4T-3

ロボットとの音声対話における発話の重なりを含む入力音の判別

杉山 貴昭 † 駒谷 和範 † 佐藤 理史 †
† 名古屋大学大学院 工学研究科 電子情報システム専攻

1 はじめに

ヒューマノイドロボット（以下，ロボット）との音声インタラクションで問題となるのは，周辺雑音が引き起こすロボットの誤り応答である．ロボットは様々な入力音からユーザの発話を判別し，応答しなければならない．我々は，Aldebaran Robotics 社製の NAO ‡を使ったシステムを構築している．ここでは，(1) マイクとスピーカがともに頭部に搭載されており，近接している，(2) 音を入出力する端子の増設は困難，というハードウェア的制約がある．このため特にユーザとロボットの発話が重なった時は，ロボットには自身に搭載されたマイクから発生する音声も入力されることになり，ユーザの発話の存在を認識することが困難である．この問題の解決法のひとつに，セミブラインド音源分離を用いる手法がある [1]．この手法は，混合音から得たい音のみを分離できる．ただし，混合音とロボットの発話を完全にで同期させる必要がある．端子の増設が難しく，かつ処理に遅延のあるロボットで，完全な同期を得るのは困難である．

そこで我々は，入力音の音響的特徴に基づき，これらを判別する手法を用いる．李らは，ユーザの発話や雑音の種類ごとに Gaussian Mixture Model（以下，GMM）を作成し，これらを判別した [2]．本稿では，この GMM に基づく判別手法を，ユーザとロボットの発話が重なった場合に適用し，その有効性を確かめる．

本研究ではまず，4 クラスの音をそれぞれ収集する．次に，それぞれの音の音響的特徴から各クラスの GMM の作成・評価を行う．その後，ロボットへの入力音を GMM によりオンラインで判別するシステムの構築を行う．本稿ではこのうち，作成した GMM の評価までを報告する．

2 入力音の判別先とする 4 つのクラス

我々はユーザの発話，雑音，ロボットの発話，発話の重なりの 4 クラスに分類する．これにより，ロボットはユーザとロボットが同時に発話した場合でも，ユーザの発話の存在を認識できるため，ユーザに聞き直すことができる．例えば，ロボットの発話中にユーザが「もう少し大きい声で話してくれませんか？」と発話した場合を考える．この時，ロボットのスピーカには自発話とユーザの発話が重なった音が入力されるため，ユーザの発話内容を認識することは困難である．このとき，入力音をロボットとユーザの発話が重なった音だと判別できれば，ロボットは発話を停止し，ユーザに「もう一度言い直してくれませんか？」と聞き直すことができる．このようにロボットが聞き直せれば，対話を円滑に行える．

分類するクラスを表 1 に示す．実環境でユーザとロボットが対話する場合，ロボットに入力されうる音は，ユーザ発話（user），ロボット自身の発話（robot），雑音（noise）である．これに加えて，我々はロボットとユーザの発話の重なり（user+robot）も判別対象とする．なぜならユーザとロボットの発話が重なった場合，ロボットのマイクとスピーカが近い位置に搭載されているため，ユーザが発話していることをロボットが認識するのは難しいからである．ロボットが robot と user+robot を判別できれば，ユーザの発話の存在を認識できる．また，その他の複数の音の重なりでもクラスを定義できるが，実環境でロボットがユーザ発話に応答するには，他の組み合わせのクラスを定義しなくても判別できると考えたため定義しない．例えば，ユーザの発話中に雑音が発生した場合，ユーザの発話と雑音が重なってロボットに入力される．この場合，ユーザ発話のほうが音源のパワーが大きいため，ロボットがユーザ発話の存在を認識することは容易であると考える．

表 1: クラスごとに収集した音と分離音の合計時間

クラス	収録した音 [秒]	分離音 [秒]
user	9831	4037.2
noise	240	174.8
robot	2790	2361.9
user+robot	1680	424.7

本研究では，簡単のため，ユーザは 1 名であり，ユーザ発話はすべてロボットに向けられたものであると仮定する．ユーザが複数人存在する場合，ユーザの発話が誰に向けた発話であるかを考慮する必要がある．例えば，2 名のユーザと 1 体のロボットの会話の場合，ロボットは，ユーザの発話が自分に向けた発話か，別のユーザに向けた発話であるかどうかも判別する必要がある．

3 学習データの作成

3.1 対象音の収集

各クラスの音データの収集を行う．本研究では，NAO のマイクを用いて，データの収集を行う．NAO には，頭部の前後左右に 4ch のマイク，同じく頭部の左右にスピーカが搭載されている．実験環境として，大学内の通常の居室を利用した．なお，収録したい音以外が入力されないようにした．

各クラスで使用した音データを以降に示す．また，クラスごとに収録した音の合計時間を表 1 に示す．

user は，ユーザ発話を外部のスピーカから再生して収録する．外部のスピーカは，ロボットの正面から 1m 離れた位置に設置した．使用したコーパスは，日本音響学会で作成された新聞記事読み上げコーパス§に収録されている ATR 音素バランス文のうち，1 セット 50 文を男女 10 名ずつの計 20 話者が読み上げた音声を使用した．

noise は，10 秒の雑音ファイルを 24 個収録した．雑音として，ロボットの動作，咳，ペンを落とす，手をたたく，ドアの開閉などを収録した．実際にロボットと対話

Classifying Input Sounds Including Overlapped Speech in Spoken Dialogues with Humanoid Robot: *Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato (Nagoya Univ.)*

‡http://www.aldebaran-robotics.com/
§http://www.mibel.cs.tsukuba.ac.jp/_090624/jnas/

する時に発生しうる環境音を収集するために，このような方法を用いた．

robot は，ロボットのスピーカから音声ファイルを再生し，同時にこの音声を収録する．ロボットの発話の音声合成には，音声合成エンジン VoiceText[¶]の女性の声を使用した．発話の内容には，ATR 音素バランス文 10 セット 503 文を用いた．

user+robot は，外部のスピーカからユーザの発話を再生しているときに，ロボットのスピーカからも音声を再生することで収録する．発話内容として，音素バランス文 10 セットのうち，それぞれ 1 セットずつランダムに選択した．ユーザは，男女 3 名ずつの計 6 話者を使用した．なお，ロボットとユーザで同じセットが選択されないようにした．再生した音声ファイルは，各セットで再生時間が異なるため，一方の音声ファイルが終了した時点で収録を止めた．

3.2 4 クラスの GMM に用いる学習データ

学習データとして，実際に音声認識で使用する音を用いる．つまり，4ch のマイクで収集した音データに対して音源分離を行い，出力された分離音を学習データとする．分離音の生成には，ロボット聴覚ソフトウェア HARK[3] を用いる．HARK では，4ch の音響信号に対して音源定位を行い，定位された方向のみの 1ch の分離音を生成できる．音源定位は，MUSIC 法に基づいており，入力される音源のパワーが閾値以上の場合，1 フレーム（0.01 秒）ごとに定位角度とそのパワーが出力される．インパルス応答は，マイクの中心から 1m の距離から，10 度ごとに 36 点計測した．したがって，定位の角度分解能は 10 度である．音源分離部での伝達関数の作成にも，音源定位部と同様に測定したデータを利用した．

クラスごとの分離音の合計時間を表 1 に示す．分離音が収集した音よりも短いのは，HARK では，音源のパワーが閾値以上の区間のみ定位・分離され，それ以外の区間の分離音は生成されないためである．また，得られた分離音を全て学習データとして使用するため，各クラスの分離音の中には，ほとんど音声が含まれていないものや，そのクラスに分類されるべきではないものを含む可能性がある．しかし，各クラスの分離音には，それらの音データよりも目的の音データのほうが多いため，それらによる影響は少ないと考える．

4 入力音の判別精度の評価

4 クラスの Gaussian Mixture Model（GMM）[4] による判別実験を行う．GMM は，特徴量の分布を複数の正規分布の重み付き和で表したものである．ある時刻に入力された特徴量に対して，各モデルの尤度を出力する．入力音に対してフレームごとに対数尤度を加算し，各クラスのうち最もその総和が大きい GMM が選択される．

表 1 で示した分離音を学習データとして，4 クラスの GMM を作成した．GMM の作成には，HTK（Version 3.4.1）[‖]を利用した．以下の 2 種類の特徴量を使用した．

- MFCC（12 次元），ΔMFCC，Δ パワーの 25 次元
- MFCC（12 次元），ΔMFCC，パワー，Δ パワーの 26 次元

表 2: 次元数の異なる GMM による判別結果

			output（個）				正解率
	次元数	class	user	noise	robot	user+robot	(%)
input	25	user	15	3	0	12	50
		noise	5	24	0	1	80
		robot	1	0	28	1	93
		user+robot	2	0	3	25	83
	26	user	21	2	1	6	70
		noise	1	25	3	1	83
		robot	0	1	29	0	97
		user+robot	0	0	1	29	97

前者は，Julius 標準の音響モデルで使用される 25 次元の特徴である．後者は，前者にパワーを加えた 26 次元の特徴である．パワーを加えた理由は，ロボットとユーザの発話を，ロボットのスピーカで収録する場合，ロボットの発話のパワーのほうが大きいため，これが特徴として利用可能と考えたからである．学習データが少ないため，どちらの GMM も混合数を 1 とした．

テストデータとして，各クラスごとに 30 個ずつ，計 120 個のファイルを用意した．user と user+robot のテストデータで用いた話者は，学習データには使用していない．

判別は，Julius[**]に-gmm オプションを付けて行った．このとき，入力音の音響的特徴から GMM ごとに尤度計算を行い，最も近い GMM が出力される．各 GMM を 1 つの HTK 形式のファイルとして作成し，これを-gmm で与えることで，認識結果とは別に判別結果を得る．テストデータに対する判別結果を，表 2 に示す．各行は入力したテストデータのクラスで，各列は Julius が出力した判別結果である．正解率は，各クラスのテストデータ 30 個のうち，適切に判別できた率を示す．特徴量の次元による違いを比較すると，全体的に 26 次元のほうが正解率が高かった．これにより，パワーが入力音判別の特徴として有効であることが示唆されている．

5 おわりに

4 章の結果より，作成したモデルによりロボットの発話と発話の重なりを判別できることを確認した．このことから，ユーザとロボットの発話が重なった場合でも，ユーザの発話の存在を認識できることを示した．今後は，作成したモデルを用いて，オンラインでロボットを動かし，実環境で利用できることを確かめる．

参考文献

[1] 武田龍, 中臺一博, 駒谷和範, 尾形哲也, 奥乃博. バージインを許容するロボット音声対話のための ICA を用いたセミブラインド音源分離. 情報科学技術フォーラム一般講演論文集, Vol. 6, No. 2, pp. 261–262, 2007.

[2] A. Lee, K. Nakamura, R. Nisimura, H. Saruwatari, and K. Shikano. Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs. *Proc. of INTERSPEECH*, pp. 173–176, 2004.

[3] K. Nakadai, T. Takahasi, H.G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and implementation of robot audition system “HARK”, advanced robotics. *VSP and Robotics Society of Japan*, Vol. 24, pp. 739–761, 2010.

[4] A. Reynolds and C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72–83, 1995.

[¶]http://voicetext.jp/

[‖]http://htk.eng.cam.ac.uk/

[**]http://julius.sourceforge.jp/