ISS-SP-248

マルチモーダル情報を用いた落語の役柄交替検出

細江 花　　酒向 慎司　　北村 正
Hana Hosoe　　Shinji Sako　　Tadashi Kitamura

名古屋工業大学
Nagoya Institute of Technology

1　はじめに

現在，文字媒体だけでなく映像による情報取得が身近になったが，聴覚障がい者など音声による取得が難しい人は字幕から情報を取得することが多い．しかし，字幕付与の作業が追い付いておらず，より効率的に付与する方法が望まれる．これまでに，落語の役柄交替の検出に関して様々な研究がなされており，川嶋らの研究では，演者の頭部動作の変化量から身体動作と発話のタイミングのモデルが提案されている．そこで，本研究では関連研究 [1] の提案したモデルをもとに，発話権の移動が最適化されていると考えられる落語を用い，マルチモーダル情報の特徴量の変化から役柄交替のタイミングを検出し，役柄交替検出に有効となり得るか検討する．

2　役柄交替のモデル

関連研究 [1] では落語を用い，発話区間と頭部動作に注目し役柄交替のモデルを作成している．視覚情報として，落語では頭部を左右に振り向ける動作を行うことで，役柄交替が発生したことを観客に知らせる．この動作は直前の役柄の発話に対し，役柄交替回数の 30 – 50 % の頻度で非同期となる知見が得られた．したがって，検出を行う際には時間が正確に合致するのではなく，幅を持たせた検出が必要である．

3　役柄交替のモデルを用いた検出手法

3.1　視覚情報による検出

頭部動作に注目し役柄交替を検出する．頭部の動き及び顔の向きを知るために，フレーム中の検出範囲を制限し，話者の頭部のみを肌色検出によって検出する．肌色領域の左端列，右端列，肌色ピクセル数が最大となる列を検出することにより，フレーム中の頭部の位置と顔の向きを決定する．

3.2　音声情報による検出

音声の物理量とそれに関連する役柄交替時の特徴として以下の 2 つが挙げられる．

- 短時間パワー：発話と発話のあいだに発生する間．
- 基本周波数：発話区間同士の基本周波数の差．

これらの特徴を用いた検出方法を考える．

役柄交替が発生する際の発話と発話のあいだに間が発生するために，短時間パワーがある閾値の絶対値より大きくなった時間を役柄交替が発生したとして検出する．

2 つ目の特徴は，発話区間内で役柄交替が発生する場合，役柄交替を行わない場合に比べ基本周波数が大きく変化する．この変化を，まず，発話区間内での基本周波数の平均を，その次に，基本周波数の平均を用いて閾値を設定し，平均を下回ったデータのうち，閾値を下回ったタイミングを，役柄交替が発生したとして検出を行う．

3.3　マルチモーダル情報による検出

3.1 節，3.2 節で検出された結果を用い，マルチモーダル情報による役柄交替の検出を行う．この時，2 章で述べた非同期性を考慮するために，視覚情報と音声情報それぞれの役柄交替が発生したとする時間の差が 0.5 秒以内である場合を，マルチモーダル情報によって検出したとする．

4　実験

実験には落語家桂枝雀が昭和 54～59 年に行った落語の寄席の映像のうち，正面から撮影された全身映像を用いた．収録したデータでは 319 回の役柄交替が行われている．視覚情報のみ，音声情報のみ，マルチモーダル情報それぞれで役柄交替検出を行った結果を表 1 に示す．

視覚情報のみの場合，0.843 という高い再現率となったが，適合率は 0.083 であるため，新たな特徴量や検出方法を考える必要がある．音声情報のみの場合，適合率は高くなったが，実際の役柄交替のうち 0.558 しか正しく検出することができなかった．マルチモーダル情報で検出した時も，適合率が高くなったが，“はい” などの短い応答で発生しやすい，頭部動作がほとんどない場合の役柄交替検出を行うことができなかった．

5　むすび

マルチモーダル情報を用いることで，視覚情報/音声情報それぞれのみの場合に比べ適合率を向上することができたが，再現率が低下した．今後，再現率を向上させるために，音声情報に含まれる役柄交替を意図しない間の扱いの検討と，頭部動作を伴わない役柄交替の検出手法の提案が望まれる．

参考文献

[1] 川嶋宏彰ほか，“落語の役柄交替における視覚的「間合い」の解析”，情報処理学会，48 巻，12 号，pp.3715–3728，2007.

表 1　検出結果

	視覚	音声	マルチモーダル
検出数	3223	1161	709
正解数	269	178	177
適合率	0.083	0.153	0.250
再現率	0.843	0.558	0.555