すべての発話を認識対象にする必要がある. 人間の同時発話認識能力は, 2, 3名であるとされている [4] ため, 本研究では, 従来研究に比べより安定な3名による同時発話認識技術を開発する.

著者らはこれまで3話者同時発話認識を行ってきた [5, 6]. 3話者同時発話を音源分離し, 分離歪を推定し, 音響特徴量の歪量に応じた重みを用いて音響尤度最大化基準で音声認識を行った. 音源分離によって個々の発話に分離すると, 1話者の認識問題に帰着する. しかし, 分離音声は, それぞれの音響特徴量が互いに干渉し相い, 音響特徴量が歪むため1話者の認識精度と比較して精度が低下する問題がある. 歪を完全に除去することが困難であるため, 歪を考慮した認識が必要である.

著者らは, MFT (Missing Feature Theory) に基き, 音響特徴量の歪量に応じた重みを用いて音響尤度を求める方法を用いている. 歪の影響を含む信頼できない特徴量をマスクして認識する手法である. このマスクは, マスク値が0から1の連続値をとることから「ソフトマスク」と呼ばれる. ソフトマスクを用いることによって0または1をとるハードマスクを用いる場合と比較して認識精度を5% 改善した [6].

ソフトマスクを用いる利点は, 音響モデルと音響特徴量に改変を加えない点であるが, 本稿では, 音響モデルを分離歪に適応した場合にもソフトマスク処理により認識精度が更に向上することを示す. MFT に基く音声認識は, 分離歪のある音響特徴値がクリーンな音響モデルとミスマッチするため, 歪のある部分をマスクすることで認識精度を高める手法である. 音響モデルを分離歪に適応すると, モデルの平均値も歪を含んだ形状になるため, 歪んだ特徴量にマッチするため, わざわざマスクする必要がないように思われる. しかし, 実際には, 分離歪への適応とマスク処理の組み合わせによって認識精度が改善することが実験により確認できた. 考察で, なぜ精度が改善されるかの仕組を述べる.

Simultaneous three talker speech recognition using soft mask and model adaptation technique : *Toru Takahashi (Kyoto Univ.), Kazuhiro Nakadai (HRI-JP), Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

2. 3話者同時発話音声認識

3話者同時発話音声認識システムが音声を認識するにはまず, マイクロフォンアレーを用いて音源定位と音源分離を行う. 次に分離音を HMM(Hidden Markov Model) で認識する. 音源分離による歪の影響で音響特徴量と音響モデルが乖離し, 認識精度が低下する. 分離歪に対処する必要がある.

分離歪に対処する方法の1つに MFT に基く音声認識がある. MFT に基く認識は, ソフトマスクを用いた認識, つまり分離音の音響特徴量に信頼度を振り, 信頼度を考慮した音声認識を行う方法である. 時刻 t における特徴量 $x(t)$ のある状態の尤度は, マスク値 $m(t)$, モデルパラメタ θ, 混合数 L とすると,

$$\sum_{l=1}^{L} m(t) \log P(x(t)|\theta) \qquad (1)$$

と表される. マスク値が常時1の時, 通常の HMM による認識と等価になる. よく知られた他の方法には, モデルを歪に適応する方法がある. 分離音が得られる場合, 分離音に音響モデルを適応し, 適応モデルで認識する. 音響モデル適応と MFT の併用も可能である.

分離歪は, 3話者の位置関係によって異なる. 3話者が近くにいる場合は, 発話間の干渉が大きく, 分離も困難になるため分離歪も大きい. そこで音響モデルを分離歪に適応するときに, 3話者の位置ごとに分離音を用い音響モデルの適応を行うと, 適応による認識精度の上限を与えることができると期待される.

本稿では, MLLR (Maximum Likelihood Linear Regression) [7] を用い教師あり話者適応と教師あり分離歪適応した音響モデルを用意し, ソフトマスク処理との任意の組み合わせで精度を比較した. ベースとなる音響モデルを用い, 認識対象となる話者適応する場合と, 認識対象の話者の発話を実際に音源分離音した音に適応する場合を比較した. 実際には, 「適応なし」, 「話者適応」, 「話者+分離歪適応」の3つのモデルを比較した.

3. ソフトマスク処理と音響モデル適応の併用

3.1 実験条件

音響モデルは, 3状態 Left-to-right モデルのトライフォン HMM で, 各状態は4混合ガウス混合分布で表し, 総状態数を約2000とした. 音響モデルの学習には JNAS の PB 中のヘッドセットデータの全ての発話を用いた. 音声波形のサンプリング周波数は16 kHz, 量子化ビット数16ビット, 線型量子化である. 1発話当りの SNR が 30dB となるようガウス雑音を加え学習用データとした. 音響特徴量は, MSLS (mel-scale logarithmic spectrum) [5] で, 24次元 (静的特徴量 12 次元, と動的特徴量 12 次元) の特徴ベクトルを用い, 特徴量の平均除去を行っている. 分析フレーム長は, 25 ms, 分析フレーム周期は, 10 ms である. この音響モデルをベースモデル (BM) として3話者同時発話認識の精度を比較する. 表1に比較条件を示す. 音響モデルの適応データの種類とソフトマスク処理の有無により6条件に分けた. 音響モデル適応に用いた音声は, ATR 音声データベース音素バランス単語中の100単

表 1: 比較条件.

BM	ベースモデルを用いて認識
BM+SM	ベースモデルを用い, ソフトマスク処理を併用して認識
MLLRC	ベースモデルを話者適応させたモデルを用いて認識
MLLR	ベースモデルを話者と分離歪に適応させたモデルを用いて認識
MLLRC+SM	ベースモデルを話者適応 (MLLR) させたモデルを用い, ソフトマスク処理を併用して認識
MLLR+SM	ベースモデルを話者と分離音に適応させたモデルを用い, ソフトマスク処理を併用して認識

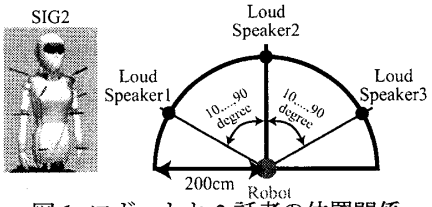

図 1: ロボットと 3 話者の位置関係.

語である. クリーンな音声 (データベースから求めた特徴量) で適応 (MLLRC) したモデルと, 実際に音源分離した音で適応したモデル (MLLR) を構築した. ロボットの正面方向, 左右からにそれぞれ m101, f101, m102 の 3 名の音声を再生し, ロボット (SIG2) 搭載の 8 チャンネルマイクロフォンアレーで音声を収録した. 収録した音声は, GSS (Geometric Source Separation) [8] により分離した. ロボットと 3 名の音声を再生したラウドスピーカの位置関係を図 1 に示す. ロボットとラウドスピーカー間の距離は常に 200cm である. 話者間の距離を角度で表し, 10 度から 90 度まで 10 度間隔で評価した. 評価用の音声は, ATR 音声データベース音素バランス単語中の 100 単語である. ただし, 音響モデル適応に用いた音声とは重複しない. ソフトマスクは, 文献 [6] に従って作成した.

3.2 実験結果と考察

図 2 に単語正解精度を角度条件毎に示す.

BM と BM+SM を比較すると BM+SM が常に高い精度を示した. 話者間角度が 50 度以下の条件では 10 % 以上精度改善が得られた. MLLRC と MLLRC+SM や MLLR と MLLR+SM を比較しても, 概して +SM の精度が高く, ソフトマスク処理による精度改善を確認できる.

BM+SM と MLLRC を比較すると, 話者間角度が 40 度以下では, クリーンな音声で話者適応するより, 適応せずに MFT に基き認識する方が高い精度である. 50 度以上では, 個別に音響モデルを適応することで認識精度改善効果が表れるが, 40 度以下では適応せずソフトマスク処理した方がよい.

MLLRC と MLLR を比較すると, 話者と分離歪に同時に適応する方が高い認識精度を示している. この二つの方法の差は, 歪に適応がない場合とある場合を比較していることになる. BM と MLLRC を比較すると話者適応なしとありを比較していることになる. 話者間角度が狭い時には, MLLR-MLLRC(MLLR と MLLRC の差) と MLLRC-BM(MLLRC と BM の差) では, MLLR-MLLRC が BM-MLLRC より大きく, 話者間角度が広い時には, 逆になる. 分離の困難な話者間角度が狭い時に, 分離歪への適応が支配的で, 分離が比較的容易な時には, 話者適応が支配的になっている.

MLLRC+SM と MLLR+SM を比較すると, 50 度以下では, 話者と分離歪に同時に適応する方が高い認識精度を示す. 60 度以上では, 同程度の認識精度を示している. これは, 60 度以上では, 3 話者 (Loud Speaker 1,2,3) が十分に離れているため音源分離が容易になり, 分離歪が減少することで, 分離音での適応が近似的に話者適応になったためと考えられる.

MLLR+SM が MLLR と比較して認識精度が高い理由

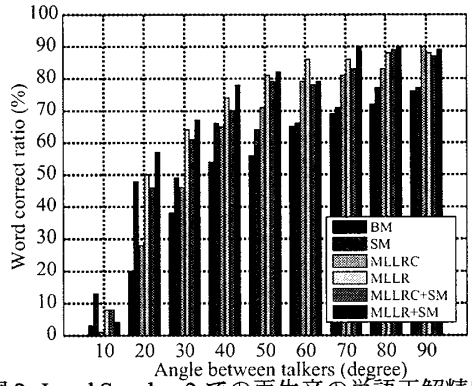

図 2: Loud Speaker 2 での再生音の単語正解精度.

は, 以下のように考えられる. 分離歪が大きいと適応によって分布の平均が大きく移動する. しかし, 適応後の平均付近の音響特徴値の信頼度は, 適応前の平均から離れているため低くなり, 尤度は高いものの, 認識の尤度計算に対する寄与が下がる. 一方, 適応前の平均付近の特徴値の信頼度は高いが, 適応後のモデルで測る尤度は低くなり, 認識の尤度計算に対する寄与が下がる. つまり, 歪量が大きい分布に対応する特徴値は, 常に尤度計算に対する寄与が低い状態になる. 歪の小さい次元の特徴量は, 適応前後で分布の平均が殆ど移動せず, 信頼度の高い特徴値は尤度も高いままで, 信頼度の低い特徴値は尤度が低い通常の対応関係が維持される. 最終的に, 歪の小さい次元の特徴の寄与が高くなる. 以上のメカニズムにより, 音響モデルを分離歪に適応したモデルを用いて MFT に基く認識を行うと精度が改善すると考えられる.

4. まとめ

分離歪に適応させた音響モデルにソフトマスクを適応すると認識精度を改善することを確認した. 音響モデルを分離歪に適応し, マスク処理と併用しても信頼度が適切に反映される仕組を考察した. 話者適応による改善も確認できた. 音響モデルを話者適応と分離歪とに適応すると話者間が中程度 (40〜60 度) の時に精度改善が大きく, ソフトマスクで更に精度が改善される結果となった. 話者間角度が 50 度以上あれば約 80 % 以上の比較的高い精度を達成可能なことを示唆できた. 音響モデル適応とソフトマスク処理により複数話者のコミュニケーションにおいて人間とロボットの立ち位置の自由度を高めることができた. 今後の課題は, 教師なし適応の性能評価などがある.

謝辞

本研究は, 科研費基盤研究 (S), および京都大学グローバル COE プログラムの支援を受けた.

参考文献

[1] *http://www.honda.co.jp/ASIMO/about/*
[2] *http://www.sony.jp/products/Consumer/aibo/*
[3] 坊農 真弓, 高梨 克也:"チュートリアル「多人数インタラクションの分析手法」連載開始にあたって", 人工知能学会誌, 22 巻 5 号, 2007.
[4] 川島 尊之, 佐藤 隆夫:"同時複数音声の分散的聴取における知覚限界", 日本音響学会誌, 65 巻 1 号, pp.3–14, 2009.
[5] S.Yamamoto, et al., "Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory", *Proc. ICRA 2005*, pp.1489–1494, 2005.
[6] T.Takahashi, et al.,"Soft Missing-Feature Mask Generation for Simultaneous Speech Recognition System in Robots", *Proc. Interspeech 2008*, pp.992–995, 2008.
[7] C.J.Leggetter and P.C.Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models", *Computer Speech and Language*, vol.9, pp.171–185, 1995.
[8] S.Yamamoto, et al.,"Making a robot recognize three simultaneous sentences in real-time", *IROS 2005*, pp.892–897, 2005.