映像コンテンツ制作におけるマルチモーダル感情表出モデルの構築に関する研究

(14208031)

平成14年度～平成16年度 科学研究費補助金 基盤研究(A)(2)
研究成果報告書

平成17年3月

研究代表者　白井 克彦
(早稲田大学理工学術院 教授)

はしがき

本報告書は，平成14年度から平成16年度の3年間にわたり，文部科学省科学研究費補助金 基盤研究(A)(2)「映像コンテンツ制作におけるマルチモーダル感情表出モデルの構築に関する研究」において得られた研究成果をまとめたものである．

本研究では，コンピュータを用いることで，専門的な技術やコストを必要とせず，個人レベルで映像コンテンツを制作することを目的としている．コンピュータを用いて映像制作をする際に，最も重要であると考えられるのが，状況に応じて複雑に変化する登場人物の動作及び発話の制御である．本研究では特に，登場人物の動作と発話時の韻律を適切に制御することで，動作と音声に感情を付与することのできるマルチモーダル感情表出モデルの構築を目的としている．

主要な研究テーマとして，以下の5つが挙げられる．

1. 音響特徴量の分析
2. 文章の感情理解
3. 動作制御モデルの構築
4. 感情動作の表出
5. 感情表現における映像と音声の役割

本研究は，映像コンテンツ制作において今後必要になると予想されるマルチモーダル感情表出モデルの構築を試みたものであり，人間の動作と音声に表出する感情情報及び感情表現における映像と音声の役割を明らかにしている．また，動作に関しては感情表出モデルを用いて動作制御が可能なシステムを実装している．以上の点から，当初の目的は達成されたものと考える．

平成17年3月

白井 克彦

課題番号

14208031

研究課題

映像コンテンツ制作におけるマルチモーダル感情表出モデルの構築に関する研究

研究組織

研究代表者　白井 克彦　早稲田大学理工学術院 教授
研究分担者　小林 哲則　早稲田大学理工学術院 教授
研究分担者　米山 正秀　東洋大学工学部 教授
研究分担者　山崎 芳男　早稲田大学国際情報通信研究科 教授
研究分担者　大平 茂輝　名古屋大学情報メディア教育センター 助手
研究分担者　村上 真　東洋大学工学部 講師

研究経費

平成14年度	19,600千円
平成15年度	9,200千円
平成16年度	3,500千円
計	32,300千円

目 次

第1章　本報告書の構成 **1**

第2章　音響特徴量の分析 **3**
2.1　感情抽出のための韻律情報の分析 3
2.1.1　目的 3
2.1.2　音声データ 5
2.1.3　分析手法 6
2.1.4　判別結果 14
2.1.5　結論 31
2.2　「笑い声」合成のための音響特徴の分析 33
2.2.1　目的 33
2.2.2　音響特徴の抽出 34
2.2.3　笑い声らしさに対する貢献度の評価 38
2.2.4　結論 44

第3章　文章の感情理解 **49**
3.1　目的 49
3.2　大局的感情分類 50
3.2.1　場面分割 50
3.2.2　感情表現語辞書と文章の大局的感情決定 53
3.3　局所的感情分類 55
3.3.1　構文解析 55
3.3.2　感情表現要素辞書と局所的感情分類の決定 57
3.4　実験 58
3.4.1　大局的な感情分類に対する実験 59
3.4.2　局所的な感情決定に関する実験 63
3.5　結論 85

第4章　動作制御モデルの構築 **87**
4.1　目的 87
4.2　歩行動作制御モデル 88
4.2.1　目標システム 88
4.2.2　歩行動作制御モデルの定義 88

4.2.3 歩行動作制御モデルの構築 89
4.3 モデル評価実験 104
4.3.1 評価実験手法 104
4.3.2 評価実験結果 105
4.3.3 考察 105
4.4 システム評価実験 107
4.4.1 実装システム 107
4.4.2 評価実験手法 107
4.4.3 評価実験結果 107
4.4.4 考察 109
4.5 結論 110

第5章 感情動作の表出 **111**
5.1 感情表出モデルの構築 111
5.1.1 目的 111
5.1.2 データ収集 112
5.1.3 感情表出モデルの構築 113
5.1.4 ディジタルアクタシステムの実装 117
5.1.5 評価実験 121
5.1.6 結論 122
5.2 心理的尺度による感情動作制御 123
5.2.1 目的 123
5.2.2 心理的尺度 123
5.2.3 実験 124
5.2.4 実装システム 128
5.2.5 結論 130

第6章 感情表現における映像と音声の役割 **131**
6.1 目的 131
6.2 感情表現収録実験 132
6.3 主観評価実験 133
6.3.1 画像と音声の組替え合成方法 133
6.3.2 実験内容 134
6.3.3 観点1評価結果 135
6.3.4 観点2評価結果 135
6.4 結論 144

第7章 あとがき **147**

研究発表（目次）

1. 学会誌論文

[1] 小林 哲則："音声対話研究の現状と動向", 人工知能学会誌, Vol.17 No.3, pp.266-270, May.2002. 157

[2] 大川 茂樹, 白井 克彦："特徴量レベルでの統合法に基づくマルチバンド型モデルによる雑音環境下音声認識", 情報処理学会論文誌, Vol.43 No.07-005, pp.2046-2054, Jul.2002. 162

[3] 菊池 英明, 白井 克彦："音声対話における心的状態変化の予測をともなうメタ発話生成機構", 情報処理学会論文誌, Vol.43 No.07-014, pp.2130-2137, Jul.2002. 171

[4] Makoto MURAKAMI, Katsuhiko SHIRAI, Masahide YONEYAMA : "Human Face Extraction using Genetic Algorithm with Similarity of Subspace Method as a Fitness Value", International Journal of Knowledge-Based Intelligent Engineering Systems, Vol.6, No.4, pp.178-185, Oct.2002. 179

[5] 村上 真, 米山 正秀, 白井 克彦："Labeled Graph Matching を用いた動画像に対する人物頭部及び表情変化を伴う部位の抽出", 電子情報通信学会論文誌, Vol.J85-D-II, No.11, pp.1656-1663, Nov.2002. 187

[6] 村上 真, 森野 岳宏, 長船 大毅, 誉田 雅彰, 白井 克彦："感情語による歩行動作制御を目的としたディジタルアクタシステムの構築", 電子情報通信学会論文誌, 査読中, ... 195

2. 国際会議論文

[1] Toru Hayakawa, Hiroaki Kato, Yoshinori Sagisaka, Katsuhiko Shirai : "Automatic phone segment alignment using statistical deviations from manual transcriptions," FIRST PAN-AMERICAN/IBERIAN MEETING ON ACOUSTICS, Dec.2002. 205

[2] Makoto Murakami, Masahide Yoneyama, Katsuhiko Shirai : "Accurate Human Face Extraction using Genetic Algorithm and Subspace Method," Frontiers in Artificial Intelligence and Applications, Vol.87, SOFT COMPUTING SYSTEMS, pp.745-754, IOS Press, Dec. 2002.210

[3] Makiko Muto,Yoshinori Sagisaka,Takuro Naito,Daiju Maeki,Aki Kondo, Katsuhiko Shirai : "Corpus-Based Modeling of Naturalness Estimation in Timing Control for Non-Native Speech," Proc. Eurospeech 2003, pp.401-404, Sep.2003. 220

[4] Shinya Fujie, Yasushi Ejiri, Yosuke Matsusaka, Hideaki Kikuchi, Tetsunori Kobayashi : "RECOGNITION OF PARA-LINGUISTIC INFORMATION AND ITS APPLICATION TO SPOKEN DIALOGUE SYSTEM," Proc. of Automatic Speech Recognition and Understanding Workshop(ASRU), pp.231-236, Dec.2003. 224

[5] Kei Yazu, Makoto Murakami, Futoshi Sugimoto, Masahide Yoneyama : "A Method to Add Emotional Expressions to Text-to-Speech Synthesis," International Conference on Intelligent User Interfaces(IUI2004), Jan.2004. 230

[6] Shinya Fujie, Daizo Yagi, Yosuke Matsusaka, Hideaki Kikuchi, and Tetsunori Kobayashi : "Spoken Dialogue System Using Prosody as Para-Linguistic Information," International Conference on Speech Prosody 2004 (SP2004), pp.387–390, Mar.2004. 233

[7] Futoshi Sugimoto, Kei Yazu, Makoto Murakami, Masahide Yoneyama : "A Method to Classify Emotional Expressions of Text and Synthesize Speech," IEEE International Symposium on Control, Communications and Signal Processing(ISCCSP2004), pp.611–614, Mar.2004.237

[8] Shinya Fujie, Daizo Yagi, Hideaki Kikuchi, Tetsunori Kobayashi : "Prosody based Attitude Recognition with Feature Selection and Its Application to Spoken Dialog System as Para-Linguistic Information," Proceedings of International Conference on Spoken Language Processing, ICSLP2004, vol.4, pp.2841-2844, Oct.2004. . 241

[9] Tosimitu Tukamoto, Futoshi Sugimoto, Masahide Yoneyama : "A Method to Classify Emotion in Text, Based on the Distribution of Emotional Words," OZCHI2004, Nov.2004. 245

[10] Shinya Fujie, Kenta Fukushima, Tetsunori Kobayashi : "A Conversation Robot with Back-Channel Feedback Function based on Linguistic and Nonlinguistic Information," Proc. International Conf. on Autonomous Robots and Agents, ICARA2004, pp.379-384, Dec.2004. 247

[11] Takehiro Morino, Daiki Osafune, Takahiro Itano, Katsuhiko Shirai, Makoto Murakami : "Construction of the Walking Motion Model for Animated Characters' Motion," Proc. of the IASTED International Conference on Internet and Multimedia Systems and Applications (EuroIMSA 2005), pp.1-6, Feb.2005. 253

3. 研究会資料

[1] 松坂 要佐, 小林 哲則 : "対面対話の収録と解析のための信号処理とパターン認識", 人工知能学会研究会資料, SIG-SLUD-A201-05, pp.27-32, Jun.2002. 259

[2] 小川 哲司, 小林 哲則 : "部分隠れマルコフモデルによる連続音声認識", 電子情報通信学会技術研究報告, SP2002-40, pp.25-30, Jul.2002. 265

[3] 松坂 要佐, 江尻 康, 小林 哲則 : "合意形成型対話における声・顔の機能的表情の分析と認識", 電子情報通信学会技術研究報告, PRMU2002-118, pp.31-36, Nov.2002. 271

[4] 大隈 健史, 長船 大毅, 森野 岳宏, 村上 真, 白井 克彦 : "人物動作における感情表出モデル構築に向けたディジタルアクタ学習システムの設計", 第 9 回画像センシングシンポジウム講演論文集, pp.427-432, Jun.2003. 277

[5] 藤江 真也, 江尻 康, 菊池 英明, 小林 哲則 : "パラ言語の理解能力を有する対話ロボット", 情報処理学会研究報告, SLP-48, pp.13-20, Oct.2003. 283

[6] 大久保 崇, 菊池 英明, 白井 克彦 : "音声対話における韻律を用いた話題境界検出", 電子情報通信学会技術研究報告, Vol.103, No.519, pp.235-240, Dec.2003. 291

[7] 藤江 真也, 福島 健太, 柴田 大輔, 小林 哲則：“FST と韻律情報を用いた相槌・復唱機能を有する対話ロボット”, 人工知能学会研究会資料, SIG-SLUD-A401-03, pp.15-20, Jun.2004. 297

4. 全国大会講演論文

[1] 塩田 大輔, 王 大可, 大川 茂樹, 白井 克彦：“様々な雑音環境下におけるマルチバンド型音声認識の性能評価”, 日本音響学会 2002 年春季研究発表会講演論文集, 3-P-4, pp.173-174, Mar.2002. 303

[2] 森野 岳宏, 大隈 健史, 長船 大毅, 村上 真, 白井 克彦：“基本動作制御モデル構築のための歩行動作の分析・生成”, 人工知能学会 全国大会 (第 17 回) 論文集, 2C3-03, Jun.2003. 305

[3] 長船 大毅, 森野 岳宏, 大隈 健史, 村上 真, 白井 克彦：“モーションキャプチャデータを用いた歩行動作の分析と制御”, 人工知能学会 全国大会 (第 17 回) 論文集, 2C3-04, (CD-ROM), June.2003. . . 308

[4] 大久保 雅史, 望月 亮, 蓑輪 利光, 小林 哲則：“波形重畳型音声合成における心理的態度の再現性評価”, 第 2 回 情報科学技術フォーラム (FIT) 一般講演論文集, vol.2, pp.265-266, Sep.2003. 310

[5] 藤江 真也, 八木 大三, 菊池 英明, 小林 哲則：“パラ言語情報を用いた音声対話システム”, 日本音響学会秋季研究発表会講演論文集, pp.39-40, Sep.2003. 312

[6] 谷津 慶, 村上 真, 杉本 富利, 米山 正秀：“テキスト音声合成における感情表現の付加”, ヒューマンインタフェースシンポジウム 2003, 2535, pp.539–540, Oct.2003. 314

[7] 八木 大三, 藤江 真也, 菊池 英明, 小林 哲則：“韻律情報を用いた肯定的／否定的態度の認識”, 日本音響学会春季研究発表会講演論文集, pp.141-142, Mar.2004. 316

[8] 大久保 雅史, 望月 亮, 小林 哲則：“心的態度表現における韻律的／分節的特徴の影響”, 日本音響学会春季研究発表会講演論文集, pp.375-376, Mar.2004. 318

[9] 大久保 崇, 菊池 英明, 白井 克彦：“韻律情報を利用した文章入力システムのための韻律制御モデル”, 日本音響学会秋季研究発表会講演論文集, 3-1-12, vol.I, pp.133-134, Sep.2004. 320

[10] 小林 季実子, 菊池 英明, 白井 克彦：“音声対話における発話の感情判別”, 日本音響学会秋季研究発表会講演論文集, 2-5-5, vol.I, pp.281-282, Sep.2004. 322

[11] 板野 貴之, 森野 岳宏, 長船 大毅, 誉田 雅彰, 村上 真, 白井 克彦：“モーションキャプチャデータを用いた多様な歩行動作制御手法の検討”, FIT2004, (CD-ROM), Sep.2004. 324

第1章　本報告書の構成

コンピュータを用いて映像コンテンツを制作する上で，最も重要であると考えられるのは，状況に応じて複雑に変化する登場人物の動作及び発話の制御である．本研究では特に，登場人物の動作と発話時の韻律を適切に制御することで，動作と音声に感情を付与することのできるマルチモーダル感情表出モデルの構築を目的としている．

音声に感情情報を付与するためには，人間の音声に感情情報がどのような形で表出するかを明らかにする必要がある．人間の音声を収録し分析する際には，台詞や場面を限定して単文を発話させたデータを使用することが多いが，本研究では演技の流れによって自然なもっともらしい感情表現がされている落語音声を使用して分析を行う．また，従来「喜び」「怒り」「疑い」といった感情を含む音声の分析は数多く行われているが，本研究では，それらに加えて生理的な発声による感情表現に着目し「笑い声」の分析を行う．第2章にその詳細を示す．

次に，台本の台詞のみから感情音声を合成するには，台詞や文章といったテキスト情報から感情を推定する必要がある．本研究では，入力された文章の感情の推定を，大局的感情分類と局所的感情分類の2つの方法によって実現する．大局的感情分類とは，入力された文章をText tiling algorithmを使用し場面に分割し，分割された各場面に存在する単語の頻度から場面毎の感情を推定する方法であり，局所的感情分類とは，予め文型毎に感情を決定するための辞書を作成し，入力された文章の一文一文に対して感情を推定する方法である．第3章でこれらについて説明する．

直感的にキャラクタアニメーションを制作可能なシステムを構築するためには，少ない指示で多様な動作を精密に生成・制御可能なモデルを構築する必要がある．人間の動作は個々の可動部位を連動させて制御されているものであることから，本研究では，可動部位の連動性に着目してモデルの構築を行う．具体的には，モーションキャプチャ装置で収録した歩行動作データを元に，各部位の空間的及び時間的連動性を考慮して，その連動性の高い部位をまとめて制御可能な関節構造モデルを構築する．これにより，少ない指示で歩行動作の制御が可能となる．次に，物理的な特徴を表す特徴量成分を実測歩行動作データから抽出し，それに動作を制御するパラメータとしての重みを掛け合わせることで歩行動作を生成・制御可能な歩行動作制御モデルの構築を行う．このことにより，多様な歩行動作を生成・制御することが可能となる．第4章で動作制御モデルの詳細を述べる．

次に，CGキャラクタに感情動作を表出させる方法について検討を行う．感情語による指示で動作を生成可能なディジタルアクタシステムを構築するには，感情語と動作との関係を記述した感情表出モデルを構築する必要がある．本研究では，感情をいく

つかのカテゴリで表現されるものとしては扱わず，より多様なものとして捉える．したがって，いくつかのカテゴリの感情動作を収録して分析することによりモデルを構築するという手法はとらず，多様性を保証した方法でデータ収集をし，モデルの構築を行う．また，多様性を持たせてモデルを構築したとしても，未知語による指示が出現することは避けられない．そのため，未知語が入力された場合に新たに感情を学習し，感情表出モデルを更新する感情学習の機能についても検討を行い，ディジタルアクタシステムを実装する．また，言語による指示とは別に，スライダの操作によって感情動作を制御するインタフェースの構築も行う．感情動作をスライダにより制御するためには，スライダの１目盛り毎に均等に感情を変化させる必要がある．そこで本研究では，心理的尺度という指標を用い，感情の変化と動作制御特徴量との関係を被験者実験により明らかにする．これら感情動作の表出については第５章で述べる．

最後に，映像コンテンツ中のキャラクタによって適切な感情表出を行うには，映像や音声といった物理信号と，人間がそこから知覚する感情との対応関係を明らかにする必要がある．そこで，映像と音声を組み合わせ，それぞれに感情表現を付与した場合，人間がそれをどのように知覚するかについて分析を行う．具体的には，映像と音声に感情表現を付与したデータを用いて被験者実験を行い，(a) 被験者が意図したとおりの感情を知覚できるかどうか，(b) 一方のモダリティ(映像または音声) に対してもう一方のモダリティ(音声または映像) の感情レベルを変更させた場合，被験者が意図したとおりの感情レベルを知覚できるかどうか，の２点に着目して分析を行う．第６章でそれらについて述べる．

巻末には，各章節で扱っている個々の問題について詳細に報告された，本研究の成果である論文及び研究会資料を添付している．

第2章　音響特徴量の分析

音声に感情情報を付与するためには，人間の音声に感情情報がどのような形で表出するかを明らかにする必要がある．そこで本章では，人間の音声を収録し分析することで，感情表現に重要な音響特徴量を明らかする．音声の分析には，台詞や場面を限定して単文を発話させたデータを使用することが多いが，本研究では演技の流れによって自然なもっともらしい感情表現がされている落語音声を使用して分析を行う．また，従来「喜び」「怒り」「疑い」といった感情を含む音声の分析は数多く行われているが，本研究では，それらに加えて生理的な発声による感情表現に着目し「笑い声」の分析を行う．

2.1　感情抽出のための韻律情報の分析

2.1.1　目的

コンピュータを用いた人間の創造活動は従来表現困難であったものを可能にする．また，個人レベルで作品を制作し，ネットワークを介して配信することが可能となれば，豊かな自己表現手段を獲得することになる．

本研究では，専門的な技術やコストを必要とせず，個人レベルで映像コンテンツを制作することを目的としている．映像コンテンツにおいて重要な要素としては，登場人物の動作や発話の制御が挙げられる．本節では，音声を適切に制御することで感情表現を付加することを目的としており，その第1段階として韻律の分析を行っている．感情表現を音声に付加した際の韻律の変化については多くの先行研究が行われている[1]-[5]．以下にその概要を説明する．

川波らの研究[1]では，種々な態度・感情における音声の韻律的特徴を分析し，朗読音声の韻律との比較を行っている．その結果，基本周波数パターンの文頭から文末への下降傾向が見られなくなること等が明らかになっている．また，模擬対話を行って感情のレベルを段階的に変化させた音声の分析では，感情のレベルと韻律的特徴との間には一定の傾向が見られるものの，話者あるいは感情の種類によって大きく異なっていることから人間が感情を表出する手段を複数有し，場当たり的に選択してレベルを表現しているのではないかと結論づけている．

また，藤井らの研究[2]では，音声に含まれている感情表現の所在解明を目指して単語と単文について調べている．特徴量としては，最高基本周波数・最大振幅・持続時間の3個のみを使って判別できるかを実験している．この結果，文のほうが単語より

も判別率が高く，母音のホルマント周波数も感情表現に大きく影響されていてその時系列が判別に寄与するという結果が得られている．

上床らの研究[3]は，4つの感情(怒り・歓喜・悲哀・驚き)により音声が受ける影響を，時間構造・振幅構造・ピッチ構造等の韻律情報について調べている．これにより得られた各感情の特徴に従って平静発話の音声を変形させた合成音声の聴取実験により，各感情の韻律制御ルールを提案している．

平賀らの研究[4]では，怒り・喜び・悲しみ・嫌悪の4感情について，豊かな感情表現の分析を行うために，演劇経験者が発生した音声やFMラジオのドラマ音声を使用している．これらの音声に対して，ピッチ周波数や振幅の変化パターンといった時間系列変化に関する分析を行っている．その結果，言い回しなどの時系列パターンに共通点が見られることから，これらが人間の感情認識に大きな影響を及ぼしていると結論づけている．

さらに，広瀬の研究[5]では，文末のアクセント成分の特徴が意図の識別に重要であるとしている．また，感情表現について，人間はある感情を表現するのにいくつかの手段を持っているが，段階を表現する際徐々にそれぞれの手段を顕著にしていくのではなく，利用する手段，あるいは割合を変更すると推測している．感情音声合成の際には，分節的特徴が重要であるとしている．

このように，感情や態度は人間の音声にいろいろな影響を及ぼすと考えられている．実際，人間同士の会話では，相手の表情やボディーランゲージなどの視覚的な情報があるとはいえ発話者の感情が話し方や声色に変化を与え，聞き手はそこから相手の感情・態度を感じ取っていると考えられる．しかしその表現方法は人それぞれであり，かなりの違いが生じる．にもかかわらず，我々は無意識のうちに臨機応変に対応している．これは，人間の優れた潜在的能力の一つであるといえるだろう．この能力について，どのようなメカニズムであるのかが解明できれば，現在の感情つき音声合成システム等における応用的な利用も可能になるであろう．

人間の感情とは非常に複雑なものであり，数えあげたらきりがないほどに微妙なニュアンスの違いを持っている．しかし対話中における感情の表現方法としては，それほど細かい分類をする必要はないであろう．なぜなら私たち人間同士の会話の中では，相手の機嫌がよさそうか悪そうか，つまり楽しそうかつまらなそうかといったことを中心としてその会話の続け方に変化が生じるからである．

そこで，本研究では基礎的な感情表現の分別をするために基本6感情を元にした6つの感情についてその判別を目指す．このとき，感情間の相違を含み，かつ，言語的情報による影響をあまり受けないものとして，基本周波数・パワー・発話速度等の韻律的特徴に焦点を当て，どのようなパラメータの使用がそれぞれの感情の判別に有効であるか分析を行うことにする．

今回は，落語音声を分析の対象データとした．これまでにも落語音声は，合成音声を作るための自然会話音声のデータとして，プロミネンスの解析などに使用されている[6]．本研究においても，落語音声は，感情表現のプロの音声であり，かつ，自然会話特有の特徴も含まれると考えられるためデータとして選択した．

本節では，落語音声を対象に，どのような韻律情報が感情判別に有効かを検討していくことにする．

2.1.2 音声データ

使用したデータ

先行研究の音声データでは，セリフや場面を限定して単文を演技させた感情表現というものが多いが，本研究では，より私たちの日常的な表現方法に近いと思われるような自然でうまい感情表現を使用したいと考えた．そこで，演技ではあるが，演技の流れによって自然なもっともらしい感情表現がされているものとして落語音声を選択した．

データの絞り込み方法

落語というジャンルの中でも，比較的有名な演目である「船徳」を題材とした．これを噺家A・B・Cの3名が演じている音声の一部を使用した．

音声の中で使用する(聴取実験を行う)部分の決定方法としては，書き起こしを行ったデータを参考にして，3名それぞれの演技がほぼ共通している部分を抜粋した．その例を図2.1に示す．

噺家A:ここにいる若旦那が，お止め申したんだが，
　　どうしても船頭になるとおっしゃるんだ
噺家B:でなあ，若旦那はこれから船頭になりてえとこういうわけだ
噺家C:若旦那が今日から船頭になりてえって
　　おっしゃるんだなあ，俺はお止め申したんだけど

図 2.1: 抜粋した発話の例

更にこの抜粋部分について聴取実験を行った．聴取実験は，

- 被験者8名(男性6名・女性2名)
- 発話単位ごとに区切った音声サンプルをスピーカーから呈示
- 回答は，基本6感情(喜び・驚き・怒り・恐れ・悲しみ・嫌悪)のうち，今回のデータでは出現頻度が少ないと思われる，悲しみ・嫌悪を除いた4感情に平静とあきれを加えた感情の中から該当するものの選択方式
- 上記6つの感情に当てはまらない場合には,「その他」として自分が思った感情を記す

として行った.

回答中の「その他」の中で，分析対象となる6つの感情に取りまとめられそうなものはそれぞれに含めて使用した．この評価結果のうち，半分(4名)以上が同じ感情を選んだ発話のみを使用可能な発話として採用した.

1. 分析対象となるデータ

   分析対象としたデータを表2.1に示す．左から順に，もとの演技時間・噺家3名の演技がほぼ共通している部分の時間と発話数・聴取実験によって採用された部分の時間数と発話数とする.

表 2.1: 分析対象データ量[分](()内は発話数)

	総演技	抜粋後	採用後
噺家A	21.45	約5.16(116)	約3.32(80)
噺家B	26.55	約3.09(85)	約2.29(63)
噺家C	20.45	約11.37(142)	約4.35(112)

2. 発話中の感情の内訳

   噺家A・B・Cの発話のうち，採用した発話中の感情の内訳は，それぞれ表2.2のようになる.

表 2.2: 採用発話中の感情の内訳[発話数]

	噺家A	噺家B	噺家C
平静	36	35	56
喜び	5	3	5
驚き	6	2	6
怒り	19	13	25
恐れ	7	8	7
あきれ	7	2	13
合計	80	63	112

2.1.3 分析手法

分析ツールには，ESPS/waves+，Excelを使用した．また，韻律的な特徴を抽出する際には以下のようなデータを使用した.

- 音声データの波形・スペクトログラム・$F0$外形

- 感情のラベルつき書き起こしファイル
- 人手による音素ラベリング情報・韻律ラベリング情報

特徴量抽出

韻律的な特徴としては，門谷らの研究 [7] では主に $F0$ を特徴量としているが，本研究では $F0$ にパワーや発話速度を加えて次のような特徴量の抽出を行った．

$F0_{max}$	:	発話中の $F0$ の最大値
$F0_{min}$	:	発話中の $F0$ の最小値
$F0_{ave.}$	:	発話中の $F0$ の平均値
$F0_{range}$	:	発話中 $F0$ の range ($F0_{max} - F0_{min}$)
$F0_{分散}$	:	発話中の $F0$ の分散
PW_{max}	:	発話中のパワーの最大値
PW_{min}	:	発話中のパワーの最小値
$PW_{ave.}$	:	発話中のパワーの平均値
PW_{range}	:	発話中のパワーの range ($PW_{max} - PW_{min}$)
$PW_{分散}$	:	発話中のパワーの分散
発話速度	:	発話中のモーラ数 /(発話長-ポーズ長)

その後，これらの韻律的な特徴を説明変数として用いて判別分析を行った．

判別分析

判別分析とは，いくつかのグループにあらかじめ分けられたサンプルデータがあるとき，それらの持っているいろいろな特徴を表す変数（説明変数）からグループの境界を予測する式を求め，この式を用いてそのサンプルデータや新たなデータがどのグループに属すかを判定するものである．サンプルデータがどの群に属すかを判別する基準としては

- 線形判別関数による判別
- マハラノビスの汎距離による判別

とがある．つまり，グループ化する上でどの説明変数が重要かを決定する手法である．

実際の線形判別関数による判別方法，マハラノビスの汎距離による判別方法としては 1 以降で述べる．

1. 線形判別関数による判別方法

   この方法は，グループで表されるカテゴリーデータを目的変数にとり，数量データで表される P 個の諸特性を説明変数にとって，関係式 (2.1) を作る．このとき，

この式を境界線とし $y>0$ と $y<0$ によってデータがどちらのグループに属するかを判別する．$y=0$ のときは判別不可能である．これが，各発話に定義された感情情報と一致するかどうかを判別するものである．

$$y = \text{切片} + a_1x_1 + a_2x_2 + \cdots + a_nx_n \tag{2.1}$$

a_i ： i 番目の説明変数の係数
x_i ： i 番目の説明変数の値

次に，最も有効と考えられるパラメータの組み合わせの選択は次のような方法で行った．まず，繰り返し回帰分析を行い式 (2.2) で表される説明変数選択基準 R_u を求めた．

$$R_u = 1 - (1 - R^2) \times (n + k + 1)/(n - k - 1) \tag{2.2}$$

R ： 重相関係数
n ： データ数
k ： 説明変数の数

式 (2.2) の R_u が最大となる場合の説明変数の組み合わせを求めることで，判別を行うのに最も効果的と思われる説明変数を抽出した．

これらの作業を繰り返し行い，各感情の判別に有効な説明変数の抽出と判別的中率を求めた．

2. マハラノビスの汎距離による判別方法

この方法はいくつかの点の中から任意の 2 つを選択し，2 点間の距離を計算する際に用いるもので，ユークリッド距離の値に全ての点のばらつきを考慮して算出するものである．つまり，ある任意のサンプルデータがどちらのグループに属するかを判定するときに用いる．

これは，個々のサンプルデータについて 2 つのグループの平均 (重心) までのマハラノビスの汎距離を求め，得られた 2 つの距離の大小関係から，どちらのグループに属するか (近いか) を判定するものである．

i 番目のサンプルデータの両群の重心までのマハラノビスの汎距離 $D^2_{i(1)}$，$D^2_{i(2)}$ は，それぞれ式 (2.3)，式 (2.4) によって求められる．

$$D^2_{i(1)} = \sum_{j=1}^{p}\sum_{k=1}^{p}(x_{ij} - \overline{x}_{j(1)})S^{jk}(x_{ik} - \overline{x}_{k(1)}) \tag{2.3}$$

$$D^2_{i(2)} = \sum_{j=1}^{p}\sum_{k=1}^{p}(x_{ij} - \overline{x}_{j(2)})S^{jk}(x_{ik} - \overline{x}_{k(2)}) \tag{2.4}$$

$\overline{x}_{j(1)}$: j 番目変数の第 1 群の平均
$\overline{x}_{j(2)}$: j 番目変数の第 2 群の平均
S^{jk} : 次に示す S' の逆行列
$S_{(1)}$: 第 1 群の偏差平方和 • 積和行列
$S_{(2)}$: 第 2 群の偏差平方和 • 積和行列
$S'_{(1)}$: 第 1 群の分散 • 共分散行列 $S'_{(1)} = S_{(1)}/(n_1 - 1)$
$S'_{(2)}$: 第 2 群の分散 • 共分散行列 $S'_{(2)} = S_{(2)}/(n_2 - 1)$
S' : $S'_{(1)}$ と $S'_{(2)}$ の加重平均

$$S' = (n_1 - 1)S'_{(1)} + (n_2 - 1)S'_{(2)}/(n_1 + n_2 - 2)$$

(ただし，n_1, n_2 は両群のサンプル数)

3. 判別分析の手順

判別分析には発話単位全体を一つのデータとしたものについての判別分析を行ったが，表 2.7-2.15 からもわかるようにそれほど良い結果を得られなかった．そこで，発話単位ごとに X-JToBI[8] による韻律ラベリングを行い (図 2.3 参照)，得られたアクセント句単位のうち文頭・文末のアクセント句それぞれ 1 つずつをデータとしたものについても判別分析を行った．X-JToBI による韻律ラベリングに関しては 4 で述べる．

また，発話単位全体の分析の際には線形判別関数のうち，もっとも有効と思われるパラメータの組み合わせによる判別とマハラノビスの汎距離による判別とを行ったが，前者は各感情のその感情データ内での判別率があまり良くなかった (表 2.7,2.8) ため，アクセント句単位の判別分析の際には手法として使用しなかった．

さらに分析を行う際，まずはじめに 6 感情全体のデータから各感情に分ける判別分析を行ったが，データ全体に対する判別率はある程度高い時でも各感情データ内における判別率が低いという傾向があり (表 2.7,2.9,2.16,2.23)，信頼性が低いと思われた．

そこで，6 感情全体のデータから各感情ひとつひとつに分けるまでの間に，喜び・驚き・怒り，平静・恐れ・あきれのグループに分け，感情表現が豊かなものとそうでないものという基準で 2 分類を行った．これによって，最初の感情分別の境界がある程度あいまいなものとなり，判別が行いやすくなるのではないかと考えた．グルーピング方法は図 2.2 のようにする．2 分類からそれぞれの感情に分け

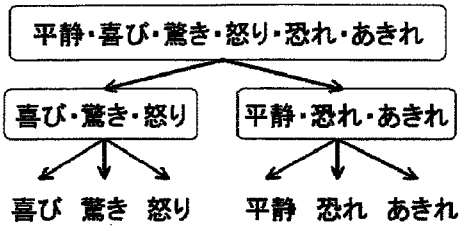


図 2.2: 判別分析でのグルーピング方法

る際の目的変数の取り方としては,「判別したい感情とそれ以外」という方法にした.

また，判別分析を行う際に使用したパラメータのうち，分析に対してどのパラメータがどのくらい効いているのかということを判断するために判別係数も求めた．判別係数の絶対値が大きいほど，その感情の構成に対してその要素が大きく影響していると考えられる.

4. 韻律ラベリングの概要

本研究では，音声データをアクセント句単位に分けるために X-JToBI による韻律ラベリングを行った．ここではその概要について述べる.

音声の韻律情報とは，声の変化の時間変化パターンのことであるイントネーションと声質から成る．ところで，音声データベースに韻律情報を登録して計算機で情報検索できるようにするためには，韻律情報を何らかの一貫した方法で記号化する必要がある．この作業が韻律ラベリングである.

本研究で使用した X-JToBI とは，日本語音声の韻律ラベリングスキームである J_ToBI の話し言葉音声への適用を目的とした拡張版で，extended J_ToBI の略称である.

発話単位の構成は下位から順に，語 (Word)→ アクセント句 → イントネーション句 → 発話単位となる．アクセント句には 0 個以上のアクセントが含まれている.

X-JToBI では，主に $Tone$, $Word$, $BreakIndex$(以降 BI) の層に各層のラベリング基準に従って記号を付与するものである．X-JToBI において用いる記号とその説明はそれぞれ次のようになる.

(a) $Tone$ 層

$Tone$ 層では，基本的に $F0$ 曲線を記述するために高 (H) 低 (L) 二種類の音韻的トーンを付与する．これは，アクセント，アクセント句の境界，イントネーション句の境界に対して付与する．そのラベルは以下のようになる.

- アクセント句頭の上昇音調
  - A : アクセント
  - $H-$: アクセント句頭の上昇のピーク

- アクセント句の境界音調

$L\%$	:	アクセント句終端での下降調の場合
$\%L$	:	アクセント句始端
$(H\%)\%L$	:	引用の形で高いトーンから発話が始まる場合のアクセント句始端
$L\%\%L, HL\%\%L,$ $H\%\%L, LH\%\%L$	:	連続するアクセント句境界で $F0$ に連続性が見られる場合

- イントネーション句末の境界音調

$H\%$	:	イントネーション句終端で上昇調の場合の $F0$ 極大値
$HL\%$	:	イントネーション句終端で上昇下降調の場合の $F0$ 極小値
$LH\%$	:	イントネーション句終端で下降上昇調の場合の $F0$ 極大値
$HLH\%$	:	イントネーション句終端で上昇下降上昇調の場合の $F0$ 極小値
pH	:	上昇下降調の場合の $F0$ ピーク位置および上昇下降上昇調の場合の $F0$ 極大値
pL	:	下降上昇調の場合の $F0$ 上昇開始位置および上昇下降上昇調の場合の $F0$ 極小値
$>, <$	:	トーンの同じレベルでの保持 (引き伸ばし) における始端あるいは終端

- その他の補助記号

x	:	$F0$ の欠損などにより位置が決定できない場合
?	:	トーンのカテゴリが不明確な場合
!	:	トーンを付与する位置が不明確な場合

$Tone$ 層の各ラベルをつけた場所の時間情報をイントネーションの情報 ($F0$ 外形) 上にプロットし，各点を直線で結ぶことによって，$F0$ 外形の曲線を直線近似として見ることができる．また，発話によっては，実際のアクセントが文法上のアクセントとは異なるところに出現したり，声量や声質などの変化によって目視ではアクセントの位置を特定できないケースも多い．そこで，アクセントの有無やアクセント句を担うモーラ位置の確定は，聴覚により知覚された位置に行う．$BPM(BoundaryPitchMovement)$ の有無も同様である．また，BPM とは，句末のイントネーションの上下動を示す．

(b) $Word$ 層

$Word$ 層では，単語ラベルを付与し，実際に知覚できた位置にアクセント“ ’ ”を記す．

(c) BI 層

BI 層では，単語間の結合度を整数値で記す．この数値が小さいほど単語間の結合度が高くなる．尚，BI 層では $BI = 1$ が形態素ごとに与えられ，これが初期値となる．表 2.3 にそのラベルを示す．

これらの規則を使って韻律ラベリングを行った例を 2.3 に示す．

表 2.3: BI 層のラベル

ラベル	使用法
1	語ないし形態素の境界
1+	1と2の中間
$1+w$	複合語内部の韻律境界
$1+p$	アクセント句内部のポーズ
2	アクセント句の境界
2+	2と3の中間
$2+p$	句末に BPM なし，ポーズが後続，ピッチレンジリセットなし
$2+b$	アクセント句末に BPM，ポーズ後続せず，ピッチレンジリセットなし
$2+pb$	アクセント句末に BPM，ポーズ後続，ピッチレンジリセットなし
3	イントネーション句の境界
3+	発話の最後
D	語断片，ピッチリセットなし
$D+$	語断片，ピッチリセットあり
P	語内部のポーズ，ピッチリセットなし
$P+$	語内部のポーズ，ピッチリセットあり
PB	境界音調の連鎖
$<F$	韻律的フィラーの始端
F	韻律的フィラーの終端

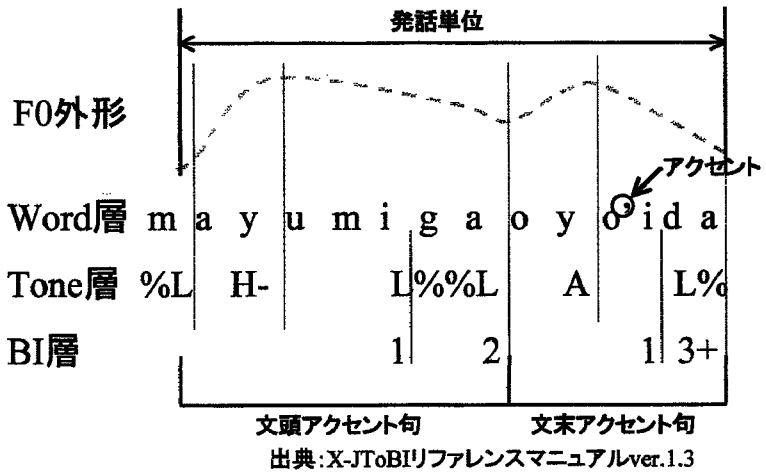


図 2.3: 韻律ラベリングの例

5. 発話単位全体の分析で実際に使用したパラメータ

(a) 線形判別関数による判別分析

分析を行う際に，他の説明変数との相関が強いものはパラメータとして適さないのではじめに除いておく必要がある．各噺家の最初に取り除くパラメータは以下のようになる．

- 噺家 A：$F0_{range}$
- 噺家 B：$F0_{range}$，PW_{range}
- 噺家 C：PW_{range}

残りのパラメータの中で最も判別率の高くなった組み合わせを表 2.4-2.6 に示す．以下，A 群：喜び・驚き・怒り，B 群：平静・恐れ・あきれとする．なお，○は採用したパラメータ，– は採用しなかったパラメータとする．

表 2.4: 採用したパラメータ (噺家 A)

感情群	$f0_{max}$	$f0_{min}$	$f0_{ave.}$	$f0_{分散}$	pw_{max}	pw_{min}	$pw_{ave.}$	pw_{range}	$pw_{分散}$	発話速度
A 群	-	-	-	-	○	○	-	-	-	○
B 群	-	-	-	-	○	○	-	-	-	○
喜び	-	○	○	-	-	-	○	○	○	-
驚き	○	-	-	-	-	-	-	-	-	-
怒り	-	-	○	-	-	-	○	○	○	○
平静	-	-	-	-	-	○	-	-	-	-
恐れ	-	-	-	-	-	-	-	-	-	○
あきれ	○	○	-	-	-	○	-	-	-	-

表 2.5: 採用したパラメータ (噺家 B)

感情群	$f0_{max}$	$f0_{min}$	$f0_{ave.}$	$f0_{分散}$	pw_{max}	pw_{min}	$pw_{ave.}$	$pw_{分散}$	発話速度
A 群	-	-	-	-	-	-	-	○	-
B 群	-	-	-	-	-	-	-	○	-
喜び	-	○	○	-	-	-	-	-	○
驚き	-	-	○	○	-	-	○	-	-
怒り	-	○	-	○	-	-	○	-	○
平静	○	-	○	○	-	-	○	-	-
恐れ	○	-	○	○	-	-	○	○	○
あきれ	-	-	-	○	○	-	-	○	-

表 2.6: 採用したパラメータ (噺家 C)

感情群	$f0_{max}$	$f0_{min}$	$f0_{ave.}$	$f0_{range}$	$f0_{分散}$	pw_{max}	pw_{min}	$pw_{ave.}$	$pw_{分散}$	発話速度
A 群	○	○	-	-	○	-	-	○	○	○
B 群	○	○	-	-	○	-	-	○	○	○
喜び	○	○	○	○	-	-	-	○	○	-
驚き	○	○	○	-	-	-	-	-	○	-
怒り	-	○	○	-	-	-	-	○	○	-
平静	-	○	-	-	○	-	-	○	-	-
恐れ	○	-	-	-	○	-	-	-	-	-
あきれ	○	-	-	-	-	-	-	-	○	-

(b) マハラノビスの汎距離による判別分析

線形判別関数による判別分析の際と同様に，各噺家ごとに取り除いたパラメータは次のようになる．

- 噺家 A：$F0_{ave.}$，$F0_{range}$，PW_{range}，$PW_{分散}$
- 噺家 B：$F0_{min}$，$F0_{range}$，$F0_{分散}$，PW_{range}
- 噺家 C：$F0_{range}$，PW_{max}，PW_{range}

6. アクセント句単位の分析で実際に使用したパラメータ

各噺家ごとにはじめに取り除いたパラメータは次のようになる．

- 文頭のアクセント句を用いた分析
  - 噺家 A：$F0_{max}$，PW_{max}
  - 噺家 B：$F0_{max}$，PW_{max}
  - 噺家 C：$F0_{max}$，PW_{max}
- 文末のアクセント句を用いた分析
  - 噺家 A：$F0_{max}$，PW_{max}
  - 噺家 B：PW_{max}
  - 噺家 C：$F0_{max}$，PW_{max}

2.1.4 判別結果

発話単位全体の判別結果

1. 線形判別関数による判別結果

全体のデータから，目的変数を各感情 (または感情群) とそれ以外という 2 グループにして，判別分析を行った結果を表 2.7, 2.8 に示す．

表 2.7: 全体のデータから 6 感情に分けた際の判別的中率 [%]

	データ全体に対する判別的中率			各感情データ内での判別的中率		
	噺家 A	噺家 B	噺家 C	噺家 A	噺家 B	噺家 C
喜び	93.8	95.2	95.5	0	0	0
驚き	93.8	98.4	94.6	16.7	50.0	0
怒り	82.5	84.1	82.1	31.6	30.8	36.0
平静	71.3	73.0	78.6	66.7	85.7	82.5
恐れ	91.3	85.7	94.6	0	0	14.3
あきれ	91.3	96.8	89.3	0	0	0

表 2.8: グルーピング後の判別的中率 [%]

	データ全体に対する判別率			各感情データ内での判別率		
	噺家 A	噺家 B	噺家 C	噺家 A	噺家 B	噺家 C
A 群	75.0	73.0	74.1	46.7	21.1	41.7
B 群	75.0	73.0	74.1	92.0	95.6	86.8
喜び	90.0	88.9	97.2	60.0	33.3	80
驚き	80.0	94.4	91.7	0	50.0	66.7
怒り	93.3	83.3	86.1	100	92.3	96
平静	72.0	88.9	81.6	100	100	96.5
恐れ	86.0	88.9	94.7	0	37.5	42.9
あきれ	86.0	97.8	86.8	0	50.0	16.7

2. マハラノビスの汎距離による判別結果

同様に，全体のデータから，目的変数を各感情 (または感情群) とそれ以外という 2 グループにして，判別分析を行った結果は表 2.9,2.13 のようになる．また，これらの各感情が実際にどの感情へと誤判別されていたかを調べるために，全体のデータを直接 6 グループ (グルーピング後は 3 グループ) に分けた場合の値をそれぞれ表 2.10-2.12,2.14,2.15 に示す．

表 2.9: 全体のデータから 6 感情に分けた際の判別的中率 [%]

	データ全体に対する判別率			各感情データ内での判別率		
	噺家 A	噺家 B	噺家 C	噺家 A	噺家 B	噺家 C
喜び	82.5	77.8	92.9	80.0	66.7	100
驚き	87.5	96.8	82.1	83.3	50.0	83.3
怒り	75.0	74.6	81.3	63.2	53.8	76.0
平静	70.0	69.8	75.9	77.8	82.9	82.5
恐れ	68.8	68.3	88.4	85.7	50.0	57.1
あきれ	77.5	88.9	67.0	100	100	75.0

表 2.10: 噺家 A 各感情への判別率 [%]

		判別された感情					
		喜び	驚き	怒り	平静	恐れ	あきれ
真の感情	喜び	60	20	0	20	0	0
	驚き	0	50	0	16.7	0	33.3
	怒り	5.3	21.1	31.6	15.8	21.1	5.3
	平静	13.9	0	8.3	11.1	44.4	22.2
	恐れ	0	0	14.3	14.3	71.4	0
	あきれ	0	0	0	0	28.6	71.4

表 2.11: 噺家 B 各感情への判別率 [%]

		判別された感情					
		喜び	驚き	怒り	平静	恐れ	あきれ
真の感情	喜び	66.7	0	0	33.3	0	0
	驚き	0	50	0	50	0	0
	怒り	15.4	0	30.8	0	30.8	23.1
	平静	17.1	2.9	5.7	51.4	20	2.9
	恐れ	12.5	0	0	37.5	25	25
	あきれ	0	0	0	0	0	100

表 2.12: 噺家 C 各感情への判別率 [%]

		判別された感情					
		喜び	驚き	怒り	平静	恐れ	あきれ
真の感情	喜び	100	0	0	0	0	0
	驚き	0	66.7	0	0	16.7	16.7
	怒り	0	8	64	12	4	12
	平静	8.8	10.5	5.3	52.6	5.3	17.5
	恐れ	14.3	0	0	0	57.1	28.6
	あきれ	0	0	16.7	25	0	58.3

表 2.13: グルーピング後の判別的中率 [%]

	データ全体に対する判別率			各感情データ内での判別率		
	噺家 A	噺家 B	噺家 C	噺家 A	噺家 B	噺家 C
A 群	76.2	76.2	78.6	63.3	55.6	75.0
B 群	76.2	76.2	78.6	84.0	84.4	80.3
喜び	83.3	83.3	100	80.0	100	100
驚き	73.3	94.4	94.4	83.3	50.0	100.
怒り	83.3	88.9	86.1	89.5	84.6	92.0
平静	62.0	80.0	78.9	52.8	82.9	82.5
恐れ	58.0	66.7	86.8	71.4	50.0	57.1
あきれ	74.0	100	77.6	71.4	100	66.7

表 2.14: グルーピング後の各感情への判別率 [%]

			判別された感情		
			喜び	驚き	怒り
真の感情	噺家 A	喜び	80.0	0	20.0
		驚き	0	100	0
		怒り	5.3	15.8	78.9
	噺家 B	喜び	100	0	0
		驚き	50.0	50.0	0
		怒り	15.4	0	84.6
	噺家 C	喜び	100	0	0
		驚き	0	100	0
		怒り	0	8.0	92.0

表 2.15: グルーピング後の各感情への判別率 [%]

			判別された感情		
			平静	恐れ	あきれ
真の感情	噺家 A	平静	33.3	44.4	22.2
		恐れ	14.3	71.4	14.3
		あきれ	14.3	14.3	71.4
	噺家 B	平静	74.3	25.7	0
		恐れ	25.0	75.0	0
		あきれ	0	0	100
	噺家 C	平静	71.9	10.5	17.5
		恐れ	28.6	57.1	14.3
		あきれ	33.3	0	66.7

文頭アクセント句の判別結果

次に，発話文頭のアクセント句をデータとして目的変数を各感情 (または感情群) とそれ以外という 2 グループにして，判別分析を行った結果を表 2.16,2.20 に示す．また，これらの各感情が実際にどの感情へと誤判別されていたかを調べるために，全体のデータを直接 6 グループ (グルーピング後は 3 グループ) に分けた場合の値をそれぞれ表 2.17-2.19,2.21,2.22 に示す．尚ここでの結果は全てマハラノビスの汎距離による判別結果による識別結果である．

表 2.16: 全体のデータから 6 感情に分けた際の判別的中率 [%]

	データ全体に対する判別率			各感情データ内での判別率		
	噺家 A	噺家 B	噺家 C	噺家 A	噺家 B	噺家 C
喜び	84.8	84.1	77.7	60.0	66.7	100
驚き	87.3	85.7	77.7	66.7	50.0	83.3
怒り	63.3	71.4	69.6	52.6	53.8	72.0
平静	57.0	61.9	68.8	68.6	74.3	75.4
恐れ	64.6	60.3	80.4	57.1	37.5	57.1
あきれ	78.5	92.1	68.8	100	50.0	83.3

表 2.17: 噺家 A 各感情への判別率 [%]

		判別された感情					
		喜び	驚き	怒り	平静	恐れ	あきれ
真の感情	喜び	60.0	0	20.0	20.0	0	0
	驚き	16.7	50	0	16.7	16.7	0
	怒り	5.3	15.8	31.6	21.1	15.8	10.5
	平静	11.4	2.9	20.0	31.4	22.9	11.4
	恐れ	0	0	0	28.6	57.1	14.3
	あきれ	0	0	0	0	14.3	85.7

表 2.18: 噺家 B 各感情への判別率 [%]

		判別された感情					
		喜び	驚き	怒り	平静	恐れ	あきれ
真の感情	喜び	66.7	0	0	33.3	0	0
	驚き	0	50.0	0	50.0	0	0
	怒り	0	7.7	38.5	23.1	23.1	7.7
	平静	17.1	5.7	14.3	34.3	25.7	2.9
	恐れ	25.0	12.5	12.5	37.5	12.5	0
	あきれ	0	0	0	50.0	0	50.0

表 2.19: 噺家 C 各感情への判別率 [%]

		判別された感情					
		喜び	驚き	怒り	平静	恐れ	あきれ
真の感情	喜び	60.0	0	0	20.0	0	20.0
	驚き	0	50.0	0	0	33.3	16.7
	怒り	4.0	28.0	16.0	12.0	24.0	16.0
	平静	19.3	12.3	1.8	49.1	5.3	12.3
	恐れ	14.3	0	0	14.3	57.1	14.3
	あきれ	0	0	8.3	25.0	16.7	50.0

表 2.20: グルーピング後の判別的中率 [%]

	データ全体に対する判別率			各感情データ内での判別率		
	噺家 A	噺家 B	噺家 C	噺家 A	噺家 B	噺家 C
A 群	67.1	60.3	71.4	71.4	50.0	69.4
B 群	67.1	60.3	71.4	60.0	64.4	72.4
喜び	76.7	88.9	94.4	60.0	66.7	100
驚き	86.7	77.8	77.8	100	50.0	88.3
怒り	73.3	83.3	66.7	73.7	84.6	64.0
平静	69.4	62.2	77.6	74.3	68.6	80.7
恐れ	67.3	57.8	85.5	42.9	37.5	57.1
あきれ	83.7	95.6	73.7	100	50.0	83.3

表 2.21: グルーピング後の各感情への判別率 [%]

			判別された感情		
			喜び	驚き	怒り
真の感情	噺家 A	喜び	40.0	20.0	40.0
		驚き	16.7	83.3	0
		怒り	15.8	15.8	68.4
	噺家 B	喜び	66.7	0	33.3
		驚き	0	50.0	50.0
		怒り	0	23.1	76.9
	噺家 C	喜び	100	0	0
		驚き	0	83.3	16.7
		怒り	4.0	24.0	72.0

表 2.22: グルーピング後の各感情への判別率 [%]

			判別された感情		
			平静	恐れ	あきれ
真の感情	噺家 A	平静	61.8	23.5	14.7
		恐れ	37.5	50.0	12.5
		あきれ	0	0	100
	噺家 B	平静	62.9	34.3	2.9
		恐れ	62.5	37.5	0
		あきれ	50.0	0	50.0
	噺家 C	平静	66.7	10.5	22.8
		恐れ	42.9	57.1	0
		あきれ	16.7	16.7	66.7

文末アクセント句の判別結果

更に，発話文末のアクセント句をデータとして目的変数を各感情 (または感情群) とそれ以外という 2 グループにして，判別分析を行った結果を表 2.23,2.27 に示す．また，これらの各感情が実際にどの感情へと誤判別されていたかをされていたかを調べるために，全体のデータを直接 6 グループ (グルーピング後は 3 グループ) に分けた場合の値をそれぞれ表 2.24-2.26,2.28,2.29 に示す．尚ここでの結果は全てマハラノビスの汎距離による判別結果による識別結果である．

表 2.23: 全体のデータから 6 感情に分けた際の判別的中率 [%]

	データ全体に対する判別率			各感情データ内での判別率		
	噺家 A	噺家 B	噺家 C	噺家 A	噺家 B	噺家 C
喜び	81.0	84.1	77.7	60.0	66.7	80.0
驚き	84.8	90.5	85.7	66.7	50.0	83.3
怒り	79.7	84.1	70.5	52.6	69.2	68.0
平静	67.1	73.0	68.8	77.1	82.9	77.2
恐れ	82.3	73.0	87.5	28.6	50.0	71.4
あきれ	78.5	76.2	69.6	71.4	100	50.0

表 2.24: 噺家 A 各感情への判別率 [%]

		判別された感情					
		喜び	驚き	怒り	平静	恐れ	あきれ
真の感情	喜び	60.0	0	20.0	0	20.0	0
	驚き	16.7	16.7	16.7	33.3	0	16.7
	怒り	5.3	10.5	42.1	31.6	0	10.5
	平静	11.4	5.7	8.6	62.9	2.9	8.6
	恐れ	0	0	0	71.4	28.6	0
	あきれ	0	0	0	42.9	0	57.1

表 2.25: 噺家 B 各感情への判別率 [%]

		判別された感情					
		喜び	驚き	怒り	平静	恐れ	あきれ
真の感情	喜び	66.7	0	0	33.3	0	0
	驚き	50.0	50.0	0	0	0	0
	怒り	23.1	0	61.5	7.7	0	7.7
	平静	8.6	8.6	2.9	51.4	8.6	20.0
	恐れ	0	12.5	0	50.0	37.5	0
	あきれ	0	0	0	0	0	100

表 2.26: 噺家 C 各感情への判別率 [%]

		判別された感情					
		喜び	驚き	怒り	平静	恐れ	あきれ
真の感情	喜び	80.0	0	0	20.0	0	0
	驚き	0	66.7	0	16.7	16.7	0
	怒り	4.0	8.0	48.0	20.0	8.0	12.0
	平静	17.5	3.5	17.5	49.1	10.5	1.8
	恐れ	0	0	0	14.3	85.7	0
	あきれ	16.7	8.3	25.0	25.0	0	25.0

表 2.27: グルーピング後の判別的中率 [%]

	データ全体に対する判別率			各感情データ内での判別率		
	噺家 A	噺家 B	噺家 C	噺家 A	噺家 B	噺家 C
A 群	75.9	77.8	68.8	79.6	61.1	69.4
B 群	75.9	77.8	68.8	70.0	84.4	68.4
喜び	86.7	77.8	77.8	80.0	100	80.0
驚き	73.3	94.4	88.9	66.7	50.0	83.3
怒り	76.7	83.3	88.9	78.9	84.6	88.0
平静	67.3	73.3	73.7	77.1	82.9	77.2
恐れ	77.6	73.3	88.2	42.9	37.5	85.7
あきれ	79.6	82.2	77.6	57.1	100	50.0

表 2.28: グルーピング後の各感情への判別率 [%]

			判別された感情		
			喜び	驚き	怒り
真の感情	噺家 A	喜び	80.0	0	20.0
		驚き	16.7	50.0	33.3
		怒り	5.3	15.8	78.9
	噺家 B	喜び	100	0	0
		驚き	50.0	50.0	0
		怒り	23.1	0	76.9
	噺家 C	喜び	80.0	0	20.0
		驚き	16.7	83.3	0
		怒り	16.0	8.0	76.9

表 2.29: グルーピング後の各感情への判別率 [%]

			判別された感情		
			平静	恐れ	あきれ
真の感情	噺家 A	平静	65.7	20.0	14.3
		恐れ	57.1	42.9	0
		あきれ	42.9	0	57.1
	噺家 B	平静	60.0	20.0	20.0
		恐れ	50.0	37.5	12.5
		あきれ	0	0	100
	噺家 C	平静	73.7	14.0	12.3
		恐れ	14.3	85.7	0
		あきれ	58.3	0	41.7

判別係数の影響度合い

次に，どのようなパラメータが判別の際に効果的であるのかを知るために判別関数の係数を調べた．噺家別の判別係数を図 2.4-2.12 に示す．この図では判別係数ごとの相対関係の値が表示されているので，噺家ごとにそれぞれ絶対値が大きいものほど判別への寄与率が高いことを表している．また，図中の $-a$ は全体のデータから直接 6 感情に分けた際の判別係数を，$-b$ は一度グルーピングした後に 6 感情に分けた際の判別係数を示す．尚，2 群分けとは，全体のデータから 2 つのグループに分けた場合のこと

である.

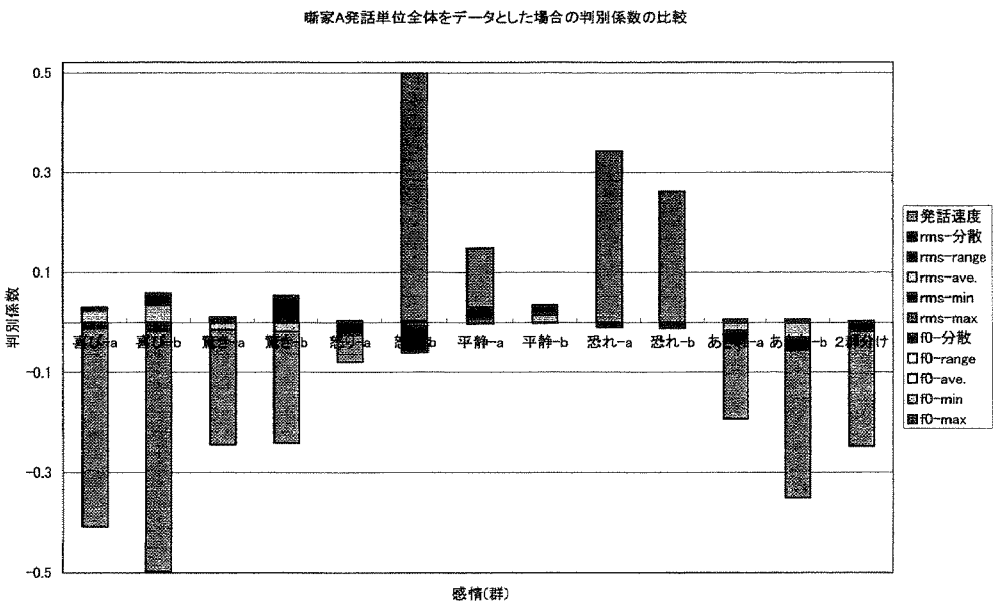


図 2.4: 噺家 A 発話単位全体をデータとした場合の判別係数の比較

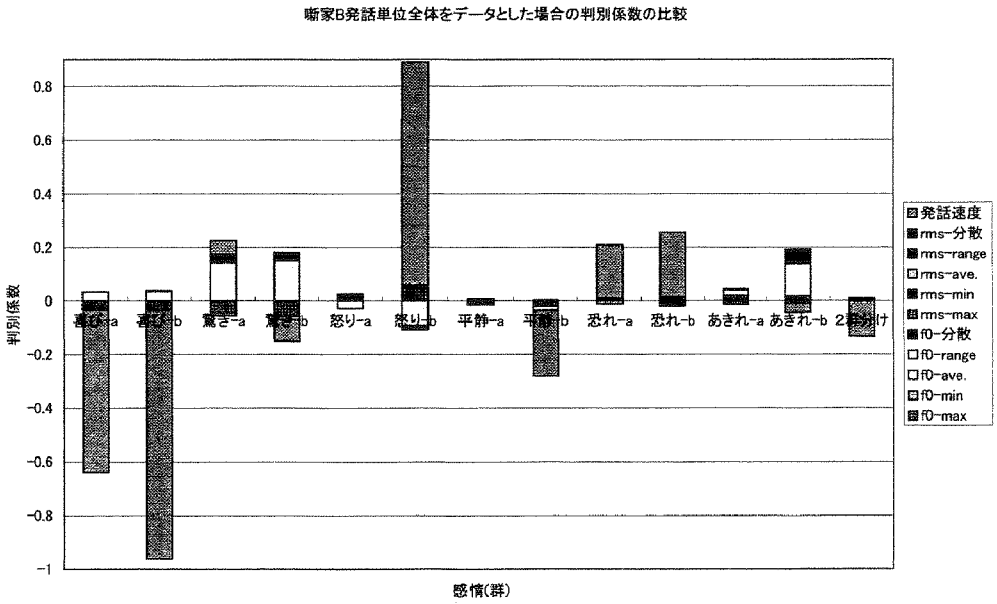


図 2.5: 噺家 B 発話単位全体をデータとした場合の判別係数の比較

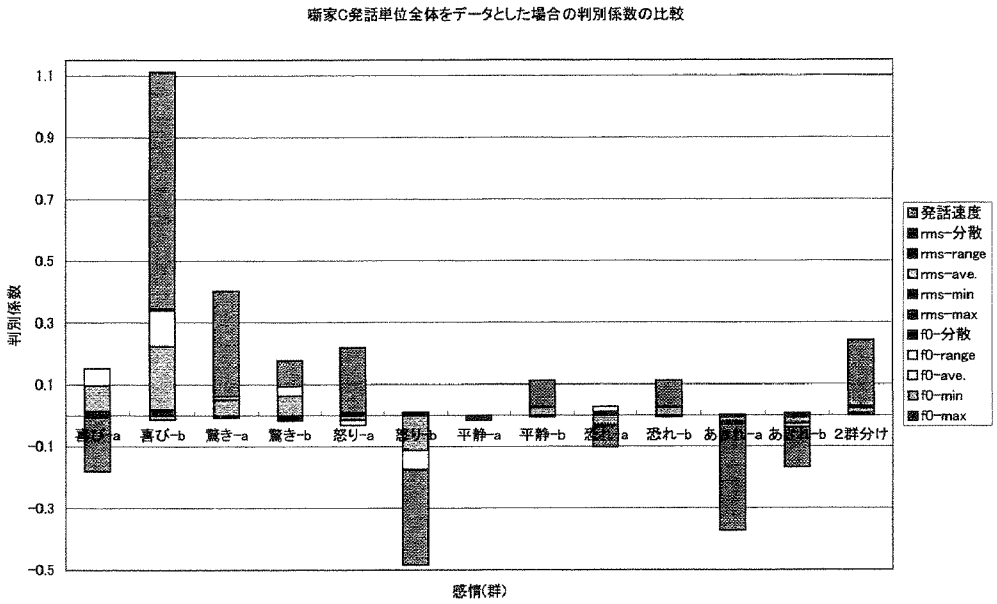


図 2.6: 噺家 C 発話単位全体をデータとした場合の判別係数の比較

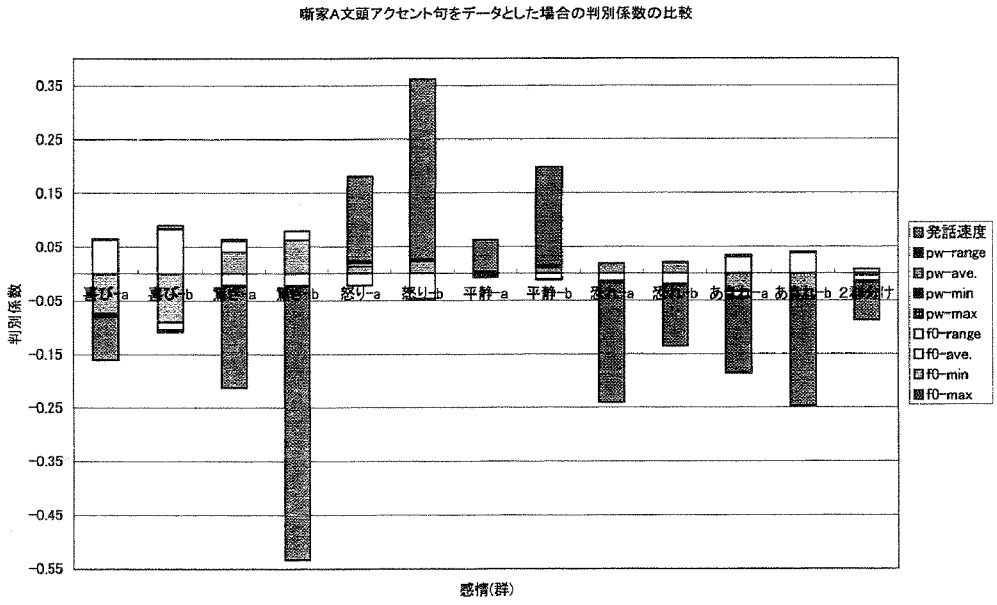


図 2.7: 噺家 A 文頭アクセント句をデータとした場合の判別係数の比較

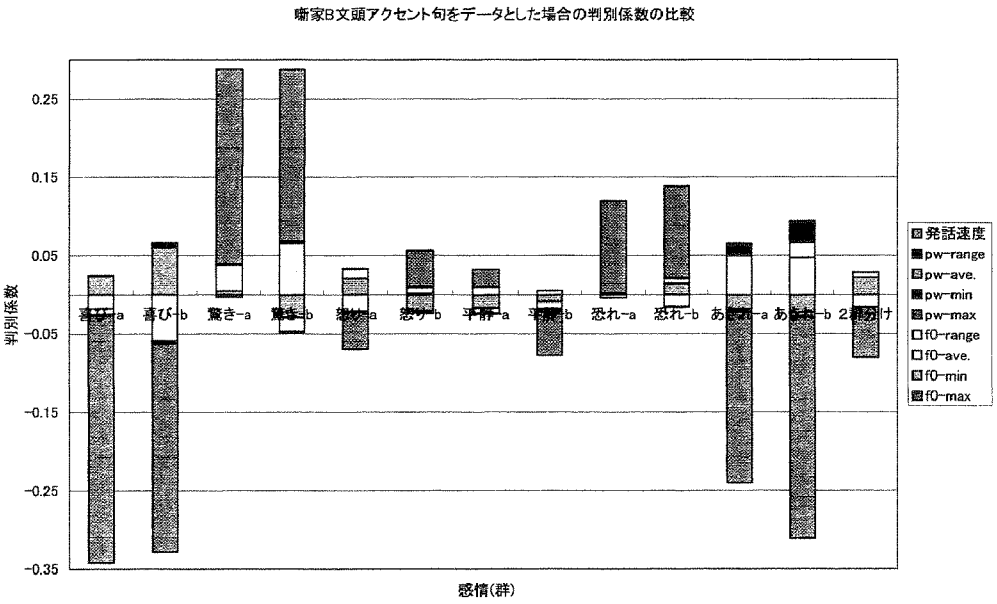


図 2.8: 噺家 B 文頭アクセント句をデータとした場合の判別係数の比較

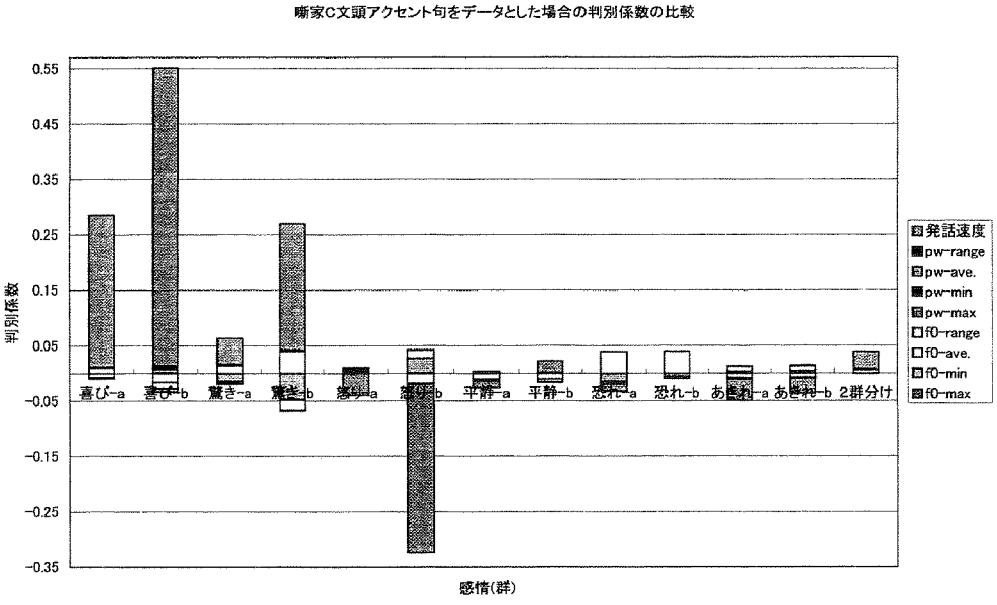


図 2.9: 噺家 C 文頭アクセント句をデータとした場合の判別係数の比較

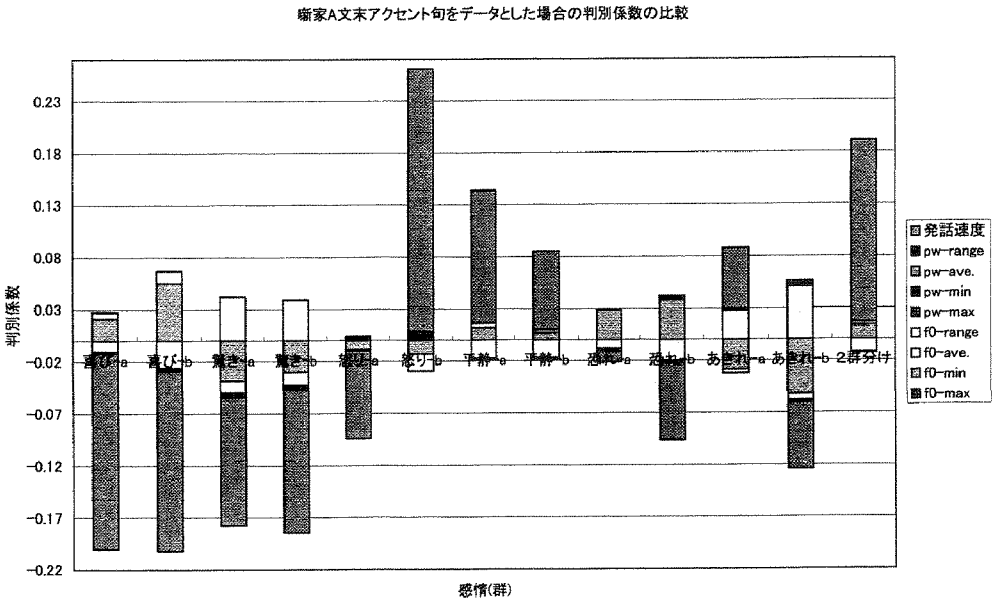


図 2.10: 噺家 A 文末アクセント句をデータとした場合の判別係数の比較

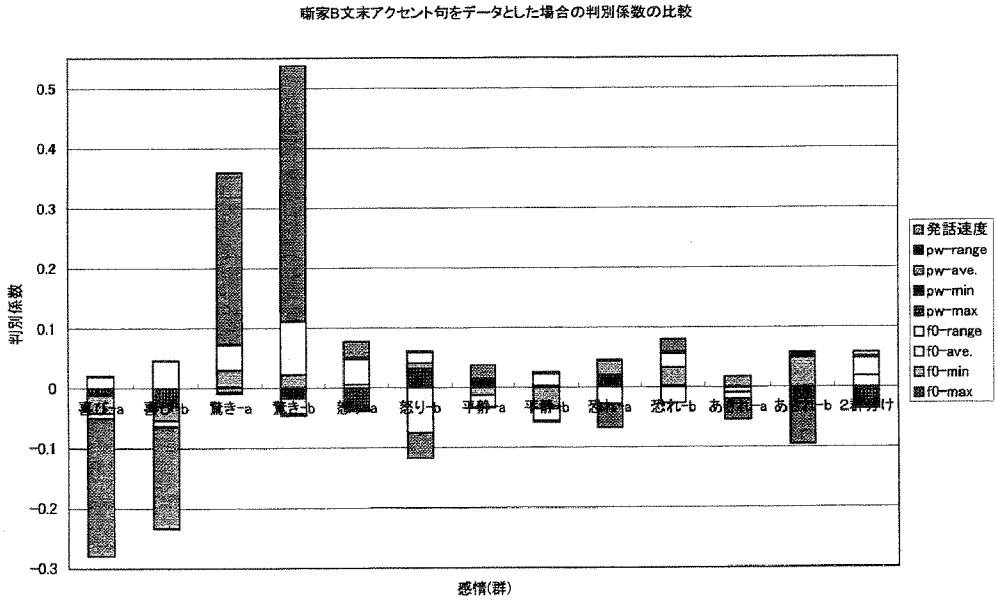


図 2.11: 噺家 B 文末アクセント句をデータとした場合の判別係数の比較

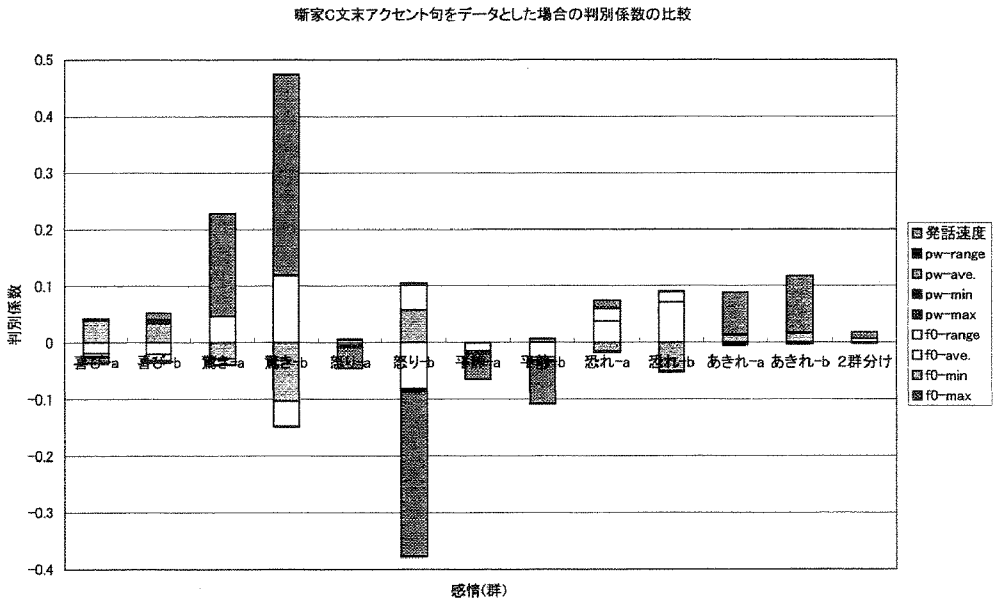


図 2.12: 噺家 C 文末アクセント句をデータとした場合の判別係数の比較

3 名ともほぼ全ての感情が主に発話速度を主要なパラメータにとるようである．そこで，噺家 A,B,C それぞれの各感情ごとに発話速度の次に有効なパラメータを調べた．結果を表 2.30,2.31 に示す．ただし，発話速度が最も有効なパラメータでない場合は，最も有効なパラメータを示す．

表 2.30: 各感情に対して効果的なパラメータ (グルーピングなしの場合)

		喜び	驚き	怒り	平静	恐れ	あきれ
噺家 A	発話単位	$F0_{min}$	$F0_{min}$	PW_{min}	PW_{min}	$F0_{max}$	PW_{min}
	文頭	$F0_{min}$	$F0_{min}$	$F0_{ave}$	PW_{min}	$F0_{min}$	$F0_{min}$
	文末	$F0_{min}$	$F0_{min}$	$F0_{min}$	$F0_{ave}$	$F0_{min}$	$F0_{min}$
噺家 B	発話単位	$F0_{ave}$	$F0_{ave}$	$F0_{ave}$	$F0_{max}$	$F0_{max}$	$F0_{max}$
	文頭	$F0_{min}$	$F0_{ave}$	$F0_{ave}$	$F0_{min}$	$F0_{ave}$	$F0_{ave}$
	文末	$F0_{min}$	$F0_{ave}$	$F0_{range}$	$F0_{range}$	$F0_{ave}$	$F0_{min}$
噺家 C	発話単位	$F0_{min}$	$F0_{min}$	$F0_{min}$	$F0_{ave}$	$F0_{min}$	$F0_{min}$
	文頭	$F0_{ave}$	$F0_{min}$	$F0_{min}$	$F0_{ave}$	$F0_{ave}$	$F0_{ave}$
	文末	$F0_{min}$	$F0_{ave}$	$F0_{min}$	$F0_{ave}$	$F0_{ave}$	$F0_{min}$

表 2.31: 各感情に対して効果的なパラメータ (グルーピングありの場合)

		喜び	驚き	怒り	平静	恐れ	あきれ	2群分け
噺家A	発話単位	$F0_{min}$	PW_{min}	PW_{min}	PW_{min}	PW_{min}	$F0_{min}$	PW_{min}
	文頭	$F0_{min}$	$F0_{min}$	$F0_{ave}$	$F0_{min}$	$F0_{min}$	$F0_{min}$	$F0_{range}$
	文末	$F0_{min}$	$F0_{ave}$	$F0_{ave}$	$F0_{ave}$	$F0_{min}$	$F0_{min}$	$F0_{min}$
噺家B	発話単位	$F0_{ave}$	$F0_{ave}$	$F0_{ave}$	$F0_{max}$	$F0_{max}$	$F0_{ave}$	$F0_{ave}$
	文頭	$F0_{ave}$	$F0_{ave}$	$F0_{min}$	$F0_{min}$	$F0_{min}$	$F0_{ave}$	$F0_{min}$
	文末	$F0_{ave}$	$F0_{ave}$	$F0_{ave}$	$F0_{min}$	$F0_{min}$	$F0_{min}$	$F0_{max}$
噺家C	発話単位	$F0_{min}$	$F0_{min}$	$F0_{min}$	$F0_{min}$	$F0_{min}$	$F0_{min}$	$F0_{min}$
	文頭	$F0_{ave}$	$F0_{min}$	$F0_{min}$	$F0_{ave}$	$F0_{ave}$	$F0_{ave}$	$F0_{ave}$
	文末	$F0_{min}$	$F0_{ave}$	$F0_{ave}$	$F0_{ave}$	$F0_{ave}$	$F0_{min}$	$F0_{ave}$

分析結果のまとめ

ここまでで得られた結果を次に示す.

1. 発話単位全体の判別結果

   - 線形判別関数における判別では，データ全体に対する判別的中率は73〜98%と高い数値だが各感情内での判別的中率は0%の場合が多い．特に喜びとあきれの感情は3名とも0%である.

   - マハラノビスの汎距離による判別では，グルーピング前よりグルーピングした後の方が各感情データ内での判別的中率はほぼ向上する．しかし，噺家Aに関しては，平静・恐れ・あきれの感情でそれぞれかなり低下している.

   - A群，B群の2群に分ける際のデータ全体に対する判別的中率は3名ともに76〜78%程度である.

   - 2群から各感情への判別的中率は58〜100%とばらつきがある.

   - 噺家Bは，平静とあきれの感情の相関が高いようである.

2. 文頭アクセント句の判別結果

   - 3名ともグルーピング前に比べグルーピング後の判別的中率が，データ全体に対しても各感情データ内においてもほぼ向上している.

3. 文末アクセント句の判別結果

- 3名ともグルーピング前に比べグルーピング後の判別的中率が，データ全体に対してはほぼ向上している．

- 噺家A,Bにおいて，恐れの感情判別的中率が非常に悪い．

4. 判別係数の影響度合い

3名とも主に発話速度が効果的である．それ以外の有効なパラメータは次のようになる．

- 噺家Aは，文頭文末のアクセント句に$F0_{min}$,$F0_{ave}$が効いており，発話単位全体ではPW_{min},$F0_{min}$が効いている．
- 噺家Bは，文頭文末のアクセント句に$F0_{ave}$,$F0_{min}$,PW_{min}が効いており，発話単位全体では$F0_{ave}$,PW_{min},$F0_{max}$が効いている．
- 噺家Cは文頭文末のアクセント句に$F0_{ave}$,$F0_{min}$が効いており，発話単位全体では$F0_{min}$,$F0_{ave}$が効いている．

2.1.5 結論

今回の実験では噺家3名について基本周波数・パワー・発話速度をパラメータとし，発話全体・文頭のアクセント句・文末のアクセント句それぞれに対して，喜び・驚き・怒り・平静・恐れ・あきれの6つの感情の判別を行った．その結果より以下の知見が得られた．

データ全体を2群に分別する際には発話単位での分析の方が有効であるが，各感情へ細かく分ける際には文末のアクセント句に注目する方が比較的良い結果を得られている．つまり，判別結果を総合的に見ると感情表現は発話の最後の部分に比較的現れやすいのではないかと推測できる．

また，2群へのグルーピングをする場合はしない場合に比べると各感情データ内の判別率は多少の向上が見られる．

各感情相互の誤判別の起こりやすさは，平静・恐れ，喜び・驚き，驚き・怒り間に多いようである．つまり，これらは特徴が類似しているものと思われる．

なお，門谷らの研究[7]と比較すると本研究は判別率が多少低下するが，噺家という感情表現のプロの音声を使用したということで

- 噺家の演技が非常に自然な感情表現であるために感情ごとの顕著な特徴というものがない
- 噺家の感情表現が豊かということで各人が何通りもの表現方法を持っている

というような可能性が考えられる．

さらに，3 名ともほぼ全ての感情が主に発話速度を主要なパラメータにとるようである．また表 2.30,2.31 より，それぞれにある程度有効なパラメータは見られるが，感情ごとの 3 人共通した傾向というものは顕著には現れていない．

最後に，韻律的な特徴としてはパワーはあまり有効性がないものと思われる．

2.2 「笑い声」合成のための音響特徴の分析

2.2.1 目的

人と人とのコミュニケーションにおいて感情音声は欠くことのできないものであるため，対話システムやCGキャラクターに人間らしさを表現させる際，感情音声の合成は当然の要求として求められる．また，CGアニメーションの作成は，専門家だけではなく一般にも普及し始めてきている一方で，その複雑さから感情音声との併用は未だ課題となっている．従って，今後はより容易な感情音声の合成技術が要求されると考える．

現在，音声合成の研究は盛んに行われており，音声に含まれる感情に関する研究もなされてきた [9][10]．従来では,「喜び」や「怒り」,「疑い」といった感情を含む対話音声を合成することが研究され，ある程度実現されてきた [11]．本研究では，生理的な発声による感情表現に着目することで人間らしさを実現するという観点から,「笑い」に注目した．

本研究では，規則合成方式を用いて「笑い声」及び笑いながら喋る「喋り笑い」のいろいろな音声を自動的に合成することを目標としている．そのために,「笑い声」を分析して音声の合成規則や制御パラメータを明らかにすることを目下の研究目的とする．なお，本合成システムを構築する際に，ユーザインタフェースの操作項目として直感的に分かりやすい操作項目を用意することも検討する．

本研究では，一般的で汎用性の高い「笑い声」の合成を目指すため，日常生活で出現頻度が高いと考えられる/ha/を基本とした笑い声を分析対象とする．分析する特徴パラメータとしては，音声パワー，ピッチ周波数，有声音・無声音の程度，スペクトル(音色)，音素タイミングを用いた．

まず，ユーザの要求に合わせた合成を行う為には,「笑い声」を単純に分類する必要がある．本研究では，大原ら [12] の行った「大爆笑」や「冷笑」といった感情での分類ではなく，その前段階として音の大小変化や高低変化といった聴覚的に分かり易い特徴からアプローチし,「笑い声」の特徴を確認していくだけでなく，ユーザインタフェースの操作項目になり得る要素を検討していく．さらに，その分類ごとに音響特徴を分析することで，分類を構成する要因を特定していく．

また,「笑い声」特有の音響特徴を知るためには,「通常発話」との比較が有効と考えられる．ここでは，ピッチ周波数，スペクトル，音素タイミングといった音響特徴を数値等によって比較することで定量的な分析を行った．

最後に，音響パラメータを必要に応じて操作して作成した合成音に対して主観評価実験を行い,「笑い声」の笑い声らしさに対する各音響パラメータの貢献度を調べた．

2.2.2 音響特徴の抽出

音声収録

対話においてどのような笑い声が発声されるかの調査及び笑い声の音響特徴量の分析に使用するため，男子大学生7人を対象に音声収録を行った．収録内容は2名または3名の被験者による30分から1時間程度の自然対話であり，収録データのサンプリング周波数は16kHzである．尚，収録にはヘッドセットマイクを用いた．これにより，話者の可動性・自由度を向上させ，より自然な対話を促すことができ，同時にマイクと発話者の口との距離が常に一定に保たれ，収録音声のパワー変化が純粋に発話の抑揚のみによって決まることになる．

次に，収録した対話音声から「笑い声」及び「喋り笑い」の抽出を行った．収録データからの笑い声の抽出数を表2.32に示す．上段には/ha/を基本とした笑い声の総数，下段には/ha/以外を基本とした笑い声の総数を示した．笑い声には，主に/ha/，/hi/，/fu/，/he/，/ho/を基本とした笑い声が存在すると考えられるが，表2.32より，どの音素を基本とした笑い声を多く発声するかには個人差があることがわかる．話者Eの場合，/fu/を基本とした笑い声の方が多く，この傾向が顕著に現れる．しかし，全体でみればほぼ半数が/ha/を基本とした笑い声であることがわかり，/ha/を基本とした笑い声が一般的な笑い声であることがわかる．一方で，収録音声には/ha/とも/fu/とも判別できない発音の笑い声も多数含まれており，今後はこれらを分類することも必要になると考える．また，/ha/を基本とした笑い声のうち，7割は「は」を2回または3回発声するものであった．

表 2.32: 話者別の笑い声の抽出数

	話者							合計
	A	B	C	D	E	F	G	
/ha/からなる笑い	26	31	38	29	9	13	11	157
その他の笑い	17	26	38	9	43	25	9	167

音素ラベリング

収録音声は音素ラベリングにより，無音区間である“Pause”，“音素/h/”，“音素/a/”に分けられる．この際，図2.13に示すように，Pauseと/h/の境界はパワーの立ち上がり開始点，/h/と/a/の境界は有声・無声の変移開始点，/a/とPauseの境界はパワーの減衰終了点としている．

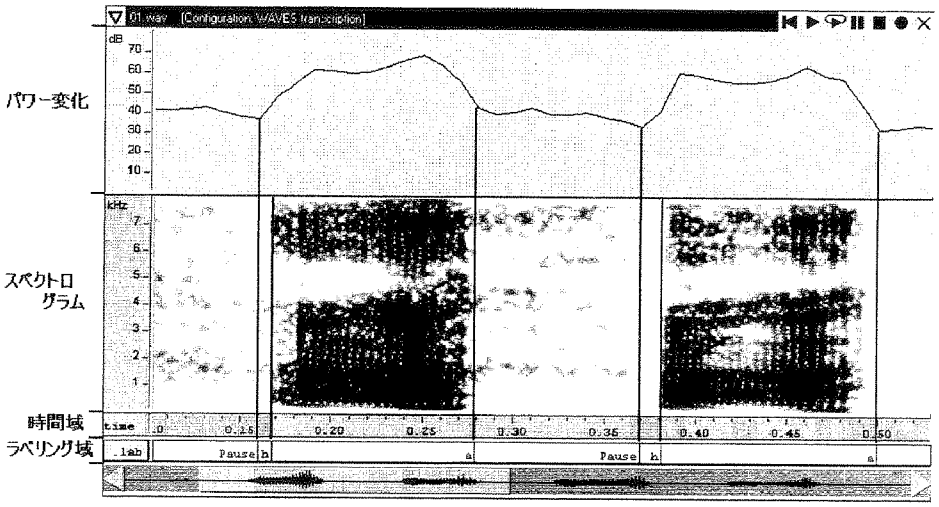


図 2.13: 音素ラベリング

音声パワー・ピッチ周波数の時間変化

「笑い声」に特有な音声パワー及びピッチ周波数の時間変化パターンを統計的に調べた．また，音声パワーやピッチ周波数といったパラメータは聴覚的にわかりやすい特徴であるので，分析の結果パターンを分類できれば，それが「笑い声」の合成をシステム化する際のユーザインタフェースの操作項目になり得る．

音声パワー及びピッチ周波数の時間変化パターンの出現割合を，/ha/の連続回数別に表2.33，2.34に示す．変化パターンの種類は以下の通りである．ただし，音声パワーなら3[dB]，ピッチ周波数なら15[Hz]の差異が直前の音節との間にあったときに“変化した”と判断するものとする．

a 下り調子…全体的に減少変化する
b 上り調子…全体的に増加変化する
c 上って下る…増加変化の後，減少変化する
d 下って上る…減少変化の後，増加変化する
e 上下過変化…増加と減少を繰り返す
f ほとんど不変…大きな増減が無い

表2.33より，音声パワーは全体的に減少変化をするか，またはあまり変化がないことがわかる．また，発声音素数が多くなると，より下り調子のパターンが増える．下り調子の生理的要因として,「笑い声」は呼気の流出が「通常発話」より多く，また，/ha/の発声回数が増えるほど肺の呼気圧が減少することが考えられる．数値でみても音声パワーは全体的な単調減少変化をしており，これは，話し声における語調成分の下降と一致している．

また，表2.34より，ピッチ周波数は発声音素数が多くなると下り調子になるパターンの割合が多くなる傾向がみられたが，全体としては変化パターンにはバラつきが見

表 2.33: /ha/の連続回数と音声パワーの変化パターンの出現割合

		変化パターン					
		a	b	c	d	e	f
/ha/の 連続回数	2回	62.7 %	3.4 %	0.0 %	0.0 %	0.0 %	33.9 %
	3回	67.3 %	6.1 %	4.1 %	2.0 %	0.0 %	20.4 %
	4回	76.0 %	4.0 %	4.0 %	0.0 %	0.0 %	16.0 %
	5回〜	73.9 %	8.7 %	0.0 %	0.0 %	0.0 %	17.4 %

表 2.34: /ha/の連続回数とピッチ周波数の変化パターンの出現割合

		変化パターン					
		a	b	c	d	e	f
/ha/の 連続回数	2回	44.1 %	35.6 %	0.0 %	0.0 %	0.0 %	20.3 %
	3回	34.7 %	30.6 %	16.3 %	12.2 %	0.0 %	6.1 %
	4回	32.0 %	12.0 %	12.0 %	20.0 %	8.0 %	16.0 %
	5回〜	56.5 %	0.0 %	13.0 %	21.7 %	8.7 %	0.0 %

られ，様々なピッチ周波数の変化が現れるので，ピッチ周波数の変化パターンでの分類は有効であると考えられる．また，パターンが分類できれば，ユーザが合成音を作成する際，ピッチ周波数の変化パターンを操作項目として笑い声を作り分けることができる可能性がある．

「笑い声」と「通常発話」の比較

「通常発話」の音声と比較することで,「笑い声」に特有な音響特徴を分析した．

1. ピッチ周波数の比較

   「笑い声」と「通常発話」それぞれの発話におけるピッチ周波数の違いを有声部分である後続母音の音素/a/について調べた．発話者2名分のデータにおけるピッチ周波数の度数分布を図2.14に示す．

   図2.14より,「笑い声」の方が全体的に50〜100Hzほどピッチ周波数が高いという傾向があることがわかる．また,「笑い声」には，さらに周波数の高い400Hz付近にもピッチ周波数の分布が現れることがわかる．これは，笑いの発声過程で裏声が現れることに起因するものと考えられる．今回対象とした2話者に関して，この高い周波数帯における分布割合には明らかな違いが見られたが，これは笑い方に個人差があることによるものと考えられる．今後は，通常の笑い声と裏声による笑い声を区別していくことも考えている．

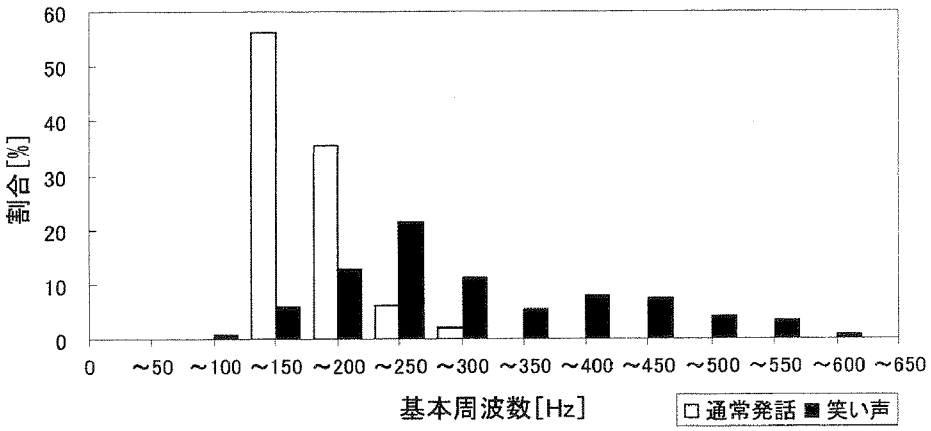


図 2.14: /a/のピッチ周波数分布

2. スペクトル包絡の比較

有声部分である後続母音の音素/a/について,「笑い声」と「通常発話」のスペクトル包絡の概形を視覚的に比較し，音色の違いを検討した．それぞれのスペクトル包絡を図 2.15，2.16 に示す．なお,「笑い声」と「通常発話」それぞれに関して，ピッチ周波数の違いによるスペクトル包絡の概形の差異が少なくなるように，ピッチ周波数が近いもの同士を比較した．

図 2.16 より，通常発話のスペクトルでは有声音であるために高調波構造がはっきり現れているが，図 2.15 の「笑い声」のスペクトルでは高調波構造が曖昧である．このように,「笑い声」のスペクトルは有声音ではなく無声音や雑音の特徴を示しているが，これは声道からの過剰な呼気の流出による影響が大きいためと考えられる．また，このことから「笑い声」は子音による影響が強く,「通常発話」と「笑い声」では音色が異なるということがわかる．

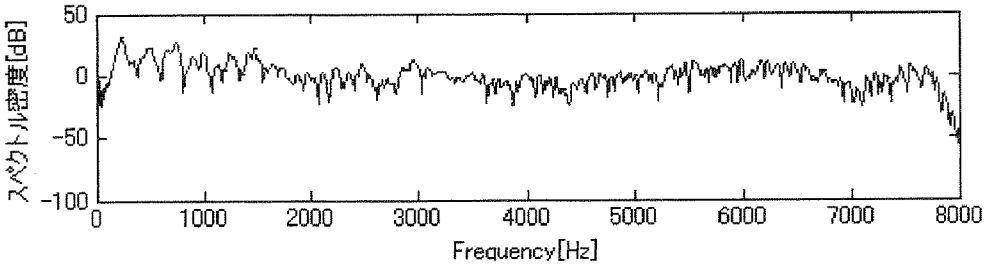


図 2.15: 笑い声の/a/のスペクトル

3. 音素タイミングの比較

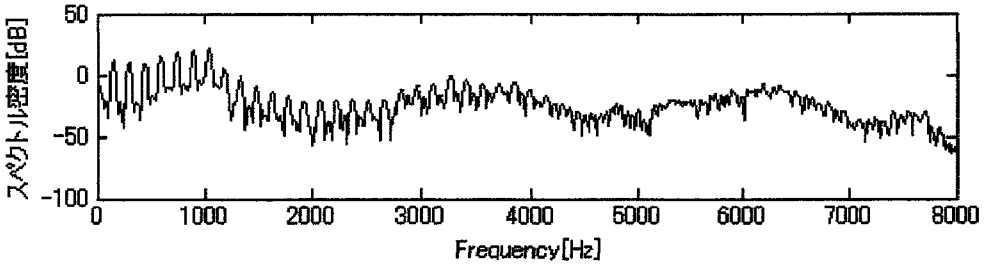


図 2.16: 通常発話の/a/のスペクトル

ラベリング情報を元にして音素タイミングを比較した．収集データから算出したPause長，音素/h/の音素持続時間，音素/a/の音素持続時間の度数分布をそれぞれ図2.17，2.18，2.19に示す．

図2.17より，Pause長は30[msec]〜110[msec]程度であり,「笑い声」でも「通常発話」でもほとんど同じ程度の長さがあるが,「笑い声」の方が「通常発話」よりもPause長が長いといえる．平均値を算出すると,「笑い声」におけるPause長は77.4[msec],「通常発話」におけるPause長は68.4[msec]程度であった．前述のとおり「笑い声」は呼気の流出が「通常発話」より多く，その分だけ呼気時間が長くなるため，9[msec]程度のわずかな差が生じていると考えられる．

図2.18より，音素/h/の音素持続時間に関しては「笑い声」と「通常発話」の間に大きな差はないといえる.「笑い声」においては，過剰な呼気の流出により無声化が強まり，子音である/h/の音素持続時間は長くなると予想されるが，前述のスペクトル包絡に対する分析結果を踏まえれば，無声化の強まりは無声部分の音素持続時間に対してではなく，有声部分の音色に対して影響を与えていると考えられる．

図2.19より，音素/a/の音素持続時間には「笑い声」と「通常発話」の間に大きな差があることがわかる.「笑い声」の方が「通常発話」よりも音素/a/の音素持続時間がかなり短いといえる．平均値を算出すると,「笑い声」における音素/a/の音素持続時間は66.7[msec],「通常発話」における音素/a/の音素持続時間は112.6[msec]程度であった．これは,「通常発話」は一つ一つの音素を明確に発声しているために後続母音の持続時間が長くなるためだと考えられる．また，この差は有声部分のものであるため，聴覚的にも違いが感じられる要因になると考えられる．

2.2.3 笑い声らしさに対する貢献度の評価

概要

一部の音響パラメータを，単に“は”を3回発声した「通常発話」の音声の音響パラメータで差し替えて合成した「笑い声」の音声と，差し替えを行わない「笑い声」の音

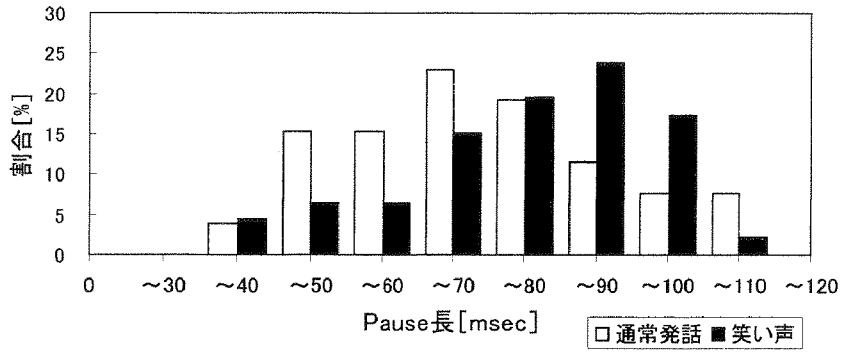


図 2.17: Pause 長の分布

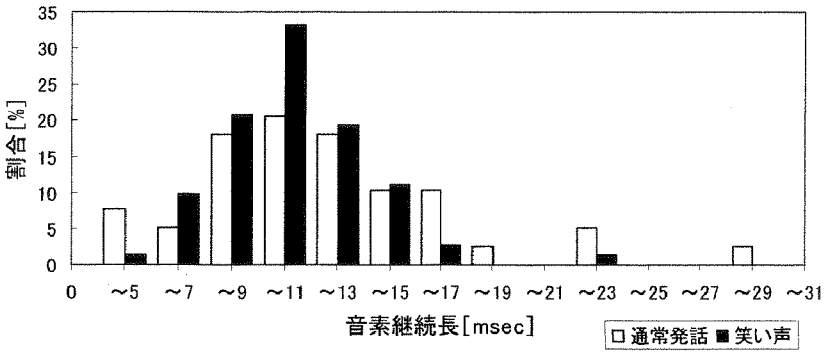


図 2.18: 音素/h/の音素持続時間の分布

声との相違を主観的に比較することで，その音響パラメータの笑い声らしさに対する貢献度を調べた．対象とする音響パラメータには，音源情報として音声パワー，ピッチ周波数，有声・無声判定情報，さらに声道フィルタ (調音) 情報として 16 次の PARCOR 係数によるスペクトル包絡情報を用いた．なお，特徴抽出の際のウィンドウサイズは 30[msec]，シフトサイズは 5[msec] である．

主観評価実験の際に分析対象となるパラメータは，前述の音源情報 3 つと声道フィルタ情報 1 つ，そして Pause の長さや音素/h/,/a/の音素持続時間に関する音素タイミングの情報を加え，以下の 5 つとなる．

- 音声パワー
- ピッチ周波数
- 有声・無声判定情報
- スペクトル包絡情報
- 音素タイミング

これらのパラメータについて，様々な差し替え方をした音声を用意する．具体的には,「笑い声」の各パラメータを「通常発話」のパラメータで順次差し替えて音声を合

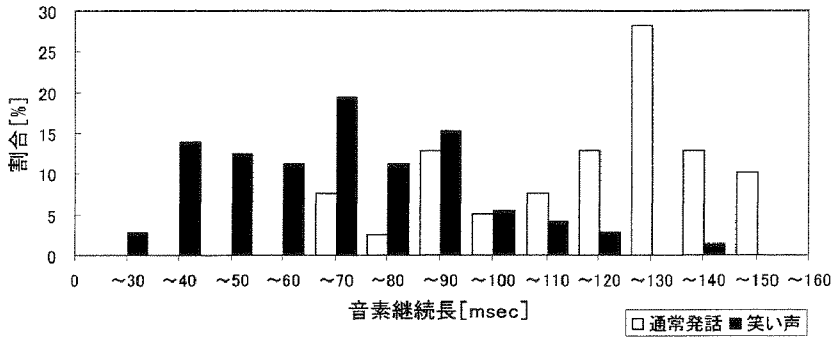


図 2.19: 音素/a/の音素持続時間の分布

成する．全てのパラメータを差し替えた場合,「笑い声」は「通常発話」に変換されることになる．また，差し替えるのではなくデータを加工することによって対象の笑い声のデータに近づけたパターンも用意した．ここでいう加工とは，対象となる笑い声のデータのピッチ周波数，音素持続時間の平均を算出し，それに合うように通常発話のデータのピッチ周波数を一様に上昇させる，または音素タイミングを一様に短くすることをさす．この加工データは，パラメータを一様に変化させたものであるので，ピッチ周波数の時間変化や音素出現の時間変化について笑い声のデータを踏襲していないことが差し替えデータとの相違点となる．また,「笑い声」と「通常発話」のパラメータを差し替える際の時間軸の対応付けは，図 2.20 に示すように，第 2.2.2 項のラベリングに基づいて音素区間を一致させて行い，各音素区間内では時間軸を線形伸縮させた．

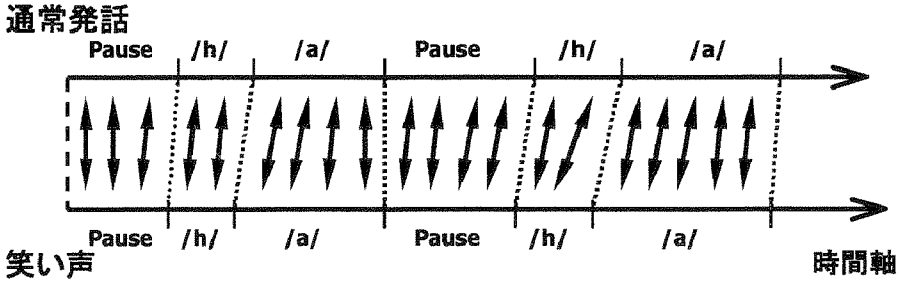


図 2.20: パラメータ差し替えのイメージ

主観評価実験

「笑い声」の笑い声らしさに対する各音響パラメータの貢献度を調べる評価実験を男子大学生 6 人を対象に行った．まず最も「笑い声」に近い音声と「通常発話」に近い音声を提示した後，作成した合成音について,「笑い声」と「通常発話」のどちらに近いかを 1～5 点の 5 段階で評価させる．評点は，最も「笑い声」と判断される場合に

5点，最も「通常発話」と判断される場合に1点とし，笑い声らしい音声ほど高い評点となる(図2.21)．評価対象データ数は，ピッチ周波数と音素タイミングの加工データを含めた5種類のパラメータ全ての組み合わせ計72種類に対して，基準となる笑い声をA,Bの2種類用いたので，合計144である．また，対象となる音声には,「笑い声」のデータ,「通常発話」のデータともに“ははは”という3モーラのデータのみを用いた．

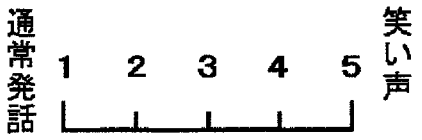


図 2.21: 評価スケール

結果と考察

評価実験の結果に対し，選ばれた音声の評点とパラメータの関係，選ばれた音声の評点とパラメータの組み合わせの関係の2点について以下に示す．

1. 評点平均からみた結果

   評点の平均を音響パラメータ及び使用したデータ毎に表2.35に示す．表2.35より，ピッチ周波数を保存した音声は，それぞれ「笑い声」と「通常発話」に近いと判断されており，両者の違いを表すパラメータといえる．すなわち，ピッチ周波数は「笑い声」の笑い声らしさに対する貢献度が高いことがわかる．しかし，ピッチ周波数を加工したデータは実際の笑い声のデータよりもやや笑い声らしくないと判断されている．つまり，笑い声らしさにはピッチ周波数の時間的変化の影響が大きいと考えられる．

   一方で，音素タイミングについては，笑い声のデータと加工したデータは近い評価値を得ており，時間的な変化の影響はないといえる．つまり，第2.2.2項でみられたPause長のわずかな差異は，聴覚的に大きな影響がでない程度の差異であるとわかる．

   また，表2.35から有声・無声判定情報は笑い声らしさに対する貢献度は非常に低いこともわかる．つまり,「笑い声」は過剰な呼気の流出により無声化が強まるが，そのために生成される無声音を聴取者が感じることは無いといえる．

2. 組み合わせからみた結果

   ここでは，高い評価が得られた音声における具体的なパラメータの組み合わせに着目する．4以上の高い評価を受けた組み合わせを表2.36〜2.39に示す．なお，表2.36〜2.39中の“●”は笑い声のデータ，“−”は通常発話のデータ，“▲”は加

表 2.35: 各パラメータの評価平均点

	使用したデータ		
	通常発話	笑い声	加工
音声パワー	2.94	3.28	—
ピッチ周波数	2.08	3.98	3.26
有声無声判定情報	3.08	3.13	—
スペクトル包絡情報	2.91	3.30	—
音素タイミング	2.38	3.48	3.46

工したデータを用いていることを表す．表中のパラメータ名称は以下の通りである．

power 音声パワー
pitch ピッチ周波数
有声/無声 有声・無声判定情報
spectrum スペクトル包絡情報
timing 音素タイミング

(a) 笑い声らしさに対する基礎的な要素の考察

5段階評価のうち4以上の評価を得られた音声は「笑い声」として認識されたとし，笑い声らしさに対する基礎的な要素を含んでいるとする．そこで，評価者6人全員が4以上の評価をした差し替えパターンに関する表2.36, 2.37について考察する．

表2.36から，ピッチ周波数が低い通常発話のデータは高い評価が得られていないことがわかる．このことから，ピッチ周波数が高いことが,「笑い声」として評価されるための最低限の要因となっていると考えられる．すなわち，ピッチ周波数は「笑い声」の笑い声らしさに対する貢献度が最も高いパラメータであり，笑い声らしさに対する基礎的なパラメータであるといえる．

また，表2.36のみにおいて笑い声A，Bに対する結果を比較すると，単純に音素タイミングが笑い声らしさに対する貢献に不可欠なパラメータであるとはいえないが，表2.37によれば，どのような笑い方であっても音素タイミングが短いほうが「笑い声」として人が認識しやすくなるといえる．つまり，音素タイミングも「笑い声」の笑い声らしさに対する貢献度が高いパラメータであるといえる．ここで，組み合わせに注目すると，表2.37の1行目と2行目，3行目と4行目のそれぞれの組み合わせは，音素タイミングが笑い声のデータであるか加工データであるかの違いだけである．5行目

についても，表2.36の笑い声Bにおいては同様のペアをつくることができ，表2.37における6つの選出パターンのうち5つがこのようなペアをつくれるということになる．このことから，音素タイミングは笑い声に特有な時間変化は無いと考えられ，特別な時間変化を重視せず，単純に短くするだけで笑い声らしく聞こえることがわかる．

次に，音声パワーとスペクトル包絡情報については，表2.36において共に全体の75%の組み合わせが笑い声のデータを使用している組み合わせである．表2.37においては全体に占める割合がさらに増え，83.3%になる．ピッチ周波数や音素タイミングに比べると「笑い声」の笑い声らしさに対する貢献度は低いようであるが，どのような笑い声であってもこのパラメータが笑い声のデータであることが少なからず笑い声らしさに影響を与えると考えられる．なお，使用した「通常発話」の音声は，最後の/ha/のみ音声パワーが減少する音声であった．これに対し，笑い声Aは同様の変化，笑い声Bは音素ごとに音声パワーが一様に減少していく単調減少変化をするものであった．表2.36において，笑い声Bよりも笑い声Aの方が音声パワーを通常発話のデータで差し替えている組み合わせが多く選出されているが，このことと変化パターンが関係していると考えられる．つまり，笑い方，ピッチ周波数の時間変化と音声パワーの時間変化に相関があるということが考えられる．

(b) 笑い声らしさを高める要素の考察

より本物の「笑い声」として人が認識するための要因を調べるため，5段階評価で最高の評価を得た差し替えパターンに関する表2.38，2.39について考察する．

まず，表2.38からもピッチ周波数が高いことと音素タイミングが短いことは笑い声らしさにおいて重要な要素であることがわかる．表2.39によれば，この2つのパラメータは笑い声によらず重要なパラメータであることもわかる．前述のとおり，「笑い声」の笑い声らしさを作り出すための最も基礎的なパラメータがこの2つであると考えられるので，当然の結論といえる．但し，ピッチ周波数に関して，加工したデータは5人以上から最高評価を得られていないことがわかる．つまり，笑い声らしさを強めるためには高い周波数であることが基礎的な要因であるが，単に周波数が高いだけでは十分とはいえない．前項で考察した通り，「笑い声」には特有の時間変化パターンがあり，主観的により本物の「笑い声」として認識するにはこの時間変化パターンを考慮する必要があるといえる．一方で，前項で考察した結果に加え，表2.39からもタイミングの時間変化パターンを考慮する必要はないと考えられる．

次に，前項で，スペクトル包絡情報と音声パワーは笑い声らしさに対する貢献が少なからずあると考えたが，表2.39より，スペクトル包絡情報の方がその貢献度は高いといえる．表2.38によれば，笑い方によらず5人以上か

ら最高評価を得られた音声はすべてスペクトル包絡情報に笑い声のデータを用いている音声であることがわかる．このことからも，より本物の「笑い声」として人が認識する合成音を生成するためにはスペクトル包絡情報を操作し，音色を笑い声特有のものにする事が有効であると考えられる．一方で，音声パワーは多少の貢献度はあるものの，スペクトル包絡情報ほど重要視する必要は無く，第2.2.2項で述べた語調成分の下降にのみ注意をすれば良いと考えられる．

2.2.4 結論

本研究では，任意の合成音を生成可能な規則合成方式による「笑い声」さらには「喋り笑い」の音声合成を目指している．そこで,「笑い声」の音声について，規則合成方式による音声合成に必要な音響特徴を分析した．本研究では,「笑い声」に関して以下のことがわかった．

笑い声一般

はじめに，男子大学生7人分の自然対話の音声を収録し，分析対象データとして笑い声の音声を抽出した．抽出結果から，どの音素を基本とした笑い声を多く発声するかには個人差があることがわかったが，全体として/ha/を基本とする笑い声は出現頻度が高いことがわかった．また，その中で出現頻度の高い笑い声は，“はは” または “ははは” という2〜3モーラの笑い声であった．

音声パワー

笑い声の発声において，音声パワーの時間変化は下り調子の変化をするか，またはほとんど変化をしないかであり，多くは全体的な単調減少の時間変化をすることがわかった．これは,「笑い声」は呼気の流出が「通常発話」より多く，/ha/の発声回数が増えるほど肺の呼気圧が減少するためであり，発声音素数が多い場合はより下り調子の変化になりやすいことが原因と考えられる．

また，主観評価実験の結果によって，音声パワーのもつ「笑い声」の笑い声らしさに対する貢献度はそれほど高くないことがわかった．合成音を生成するときは，音声パワーの時間変化パターンが前述のように全体的な単調減少の変化となるように注意すれば十分であるといえる．但し，ピッチ周波数の変化パターンとの相関にも着目する必要があると考えられる．

表 2.36: 6 人全員が 4 以上の評価をつけた組み合わせ

	パラメータ				
	power	pitch	有声/無声	spectrum	timing
笑い声A	●	●	●	●	●
	●	●	●	●	▲
	●	▲	●	●	▲
	●	▲	●	–	▲
	●	▲	●	–	●
	●	●	–	●	●
	●	●	–	●	▲
	●	▲	–	●	●
	–	●	●	●	▲
	–	●	–	●	●
	–	●	–	●	▲
	●	●	–	–	▲
	–	●	–	–	●
笑い声B	●	●	●	●	●
	●	▲	●	●	●
	●	●	●	●	▲
	●	●	●	●	–
	●	●	●	–	●
	●	●	–	●	●
	●	●	–	●	▲
	–	●	●	●	●
	–	●	●	●	▲
	●	●	–	●	–
	●	●	–	–	▲

表 2.37: 6 人全員が笑い声A，B共通して 4 以上の評価をつけた組み合わせ

パラメータ				
power	pitch	有声/無声	spectrum	timing
●	●	●	●	●
●	●	●	●	▲
●	●	–	●	●
●	●	–	●	▲
–	●	●	●	▲
●	●	–	–	▲

表 2.38: 5 人以上が最高評価をつけた組み合わせ

	パラメータ				
	power	pitch	有声/無声	spectrum	timing
笑い声 A	●	●	●	●	▲
	●	●	–	●	●
	–	●	●	●	●
笑い声 B	●	●	●	●	▲
	●	●	–	●	●
	–	●	●	●	▲
	●	●	●	●	–
	●	●	–	●	–

表 2.39: 4 人以上が笑い声A，B共通して最高評価をつけた組み合わせ

パラメータ				
power	pitch	有声/無声	spectrum	timing
●	●	●	●	▲
●	●	–	●	●
–	●	●	●	●
–	●	●	●	▲

ピッチ周波数

「笑い声」と「通常発話」の/ha/の音についてピッチ周波数を比較したところ,「笑い声」は「通常発話」よりも全体的に50〜100Hzほどピッチ周波数が高く，話者によっては裏声がよく現れ，400Hz付近の音声になることがあるとわかった．話者による裏声の出現頻度の違いは，笑い方に個人差があるためと考えられる．

また，音声パワーの減少に伴い，発声音素数が多くなるとピッチ周波数も減少していくが，ピッチ周波数の時間変化には全体として様々な変化パターンが存在し，それらのパターンの出現頻度に大きな偏りは無かった．そこで，ユーザが合成音を作成する際，聴覚的にわかりやすいという点からもピッチ周波数の変化パターンを操作項目にして笑い声を作り分ける事が有効であると考えられる．しかし一方で，主観評価実験の結果から，変化パターンによって笑い声らしさに差が出るということがわかった．合成音を生成する際，より本物の「笑い声」として人に認識させるためには，笑い声特有の時間変化パターンを考慮する必要があると考えられる．

さらに，この主観評価実験の結果から，ピッチ周波数の低いものは「笑い声」として認識され難く，ピッチ周波数が高いことが笑い声らしさのための重大な要因であることがわかった．このことから，ピッチ周波数は「笑い声」の笑い声らしさに対する最も基礎的なパラメータであるといえる．

有声部・無声部の差異

「笑い声」は声道からの過剰な呼気の流出により無声化が強まるが，主観評価実験の結果によれば，そのために生じる無声音を聴取者が感じることは無いといえる．無声化の影響は，主に音色に影響していると考えられる．

スペクトル包絡・音色

「笑い声」と「通常発話」の/ha/の音について，有声部である後続母音の音素/a/のスペクトル包絡の概形を比較したところ,「笑い声」の場合は無声音や雑音のように高調波構造が曖昧になっていることがわかった．これは，声道からの過剰な呼気の流出によって無声化が進み，子音の影響が強まることによって「通常発話」の音色とは異なる音色になるためだと考えられる．

主観評価実験の結果から，このスペクトル包絡の構造の違いによる音色の違いが主観的にも感じられるものであるとわかった．合成音を生成する際，より本物の「笑い声」として人に認識させるためには，スペクトル包絡情報を操作することにより音色を「笑い声」特有の音色に近づける必要があると考えられる．

音素タイミング

定量的な分析により,「笑い声」の音声と「通常発話」の音声では，Pauseの長さや音素/h/の音素持続時間に差はほとんどなく，音素/a/の音素持続時間に大きな差異があることがわかった．音素タイミングに関して，音素/a/の音素持続時間の差が「通常発話」との違いであるといえる．主観評価実験の結果により，この差が「笑い声」と「通常発話」の聞こえ方の違いに対する重大な要因になっていることがわかった．音素タイミングは「笑い声」の笑い声らしさに対する基礎的なパラメータの一つであるといえる．また，合成の際に音素タイミングの時間変化に注目する必要はなく，笑い声らしさを強調するためには単に音素タイミングを一様に短くすれば良いと考えられる．

第3章　文章の感情理解

本章では，感情表現可能な音声合成システムを構築する上で必要となる文章の感情理解について述べる．本研究では，入力された文章から感情を決定する方法として2通りの方法を用いる．1つは，文章を場面ごとに分割し，分割された場面単位で感情を決定する大局的感情理解であり，もう1つは，文単位で感情を決定する局所的感情理解である．

3.1　目的

音声合成技術は，機械から人へと情報を伝えるインターフェースとして，様々な場面で利用されてきている．とりわけ，音声合成技術の中でもテキスト音声合成は，任意の文章を正確に音声合成するという汎用性から，電子メールの文章読み上げシステムやコンピュータゲームでのキャラクタの音声などに利用されてきた．そして，音声合成技術は，計算機の進歩によりその正確性をますます上げている．そうした背景の下，より高品質な音声合成を目指すためには，従来のようなニュース調の読み上げ式の音声合成ではなく，感情表現豊かな音声合成を実現することが必要となる．

感情表現豊かな音声合成を実現する試みは過去にいくつか行われている．それらの手法の多くは，飯田ら [11] の研究に代表されるように感情ごとの音声のコーパスを用いることによって，感情豊かな音声合成を実現するというものである．しかし，これらの研究では，あらかじめ文章の感情がわかっていなければ音声合成が行えない．つまり，任意の文章を与えたときに，全自動で感情表現豊かな朗読システムを実現するには問題がある．そこで本章では，入力された文章から感情を理解することを目的とする．

本章で提案する感情理解方法は次のようなものである．まず，入力された文章を感情分類する．入力された文章の感情分類を行う方法として，2通りの方法を用いる．1つは場面ごとに緩やかな感情表現を行うための大局的な感情分類であり，もう1つは一文一文を強い感情で表現させるための局所的な感情分類である．大局的な感情分類では，文章を場面ごとに分割し，分割された場面ごとに感情分類を行う．場面分割の方法には，自然言語の分野で研究されている Marti A.Hearst により提案された Text tiling algorithm[13] を修正したものを用いる．場面分割を行った後に場面内の単語の感情に関わる属性により大局的な感情分類を行う．局所的な感情分類では，文法ごとに，感情を決定するための辞書について作成する．

3.2 大局的感情分類

本節では，大局的な感情分類について説明する．まず，大局的な感情分類をするための方法の第一段階である場面分割の際に用いられる Text-tiling algorithm について説明し，次に，分割された各場面の感情を決定する際に用いられる感情表現語辞書の作成方法とそれを用いた場面の感情決定について説明する．

3.2.1 場面分割

望月 [14] によれば，文書は単なる文の集まりではなく，文書中の各文は互いに何らかの意味的関係を持っている．特に意味的関係の強い文が集まって談話セグメントと呼ばれる単位を形成する．ここでいう談話とは，物語，講演，説教など様々なものが含まれる．よって，文が互いに意味的関係を持つように，これらの談話セグメント間にも意味的な関係が存在する．文書の全体的な談話構造はこの談話セグメント間の関係によって形成される．ここではこの談話構造を場面として捉える．

こうした文間やセグメント間の結びつきを扱う重要な言語的性質として，一貫性と結束性がある．一貫性とは，文章中の全体的な談話構造を表す性質である．Hobbs は，一貫性を原因，評価，背景-描写および説明，拡張:並行，詳細化，例示，対照，反対の予想という意味的な関係に分類している．文書中の各文がこのような意味的な関係によって結びつき一貫した談話セグメントを形成する．また，各セグメント間の関係にも意味的な関係が存在することによって，文書全体としての談話構造である序論，本論，結論や起承転結というような構造，主題-副主題構造を形成する．一貫性は談話構造の表現力が高いが，計算機で扱うことは困難である．一貫性を認識するためには，文書内容の完全な理解とセグメント間の関係についての複雑な推論が必要となるためである．また，意味的関係の分類基準が曖昧な点も問題になる．分類基準は，文脈，知識，および分類者の信念に基づく場合が多くシステマティックに関係を決定することができない．

一方，結束性は，文と文の表層的な関係による結びつきのことを言う．Haliday と Hassan は，結束性を理論的な面から，代名詞，省略，接続，代入，語彙的結束性の 5 つに分類している．結束性を自動的に計算するという面からは，代名詞と省略は指示関係にまとめることができる．接続は，接続詞が出現する前後の文がセグメント間の関係を結び付ける関係である．また，代入は，既出の語を他の語に言い換える照応のことをいうため，指示関係に含めて考えることができる．したがって，結束性は，例示，接続詞，語彙的結束性の 3 つに大きく分けられる．この 3 つの分類の内，接続詞以外の 2 つである指示関係と語彙的結束性は，意味的に同じあるいは関連する 2 つの語によって示される文と文の関係であるという点で類似していると考えられる．しかし，語彙的結束性の場合は，関連する 2 つの語が両方とも文書内に出現するため，自動的な計算が比較的容易であるのに対して，指示関係では，関連する 2 つの語の片方は代名詞であったり省略されていたりするため，自動的に計算するためには照応解析

などによって省略された語や指示対照を特定する必要がある．

ここでは，結束性の中で計算機によって比較的容易で自動的に認識できる語彙的結束性について扱う．

Morris と Hirst は，語彙的結束性には，次の五つに分類している．

1. 同一の参照の繰り返し

   例 1. Mary bit into a peach.Unfortunately the peach wasn't ripe.

2. 異なる参照の繰り返し

   例 2. Mary ate some peaches.She likes peaches very much.

3. 上位の概念

   例 3. Mary ate a peach.She likes fruit.

4. 規則性のある意味的な関係

   例 4. Mary likes green apples.She does not like red ones.

5. 規則性のない意味的な関係

   例 5. Mary spent three hours in the garden yesterday.She was digging potatoes.

例 1 から 3 は Haliday と Hassan の分類では，reiteration と呼ばれる．例 1 では，2 つの peach がどちらも同じ peach をさすが，例 2 では Mary が食べた peaches と後の peaches は同じものをさしてはいない．また，例 3 では fruit と peach は概念的に上位下位の関係にある．このように，reiteration には，同じ語の繰り返しだけでなく上位関係，下位関係および類義語が含まれる．

例 4 と 5 は Haliday と Hassan の分類では，collocation と呼ばれる．例 4 は，規則性のある意味的な関係による collocation である．このタイプの関係は反対語，順位性のある語，順位性のない語，全体部分関係，に分類できる．例 5 は，規則性のない意味的関係による collocation である．このタイプの関係は計算が難しい．例えば例 5 の，garden と digging には意味的な関係があるが規則性はない．このタイプの結束性は，類似した状況や内容を記述する場合に共起する傾向の高い語と語の間に存在すると考えられる．例えば，teacher, school, examination は学校についての内容を持つ文書中では，結束性を持つ関係であると考えられる．

しかし実際に語彙的結束性を計算する上では，1 と 2 を厳密に区別することは難しい．どちらも表層上は同一の語で構成されているためである．同一の語の繰り返しについて，それが 1 の結束性であるか 2 の結束性であるかを決定するためには，他の語による情報を取り入れる必要がある．しかし，同一の文書中に出現した同じ語は参照が同一か異なるかを別にすれば同じ意味で使用されていると考えることができる．

上記の事柄の具体的な実現案として，Marti.A.Hearst が提案した Text tiling algorithm がある．algorithm は次の 3 つの部分から成り立っている．

1. 印付け

2. 類似度の探索

3. 場面境界の決定

印付けとは，入力された文章を分割するのに使用する一個一個の語彙的な単位を参照するために行う．語彙的な単位とは，語彙的結束性という観点から意味のあるものである．このアルゴリズムを用いた多くの研究では，内容語（名詞，動詞，形容詞）を用いている．この印付けのアルゴリズムは，次のようにして行う．まず，単語を形態素ごとに分割する．形態素とは，自然言語処理の分野で考え出された概念で，自然言語学者である長尾[15]は,「意味をもつ最小の言語単位，一つ以上の音素からなる」と定義している．形態素解析を使って形態素に分けた結果を用いて，それぞれの形態素の品詞を決定して内容語を抽出することを行う．これらの処理は，奈良先端大学院大学で開発されたフリーソフトウェア "茶筅"[16] を用いて行う．本で "茶筅" を用いた理由としては，次の二つが挙げられる．

1. 自然言語処理の分野で形態素解析を行う際に広く用いられ高い評価を受けている．

2. 形態素を分解できるばかりでなく，品詞も同定してくれる．

印付けの後に，印付けの列の塊ごとの語彙的な類似度を比較する類似度の探索を行う．これは，次のようにして行う．概念図を図 3.1 に示す．まず，テキストの境界の類似度を測定するためには，その境界の単語の幅を持つ二つの窓を入力されたテキストの上にのせる．便宜上，テキストの先頭に近い側の窓を左窓，テキストの終端に近い側を右窓と呼ぶことにする．次に，二つの窓の類似度を計算する．類似度には，式 (3.1) に示すコサインメジャーを使う．

$$sim(b_l, b_r) = \frac{\sum_t w_{t,b_l} w_{t,b_r}}{\sqrt{\sum_t w_{t,b_l}^2 \sum_t w_{t,b_r}^2}} \tag{3.1}$$

ここで，t は，ブロック内の印付けられた単語を表している．b_l と b_r は，それぞれ左窓と右窓を表している．w_{t,b_l} は，左窓の印付けられた単語 t の数を表し，w_{t,b_r} は，右窓の記し付けられた単語 t の数を表している．式 (3.1) は，値が高いほど，境界において，2 つの窓の類似度が高い，つまり，2 つの窓が似ているということであるので 2 つの窓の境界は，場面の境界ではなくなる．逆に，この値が低いほど，場面の境界となる可能性が高いことを示している．類似度を計算した後，刻み幅と呼ばれるもので窓をずらし，その後二つの窓の境界点で類似度を計算する．以上のことをテキストの最初から最後まで行う．

最後に，類似度の探索の計算結果に基づいて，場面境界の決定を行う．基本的に場面境界の決定には，類似度の極小点，つまり，類似度が落ち込むところを利用することにより行う．しかし，類似度の系列には，微弱なゆれがあり，誤った場所を場面境界と認定してしまうことがある．そこで，誤った場所を場面境界として認定しないようにするために，*depth_score* という概念を導入する．*depth_score* は，極小点におい

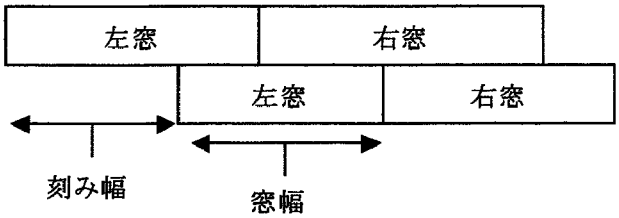


図 3.1: Text tiling algorithm の概念図

てどの程度類似度の変化が大きいのかを調べるための量である．これは式 (3.2) に基づいている．

$$depth_score = (C_{sp} - C_{mp}) + (C_{fp} - C_{mp}) \tag{3.2}$$

ここで，C_{mp} は極小点での類似度，C_{sp}, C_{fp} はそれぞれ C_{mp} での開始点と終了点である．

この $depth_score$ が，ある評価基準 h より大きいとき，その極小点が場面の境界であると判定する．この評価の基準 h には式 (3.3) を使用する．

$$h = \bar{C} - \frac{\sigma}{2} \tag{3.3}$$

ここで，$\bar{C}$ は類似度の平均を表し，σ は類似度の標準偏差を表している．

Hearst の研究の目的は，説明文の談話構造を捉えるということであった．しかし，本研究では，これを物語文に対して適用する．また，このアルゴリズムを利用するために，窓幅の刻み幅のパラメータの決定について実験を行った．実験の方法と結果については，3.4 節で説明する．

3.2.2 感情表現語辞書と文章の大局的感情決定

本項では，場面分割した後に場面ごとに快・不快に決定していく方法を説明する．場面ごとの感情決定には，単語の感情が記述されている感情表現語辞書に基づいて行う．

単語の感情表現語辞書には，大日本図書から出版されている「分類語彙表」[17] を使用した．ここで，本研究で用いる「分類語彙表」について説明する．「分類語彙表」は，約 9 万 6 千語の単語を「単語が表し得る意味の世界を分類して，その分類の各項にそれぞれの単語を配当したもの」である．この単語の量は，一般の小型国語辞典に匹敵する量である．図 3.2 に分類語彙表の概略図 (a) と分類語彙表の中に書かれている単語ごとの番号の例 (b) を示す．以下でそれぞれの分類について説明する．

- 類
  この分類は，一番大きな分類で 4 つの類からなる．すなわち，1. 体の類，2. 用の類，3. 相の類，4. その他の類である．体の類は名称を表す語で名詞の類，用の類は存在・活動を叙述する語で動詞の類，相の類は状態を叙述する語で形容詞・

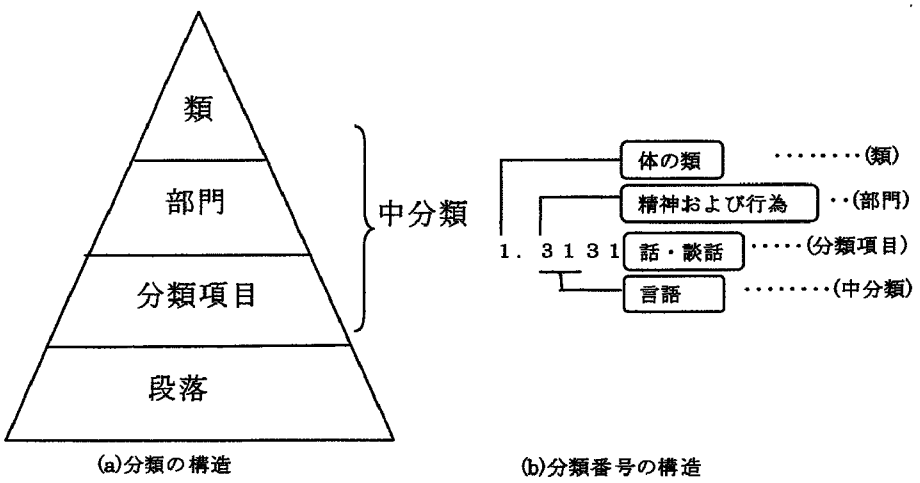


図 3.2: 分類語彙表の構造

形容動詞・副詞・連体詞の類，その他の類は一部の副詞・接続詞・感動詞の類である．図 3.2(b) では，整数部分を表している．

- 部門
  体・用・相の各類の中に，大きな意味的まとまりとして,「抽象的関係」,「人間活動の主体」,「人間活動—精神および行為」,「生産物および用具」,「自然物および自然現象」という五つの部門がある．ただし，用・相の類は,「人間活動の主体」,「生産物および用具」の部門を欠く．図 3.2(b) では，小数点以下一桁目の数字をあらわしている．

- 分類項目
  体・用・相それぞれの類の中に，ほぼ同様の分類項目を設け，同様の配列をとっている．三類相互の参照の必要から，原則として三類共通の項目名と項目番号を付いているため，類によっては，図 3.2(b) に示す番号が飛んでいる場合がある．

- 中項目
  中項目は部門をより具体的に細分したもので，その下位に位置する分類項目をまとめたものである．図 3.2(b) においては，小数点以下一桁目と二桁目を合わせた部分に相当する．

- 段落
  分類項目の中に意味上の語集団を設けたもの．これは，図 3.2(b) とは別に番号が振られているものである．

次に，単語の感情分類について説明する．単語の感情分類には，上で説明した辞書の分類項目を使い，体，用，相の類の単語を快と不快に分類する．快と不快に分類する際の基準は，心理学で提唱されている schulsberg の円感情 [18] を使用した．図 3.3 に

schulsbergの円感情モデルを示す．この図は，縦軸に快・不快をとり，横軸に拒否・注目をおいている．しかし，この図において，快に感情分類されるものは,「愛・喜び・幸福」のみである．

他の感情分類は,「驚き」を除いてすべてが不快に分類されている．例外的に「驚き」については，快・不快どちらにも属すると考えられている．この分類基準に基づいて，分類語彙表の単語を快・不快，無関係に単語の分類を行うことで辞書の作成を行った．

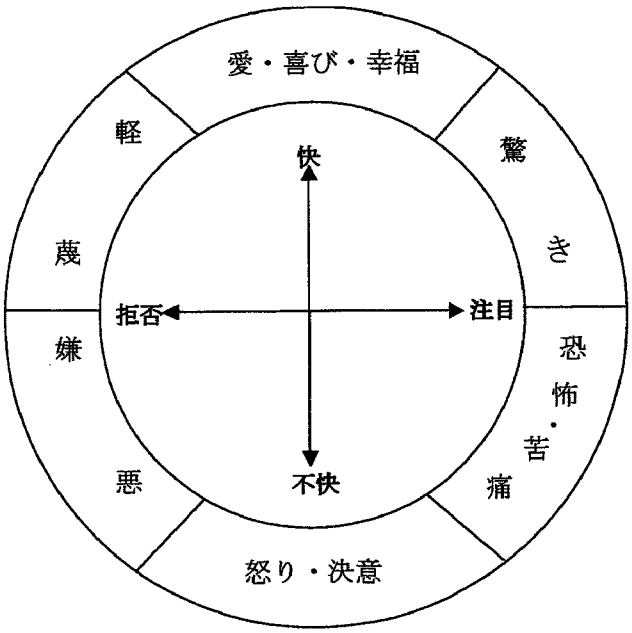


図 3.3: schulsbergの円感情

本研究では，この快不快に分類された場面についての検証実験を行った．結果については，3.4節で説明する．

3.3 局所的感情分類

3.3.1 構文解析

局所的感情分類を行う際には，文章の構文解析と形態素解析を行う．構文解析とは，文法的規則および種々の優先規則によって入力文を解析し，その構造を明らかにすることである．英語の構文解析によく用いられる手法としては，入力された文に単純に文法規則を数あたり的に割り当てていって，当てはまる文法を探していくトップダウン方式の手法と単語の品詞から文法の規則を探っていくボトムアップ方式の手法，または，トップダウン方式とボトムアップ方式をあわせた手法などがある．しかし，日本語の構文解析においては，以下のような理由のために英語の構文解析をそのまま適用することができない．

- 語順が比較的自由である．
  例：
  語順が正しい形：私は，机の上にペンを置いた．
  語順が不完全な形：私は，ペンを机の上に置いた．

- 各要素の省略が可能である．
  例：
  省略されていない形：私は，机の上にペンを置いた．
  省略された形：私は，ペンを置いた．

そこで，日本語の場合には，構文解析の中でも依存関係を考慮した係り受け解析と呼ばれる手法を用いる．係り受け解析の解析結果を図 3.4 に示す．

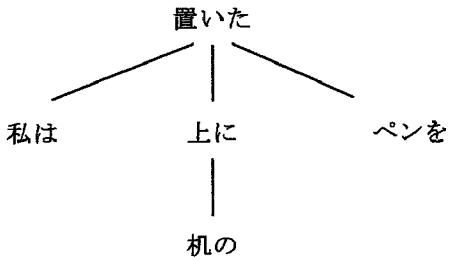


図 3.4: 係り受け解析の解析例

係り受け解析では，文の構造を，係り受け構造とよぶ二文節間の係り受け関係の集合として捉える．文中の二つの文節が係り受けとよぶ修飾・非修飾の関係にあるとき，文頭側の文節が文末側の文節に係るといい，文末側の文節は文頭側の文節を受けるという．係り受け解析は，次のような係り受け構造の制約を設ける．

- 係りの唯一性
  文節は，文末側にある唯一つの文節に係る．

- 係り受けの非交差性
  係り受けは基本的に交差しない．

- 係り受け関係の非重複性
  文節は同じ係り受け関係で二つ以上の文節を受けない．

この係り受け解析を行う部分は，奈良先端大学院大学で開発された Cabocha[19] を使用する．以下では，Cabocha で行っている係り受け解析の仕方について説明する．

Cabocha において係り受け解析には，Abney が英語の構文解析の際に適用した Cascaded chunking parsing を使っている．以下で，そのアルゴリズムについて説明する．

1. すべての文節に O タグを置く．O タグとは，現在の文節の依存関係が決定されていないことを示している．

2. Oタグを持つそれぞれの文節に対して，その文節が隣接する右側にある文節を修飾するかどうかを決定する．もし，修飾するようであれば，Oタグに，Dタグを代入する．

3. Oタグの右隣にDタグがあるならDタグを削除する．

4. もし，一つの文節のみが残っているのならこのアルゴリズムは終了する．そうでない場合には，ステップ2に戻って繰り返し同じことを行う．

図3.5にこのアルゴリズムによって解析した例を示す．

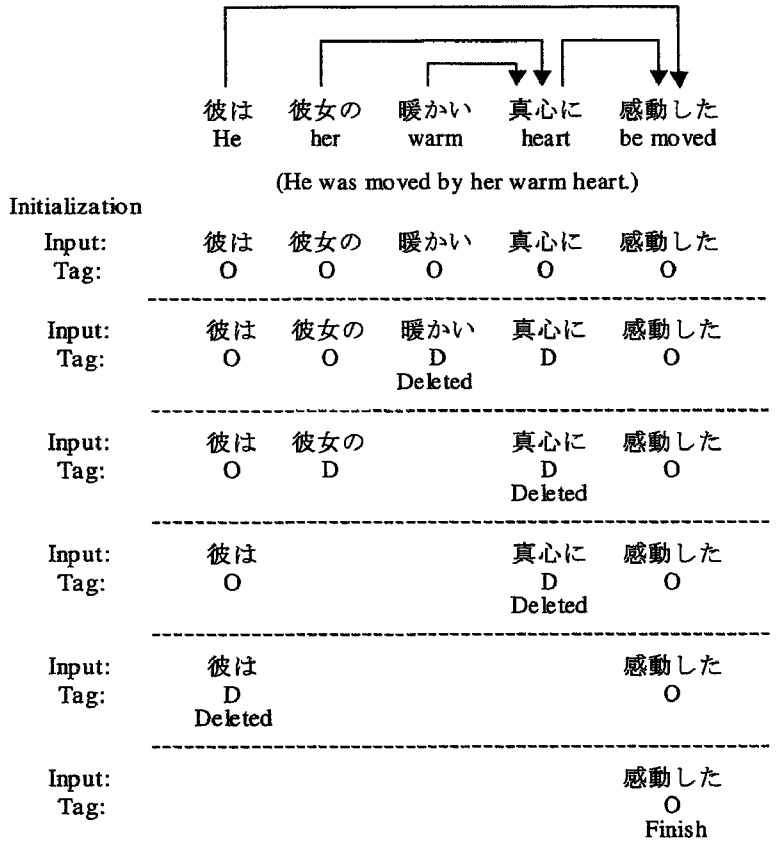


図 3.5: Cabochaにおける係り受け解析の例

3.3.2 感情表現要素辞書と局所的感情分類の決定

本項では，3.3.1項で行った結果に対して，局所的な感情分類を行うために使用する感情表現要素辞書について説明する．本研究では，形容詞で終わる形容詞文，形容詞的な名詞で終わる名詞文，さらに動詞文について辞書を作成する．ここで，形容詞的な名詞というのは,「明らか」や「かわいそう」といった名詞に係ってもおかしくないような名詞のことをいう．

感情表現語辞書を作成する際に，情報処理振興事業協会(IPA)が出版しているIPAL[20]を用いた．IPALとは，語彙体系上ならびに使用頻度上重要であると考えられる基本的な単語である，1081語の名詞，861語の動詞，136語の形容詞を意味および統語的な特徴に基づいて区分し，それを1つの単位として，形態，意味，統語，慣用表現などに関する情報が記載されている辞書である．特にこの辞書の中では，名詞には「意味素」という項目がついている．意味素のいくつかの例を表3.1に示す．意味素とは，分解意味論で用いられる意味の最小単位のことである．形容詞，動詞については，単語の意味ごとにどの意味素がつくのかで分類されている．本研究では，意味ごとに辞書を作成する際にこの意味素を使用した．

感情を分類する際の基準としては，3.2.2項で説明したschulsbergの円感情モデルから,「苦痛」を「悲しみ」と変えて,「喜び，怒り，悲しみ，驚き，恐怖，嫌悪」の6感情と「無感情」を使用した．

表 3.1: 意味素性の例

意味素	例
人間	医者, 患者
測量	体重
組織	政党
場所	ホテル

先行して行った少数の研究では，形容詞文は，文全体が形容詞に依存することを示唆するデータがある．また，動詞文については，名詞と動詞の組み合わせが文の感情を決定することを示唆するデータがある．

そこで，形容詞文や形容詞的な名詞文については，形容詞を意味ごとに感情分類することを行った．一方，動詞文については，前にかかる名詞と動詞の組み合わせが，文全体を支配するとした．また，すべての組み合わせについて感情を被験者実験によって調べるのは困難であり，瀬田らの主張[21]によれば,「対人関係」が感情に与える重要な成分となっている．したがって,「対人関係」が動詞文の感情に大きな影響を与えると仮定して，前にかかる「名詞」は,「人」についてのみ着目した．

3.4 実験

本節では，3.2節及び3.3節で説明した理論に基づき行った，実験方法と結果について述べる．3.4.1項では，大局的な感情分類に対して行った実験について述べる．3.4.2項では，局所的な感情分類に対して行った実験について述べる．

3.4.1 大局的な感情分類に対する実験

場面分割に関する実験

1. 実験1 パラメータ推定

   (a) 実験方法
   本実験は，Text tiling algorithm で使われる窓幅と刻み幅のパラメータを推定するための実験である．パラメータ推定は，以下のような方法で行った．

   i. 被験者に文章を読んでもらい，場面分割を行ってもらう．

   ii. i で取得したデータにおいて，適切な人数が共通して分割した場面境界を教師データとして，Text tiling algoritmm のパラメータを試行錯誤的に探索することで，評価基準の高いパラメータを選択する．

   実験で用いた文章は，web 上にある「NTT ドコモ電子図書館」[22] から昔話を1セット選択した．昔話を選択した理由は，以下のことである．

   - 語彙数が少ない．
   - 文の感情表現がはっきりしている．

   被験者の 15 人の学生であり，被験者には「次に文章の場面の区切れだと思うところで，区切ってください」と指示した．ここでいう「場面」を厳密に以下のように定義する．

   i. “同じ話題である” とか “同じ事柄や状況を説明している” というように他の部分の文章と明確に区別できる文章の単位．

   ii. 場面は，“」” か “。” の部分で区切る．

   評価基準は，情報検索などの分野で用いられている一致度である．一致度とは，再現率と精度によって表す．ここでいう再現率とは，Text tiling algorithm により，被験者の場面境界がどれだけ一致するのかを表す指標であり，式 (3.4) のように定義される．

$$\text{再現率}\ (precision) = \frac{\text{被験者と}\ algorithm\ \text{が一致した場面境界数}}{\text{被験者の場面境界数}} \tag{3.4}$$

   一方，精度は，Text tiling algorithm によって検索された場面境界の中に場面境界がどれだけ含まれるかを表す指標であり，式 (3.5) のように定義される．

$$\text{精度}\ (recall) = \frac{\text{被験者と}\ algorithm\ \text{が一致した場面境界数}}{algorithm\ \text{の場面境界数}} \tag{3.5}$$

   ただし，ここでいう一致とは，ある程度の誤差を許した中での一致である．また，基準となる Hearst の実験結果では，再現率と精度がそれぞれ，約 70%，約 50%である．

(b) 結果

まず，被験者に選択してもらった場面分割の結果は，図3.6のようになった．ここで，横軸の「文の番号」とは，文章中の文の番号を表しており，縦軸の「被験者が分けた頻度」とは，その文の終わりが場面の境界であると判断した被験者の人数を表している．

本実験では，被験者4人以上が場面の境界であると判断した部分訓練データとして使用し，Text tiling algorithm のパラメータを決定する．パラメータの誤差は，1文とした．つまり，前後1文の範囲で場面が一致したと認定する．実験の結果，窓幅を102単語，刻み幅を16単語のときに，再現率が89%，精度が57%と最もよかった．

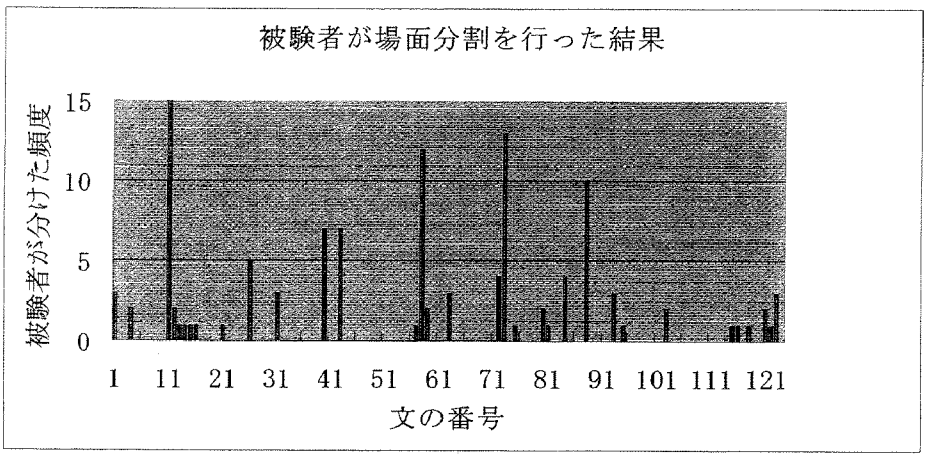


図 3.6: 被験者が場面分割を行った結果

2. 実験2 検証実験

(a) 実験方法

検証実験は，次のように行った．

i. 実験1と同様に被験者に評価用の文章を読んでもらう．

ii. 実験1で用いたパラメータを用いて，評価用の文章に対して場面分割を行う．

iii. iより被験者が4人以上選んだ場面と Text tiling algorithm が，選んだ場面に対して一致度の評価を行う．一致度を評価する際の誤差は，3文とした．

実験には，実験1同様，昔話を用いた．用いた文章の数は2セット，被験者の数は，実験1同様15名である．

(b) 結果

文章1を用いた実験では，再現率が78%，精度が78%であり，文章2を用

いた実験では，再現率が44%，精度が70%であった．文章1は，望ましい結果となった．文章2の結果がよくなかった原因は，文章2の場面の展開が早く，Haerst法の特徴量である語彙的結束性があまり生じなかったためであると考えられる．

大局的な感情決定に関する実験

次に「場面ごとの快・不快の決定は，単語の快・不快の頻度が大きい方の場面である」というルールが正当なものであるかを評価するための実験を行った．

1. 実験方法
   まずデータセットを以下のように用意した．

   (a) 文章をText tiling algorithmにより場面に分割する．

   (b) 感情表現語辞書に基づいて，文章中の単語に快・不快ラベルをつける．

   (c) 文章の各場面において,「場面内の単語が快のみのもの」,「場面内の単語が不快のみのもの」,「場面内に快の単語が多く不快の単語が少ないもの」,「場面内に不快の単語が多く快の単語が少ないもの」をそれぞれ5セットずつ抽出する．

   場面分割のために用いた文章は，web上にある児童文学，昔話，童話など[22]-[31]の比較的語彙数が少ないと考えられる文章を選択した．

   アンケートは，被験者17名の学生に対してweb上で行った．アンケートを取る際には，“快(喜び)”,“不快(怒り，悲しみ，嫌悪，恐怖)”，“快, 不快には関係ない”のどれに属するかを選択してもらった．

2. 結果
   快のみの文章の結果を表3.2と図3.7に示す．結果は，本研究の予想を裏付けるデータを示唆している．ただ，文章2において無関係が，ほぼ同じ量を示している．これは，文章2の快の同じ単語が繰り返し用いられており，反復される間隔が短いためだと思われる．具体的には「あッ，桃売りが来た！面白い，面白い，面白い，桃売りが来た．うれしい，うれしい，うれしい，桃売りが来た！」(うれしいが感情表現語辞書内での快)のような文である．

   不快のみの文章の結果を表3.3と図3.8に示す．結果は，本研究における予想を裏付けるデータとなった．さらに，快のときの文章に比べ，データにはっきりとした偏りが出ている．

   快が多く，不快が少ない文章の結果を表3.4と図3.9に示す．図3.9を見ると，全体的に快と判断される文章が多く，予想通りになった．また，不快あるいは無関係と判断された場面数との差は，図3.7と比較して小さいものが多かった．これは，不快の単語により，快の性質が弱まってしまったためだと考えられる．また，

表 3.2: 大局的感情分類の評価結果 (快のみ)

	快	不快	無関係
No.1	13	1	3
No.2	8	2	7
No.3	12	1	4
No.4	12	2	3
No.5	14	0	3

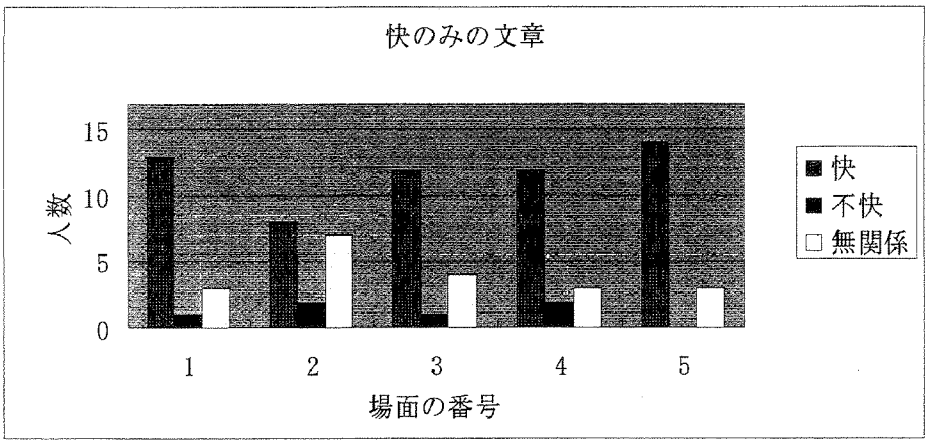


図 3.7: 大局的感情分類の評価結果 (快のみ)

表 3.3: 大局的感情分類の評価結果 (不快のみ)

	快	不快	無関係
No.1	1	15	1
No.2	0	12	5
No.3	1	13	3
No.4	0	12	5
No.5	0	17	0

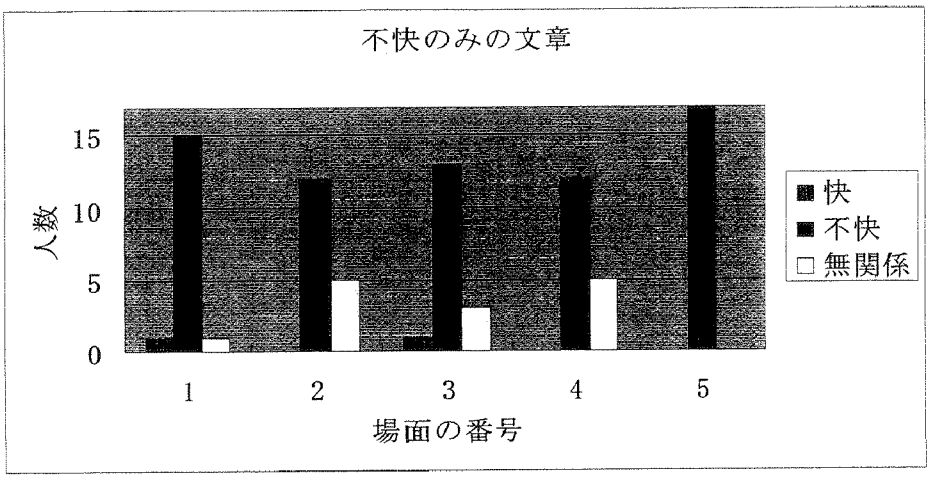


図 3.8: 大局的感情分類の評価結果 (不快のみ)

文章 4 において，快と不快が全く逆の結果が出ている．これは，文章 4 内の単語で最も多く出現した "明るい" という単語が，分類語彙表の項目の性格が明るいという意味ではなく，光が明るいという意味で使われていたためだと思われる．

表 3.4: 大局的感情分類の評価結果 (快が多く不快が少ない)

	快	不快	無関係
No.1	9	1	7
No.2	9	1	7
No.3	11	3	3
No.4	1	9	7
No.5	8	3	6

不快が多く，快が少ない文章の結果を表 3.5 と図 3.10 に示す．ほぼ予想通りの結果が得られた．また，不快と判断された場面数と無感情と判断された場面数との差は，図 3.7 と比較して小さいものが多かった．これは，不快の単語により，快の性質が弱まってしまったためだと考えられる．

3.4.2 局所的な感情決定に関する実験

本研究では，形容詞の感情表現要素辞書，名詞の性質分類辞書，動詞の感情表現辞書の作成に関する 3 つの実験を行った．

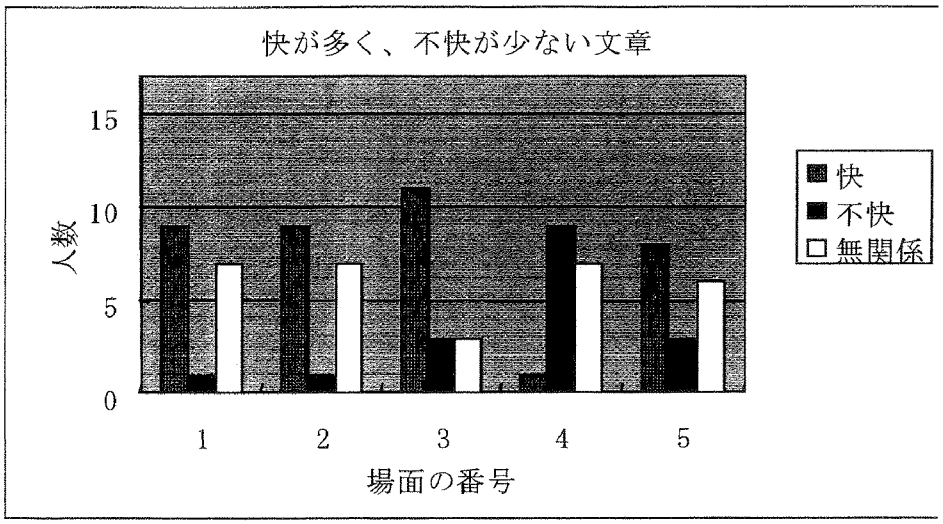


図 3.9: 大局的感情分類の評価結果 (快が多く不快が少ない)

表 3.5: 大局的感情分類の評価結果 (不快が多く快が少ない)

	快	不快	無関係
No.1	4	7	6
No.2	0	13	4
No.3	4	6	7
No.4	1	11	5
No.5	4	3	10

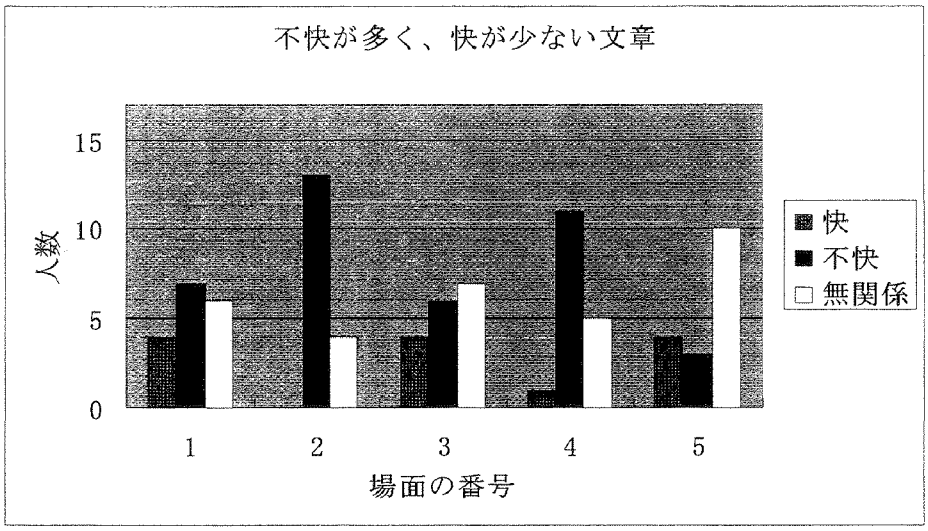


図 3.10: 大局的感情分類の評価結果 (不快が多く快が少ない)

形容詞及び形容詞的な名詞の辞書作成

被験者アンケートにより，形容詞及び形容詞的な名詞の辞書を作成した．アンケートの際には，被験者に「怒り」,「喜び」,「悲しみ」,「驚き」,「恐怖」,「無感情」の7感情から，ひとつを選択してもらった．被験者の数は，データを4分割し，それぞれのデータに対して，20名ずつにアンケートを行った．

表3.6に形容詞の感情分類結果を示す．表中の数値は，(ある感情を選択した人数) ÷ (全体の人数=20人) を表している．

表 3.6: 形容詞の感情分類結果

単語	喜び	悲しみ	怒り	驚き	恐怖	嫌悪	無感情
有難い	1.00						
嬉しい	1.00						
親しい	1.00						
明るい (陽気)	1.00						
面白い	1.00						
めでたい	1.00						
楽しい	1.00						
好ましい	0.95						0.05
輝かしい	0.95			0.05			
美しい	0.95			0.05			
かわいい	0.95						0.05
おいしい	0.95						0.05
懐かしい	0.95			0.05			
温かい (親切)	0.95		0.05				
うまい	0.95						0.05
やさしい (心)	0.95					0.05	
よい	0.95						0.05
頼もしい	0.95			0.05			
広い (心)	0.95						0.05
素晴らしい	0.95						0.05
快い	0.90			0.05			0.05
暖かい	0.85			0.05		0.05	0.05
甘い (味)	0.85						0.15
軽い（楽）	0.75						0.25
やさしい（簡単）	0.75			0.05			0.20
新しい	0.70			0.10			0.20
涼しい	0.70	0.05			0.05		0.20
明るい (視覚)	0.70						0.30

表 3.6: 形容詞の感情分類結果

単語	喜び	悲しみ	怒り	驚き	恐怖	嫌悪	無感情
柔らかい (刺激)	0.65			0.05			0.30
おかしい (滑稽)	0.65		0.05	0.10			0.20
深い (情)	0.65	0.10					0.25
正しい	0.65					0.05	0.30
熱い (心)	0.65	0.05	0.05	0.05		0.05	0.15
ふさわしい	0.60			0.05			0.30
強い (心)	0.60			0.10			0.30
若い	0.60			0.05			0.35
やすい	0.60	0.05		0.05		0.05	0.25
柔らかい (気性)	0.55			0.10			0.35
温かい (温度)	0.50			0.05			0.45
たくましい	0.50			0.10			0.40
欲しい	0.45	0.10				0.05	0.40
勇ましい	0.45			0.25			0.30
えらい	0.40		0.05	0.15		0.05	0.35
深い (関係)	0.40	0.10		0.05			0.45
悲しい		0.90	0.05			0.05	
寂しい		0.85				0.05	0.10
惜しい	0.05	0.80	0.10				0.05
薄い (髪)		0.75		0.05	0.10	0.05	0.05
冷たい (心)		0.70			0.10	0.20	
乏しい		0.70			0.05	0.20	0.05
みすぼらしい		0.60			0.05	0.30	0.05
貧しい		0.60	0.10		0.05	0.20	0.05
少ない		0.55	0.10	0.05		0.10	0.20
悔しい		0.50	0.40				0.10
つらい		0.45			0.10	0.45	
やむおえない		0.45				0.10	0.45
深い (深刻)		0.40			0.35	0.25	
苦しい		0.40	0.05		0.15	0.40	
古い	0.05	0.40		0.10			0.45
狭い (心)		0.40	0.10			0.40	0.10
羨ましい		0.40	0.15	0.05		0.10	0.30
うるさい			0.75			0.25	
騒がしい			0.60	0.05		0.35	
ずるい			0.60			0.35	0.05

表 3.6: 形容詞の感情分類結果

単語	喜び	悲しみ	怒り	驚き	恐怖	嫌悪	無感情
憎い		0.05	0.50	0.05	0.05	0.35	
憎らしい			0.50		0.05	0.45	
やかましい			0.45	0.05		0.40	0.10
すさまじい			0.10	0.80	0.05		0.05
珍しい	0.05			0.80			0.15
とんでもない			0.10	0.75		0.05	0.10
激しい (勢い)	0.05			0.70	0.15	0.05	0.05
著しい	0.05			0.60			0.30
からい		0.05		0.50	0.05	0.10	0.30
大きい	0.15			0.50			0.35
高い (金額)	0.05	0.05	0.05	0.45	0.10	0.05	0.25
鋭い (勘)	0.30			0.45		0.05	0.20
恐ろしい				0.05	0.95		
怖い				0.05	0.85	0.10	
危ない		0.05	0.05	0.25	0.60		0.05
鋭い (角)			0.05	0.15	0.55		0.25
荒い (乱暴)		0.05	0.25		0.55	0.15	
かたい (緊張)				0.20	0.45		0.35
険しい (緊張)		0.05	0.05	0.05	0.40	0.35	0.10
わずらわしい		0.05	0.05			0.90	
汚い		0.10	0.05			0.85	
臭い（におい）		0.05	0.05	0.10	0.05	0.75	
かゆい		0.05	0.10		0.05	0.75	0.05
難しい	0.05		0.10	0.05		0.75	
苦い (不快)		0.10	0.10		0.10	0.70	
みにくい		0.25	0.05			0.70	
悪い		0.10	0.05		0.20	0.65	
たまらない (不快)		0.05	0.20		0.10	0.65	
忙しい	0.10	0.10	0.10			0.65	0.05
みっともない		0.10	0.10	0.10	0.10	0.60	0.10
臭い (疑わしい)		0.05		0.05	0.10	0.60	0.20
しつこい			0.35		0.05	0.60	
苦い (味)		0.05		0.15	0.10	0.55	0.15
渋い (味)		0.05		0.15	0.10	0.55	0.20
まずい (味)		0.15	0.10	0.10		0.55	0.10
黒い (心)		0.10	0.10		0.25	0.50	0.05

表 3.6: 形容詞の感情分類結果

単語	喜び	悲しみ	怒り	驚き	恐怖	嫌悪	無感情
きつい (苦しい)		0.20	0.05		0.15	0.50	0.10
暗い (陰気)		0.15			0.15	0.50	0.20
ひどい		0.30	0.20		0.05	0.45	
暑い			0.25			0.45	0.30
細い (神経)	0.05	0.10			0.10	0.40	0.35
鈍い		0.15	0.05	0.10	0.05	0.40	0.25
つまらない		0.20	0.20			0.40	0.20
かたい (まじめ)				0.10		0.10	0.80
眠い			0.05	0.05		0.15	0.75
ゆるい	0.05	0.05				0.15	0.75
渋い (人柄・顔)	0.05	0.05	0.05	0.05		0.05	0.75
浅い (空間的広がり)	0.10	0.05			0.10	0.05	0.70
赤い	0.05		0.15		0.10		0.70
粗い (触覚)				0.15	0.05	0.30	0.50
甘い (手ぬるい)	0.05		0.20	0.10		0.15	0.50
青い (色)	0.20	0.30					0.50
浅い (不十分)		0.20		0.05		0.30	0.45
高い（プライド）	0.05	0.10	0.10	0.05		0.25	0.45
ない（存在）		0.30			0.15	0.10	0.45
冷たい (温度)		0.05		0.35	0.10	0.10	0.40
はやい (時)	0.10	0.15		0.30	0.05		0.40
まずい (不適当)		0.20	0.05	0.05	0.05	0.30	0.35
険しい (困難)		0.05	0.05	0.20	0.10	0.25	0.35
青い (未熟)	0.10	0.25	0.05		0.05	0.20	0.35
狭い (範囲)	0.10	0.05		0.10	0.25	0.15	0.35
多い	0.30			0.30	0.05		0.35
甚だしい			0.15	0.20		0.35	0.30
遅い		0.15	0.20			0.35	0.30
寒い		0.35	0.05		0.05	0.25	0.30
広い（範囲）	0.15	0.05		0.30	0.10	0.10	0.30
熱い (温度)		0.05		0.30	0.05	0.35	0.25
粗い (雑)		0.10	0.30			0.35	0.25
いけない (ダメ)	0.10	0.10	0.20	0.05	0.05	0.25	0.25
もろい		0.20	0.05	0.20	0.15	0.20	0.20
厳しい	0.05		0.15	0.10	0.35	0.15	0.20
怪しい	0.10			0.10	0.35	0.30	0.15

表 3.6: 形容詞の感情分類結果

単語	喜び	悲しみ	怒り	驚き	恐怖	嫌悪	無感情
恥ずかしい	0.15	0.10	0.05	0.15	0.10	0.30	0.15
おかしい	0.10			0.25	0.20	0.35	0.10
(変・怪しい)							
慌しい		0.05	0.25	0.15	0.20	0.25	0.10
痛い		0.15	0.10	0.05	0.35	0.30	0.05
激しい (気性)	0.05		0.30	0.20	0.30	0.10	0.05

形容詞における辞書中の感情の割合を，表 3.7 に示す．表中の分散とは，一つの感情に決定できずに他の感情にばらついてしまい，感情が特定できなかった単語である．表を見ると，無感情にくらべ，喜びや嫌悪の方が辞書の中の割合が大きい．図 3.3 に示す schulsberg の円感情により，喜びを快，喜び以外の有感情を不快と考えると，不快の有感情において，嫌悪と悲しみが多く，逆に怒り，驚きと恐怖は少ないという結果になった．

表 3.7: 形容詞における辞書の中の感情の割合

感情	割合
喜び	31%
悲しみ	12%
怒り	4%
驚き	6%
恐怖	5%
嫌悪	17%
無感情	10%
分散	14%

表 3.8 に，形容詞的な名詞の感情分類結果を示す．数値は，形容詞の場合と同様の方法で算出した．

表 3.8: 形容詞的な名詞の感情分類結果

単語	喜び	悲しみ	怒り	驚き	恐怖	嫌悪	無感情
幸福	1.00						
快適	1.00						
快活	1.00						
綺麗　奇麗	0.85			0.05			0.10
可能	0.80						0.20

表 3.8: 形容詞的な名詞の感情分類結果

単語	喜び	悲しみ	怒り	驚き	恐怖	嫌悪	無感情
活発	0.80						0.20
円満	0.80			0.05			0.15
健全	0.75						0.25
穏やか	0.75						0.25
甘味	0.65			0.10		0.05	0.20
鮮やか	0.65			0.10			0.25
明らか	0.65						0.35
謙虚	0.55	0.05				0.05	0.35
勤勉	0.55	0.05		0.05		0.05	0.30
気の毒		0.90					0.05
かわいそう		0.90				0.05	0.05
哀れ　憐れ		0.80		0.05	0.15		
孤独		0.70			0.20		0.10
苦労	0.10	0.50	0.05			0.15	0.20
勝手		0.05	0.45	0.10		0.30	0.10
頑固			0.40			0.25	0.35
緊急			0.05	0.65	0.15	0.05	0.10
巧妙	0.10		0.05	0.55		0.10	0.20
残酷		0.25	0.05		0.50	0.20	
嫌		0.05	0.10			0.85	
嫌い		0.15	0.05			0.75	0.05
窮屈			0.15		0.10	0.65	0.10
悪質		0.05	0.25		0.10	0.60	
軽率	0.10	0.10	0.15			0.60	0.05
いんちき		0.10	0.20	0.15		0.55	
広さ	0.40				0.05		0.55
公平	0.40					0.05	0.55
堅実	0.40			0.05			0.55
間接			0.05			0.05	0.90
簡略	0.20					0.05	0.75
いろいろ	0.20		0.05				0.75
厚め	0.10			0.10		0.05	0.75
肝心	0.10		0.10	0.05	0.05		0.70
細かさ	0.05	0.10		0.05		0.10	0.70
簡潔	0.25	0.05		0.05			0.65
軽さ	0.25					0.10	0.65

表 3.8: 形容詞的な名詞の感情分類結果

単語	喜び	悲しみ	怒り	驚き	恐怖	嫌悪	無感情
かすみ		0.05		0.05	0.15	0.10	0.65
広め	0.35			0.05			0.60
啓蒙	0.15			0.25			0.60
かすか	0.10	0.15		0.10	0.05		0.60
粗め　荒め	0.05	0.05	0.10	0.05		0.15	0.60
重き		0.15	0.05		0.05	0.15	0.60
軽め	0.30		0.05	0.05		0.05	0.55
薄め	0.15	0.20	0.05	0.05			0.55
かなり	0.05			0.30	0.05	0.05	0.55
感心	0.50			0.10			0.40
滑稽	0.25			0.25	0.05		0.45
重み	0.10	0.10		0.05	0.15	0.20	0.40
急ぎ	0.05		0.05	0.25		0.25	0.40
強情	0.10		0.25	0.10		0.35	0.20
荒くれ			0.30		0.35	0.30	0.05
臆病		0.25	0.05		0.30	0.30	0.10
安易	0.35	0.10	0.05	0.05		0.25	0.20
狭め		0.10	0.10	0.05	0.20	0.20	0.35
厳重	0.05		0.15	0.15	0.20	0.15	0.30
強力	0.30		0.05	0.30			0.35

形容的な名詞における辞書中の感情の割合を表 3.9 に示す．表 3.7 と比較すると，無感情の比率が高かった．これは，形容詞の方が，名詞に比べ，感情を含む単語の頻度が高いことを示唆するデータである．

表 3.9: 形容詞的な名詞における辞書の中の感情の割合

感情	割合
喜び	23%
悲しみ	8%
怒り	3%
驚き	3%
恐怖	2%
嫌悪	10%
無感情	38%
分散	12%

名詞の性質分類辞書

動詞に強く影響を与えると考えられる「人」の性質を持つ名詞についての辞書を作成した．実験方法は，名詞から IPAL により「人」の意味素性を持つ単語を抜き出す．次に，被験者に各名詞について「快」と「不快」,「友好」と「敵対」,「主」と「従」のそれぞれについて該当するならば複数回答ありで回答してもらい，どれにも該当しないならば「該当なし」を選択してもらうアンケートを行った．被験者の数は，15 名の男女である．表 3.10 に結果を示す．またここでは，6 名以上が回答した場合に該当する性質があったと判断する．

表 3.10: 名詞の性質分類アンケート結果

単語	主	従	友好	敵対	快	不快	該当なし
父	11	0	5	1	1	1	3
飼い主	10	0	4	0	2	0	8
社長	10	0	2	2	1	2	6
議長	9	0	2	0	2	0	7
主治医	9	1	2	2	1	2	6
主人	9	0	4	0	3	0	6
先生	9	2	4	2	2	2	4
親	8	1	6	0	2	0	8
幹部	8	0	3	2	2	3	7
先輩	8	2	5	1	1	2	4
夫	7	0	6	0	1	0	9
大人	7	0	2	1	1	1	8
母	7	0	9	0	4	0	2
兄	6	0	7	0	4	1	12
姉	6	0	7	0	5	0	12
委員	6	2	1	1	2	2	11
医者	6	3	2	1	0	4	11
担当	6	0	1	1	1	1	4
おじいさん	5	0	8	0	2	1	9
自分	5	0	5	1	3	1	5
男	4	0	3	1	1	1	8
おばさん	4	0	5	2	1	2	8
専門家	4	1	2	1	1	1	4
相談役	4	2	9	0	4	0	4
天才	4	0	3	2	4	2	3
僕	4	1	2	0	1	0	2
役員	4	2	1	1	1	2	1
私	4	0	4	1	3	2	0

表 3.10: 名詞の性質分類アンケート結果

単語	主	従	友好	敵対	快	不快	該当なし
客	3	2	4	1	1	1	7
警官	3	0	1	3	1	6	7
警察	3	0	1	4	1	6	7
堅物	2	0	1	4	0	6	7
作者	2	1	3	0	3	0	6
選手	2	3	3	0	2	0	4
人間	2	1	6	1	2	1	2
人	2	0	5	1	1	2	2
赤ちゃん	1	8	14	0	10	1	13
あなた	1	2	4	0	2	0	12
絵描き	1	0	3	0	2	1	9
家族	1	1	11	0	6	0	8
学生	1	3	4	1	1	1	7
兄弟	1	1	8	2	3	1	7
国民	1	3	3	0	2	0	6
職人	1	0	3	1	2	1	5
住民	1	3	5	0	1	0	5
人類	1	0	5	1	1	1	5
妻	1	4	8	0	7	0	3
年寄り	1	1	10	1	3	2	3
友達	1	1	13	1	8	0	3
日本人	1	1	5	1	1	1	2
被害者	1	2	3	1	2	5	2
妹	0	6	9	0	1	0	10
嘘つき	0	0	0	10	0	12	9
生まれ	0	1	2	0	2	0	9
弟	0	7	8	0	2	0	8
女	0	1	6	1	1	1	8
敵，仇	0	0	1	13	1	6	8
子，	0	6	9	1	5	1	6
恋人	0	0	12	0	11	0	6
子供	0	6	9	0	5	0	6
生徒	0	7	4	1	1	1	4
他人	0	1	2	1	1	2	4
大工	0	0	2	1	1	1	4
通	0	0	1	1	0	1	3

表 3.10: 名詞の性質分類アンケート結果

単語	主	従	友好	敵対	快	不快	該当なし
当番	0	4	4	0	3	1	3
二年生	0	1	3	0	1	1	2
病人	0	2	3	0	1	5	2
夫婦	0	0	12	0	4	0	2
保証人	0	0	12	0	4	0	2
孫	0	4	10	0	7	0	2
迎え	0	2	3	0	3	1	1
娘	0	5	8	0	7	0	1
者	0	0	2	1	1	1	1

友好関係において，性質ありと判定された単語を表3.11に示す．表3.11を見ると，友好に選ばれた単語は，肉親関係，利害関係において自分に利益をもたらす単語であった．

表 3.11: 友好の性質を示す名詞

名詞	友好	名詞	友好
赤ちゃん	14	子	9
友達	13	娘	8
恋人	12	弟	8
保証人	12	妻	8
夫婦	12	兄弟	8
家族	11	おじいさん	8
年寄り	10	姉	7
孫	10	兄	7
妹	9	夫	6
母	9	人間	6
相談役	9	親	6
子供	9	女	6

表3.12に敵対関係を示す名詞を示す．敵対関係を示す名詞は，利害関係において，自分に害を及ぼす単語であった．

表3.13及び表3.14にそれぞれ主の関係，従の関係を示す単語を示す．主の関係は，主に社会的地位や家庭内での地位が高い単語が選ばれている．

表 3.12: 敵対関係の性質を示す名詞

名詞	敵対
敵，仇	13
嘘つき	10

表 3.13: 主の関係の性質を示す名詞

名詞	主	名詞	主
父	11	先輩	8
飼い主	10	夫	7
社長	10	大人	7
議長	9	母	7
主治医	9	兄	6
主人	9	姉	6
先生	9	委員	6
親	8	医者	6
幹部	8	担当	6

表 3.14: 従の関係の性質を示す名詞

名詞	従	名詞	従
赤ちゃん	8	弟	7
生徒	7	妹	6
子	6	子供 6	

動詞の分類アンケート

動詞の分類辞書を作成した．本研究では，3.3.2 項で説明したとおり，名詞と動詞の組み合わせにより動詞を分類する．そのためのアンケートは以下のようにして作成した．まず，IPAL 内にある動詞から自動詞を選び，自動詞から IPAL 内の意味を参照して,「人」を主語としたときに意味を持つ動詞を選択した．その結果，62 種類の動詞が選択された．

次に，名詞の感情分類アンケートにおいて，快と友好，不快と敵対がそれぞれ類似した性質を持つという結果が得られたため，それぞれを一つとして，名詞の分類辞書から，友好・敵対，主・従を抜き出した．抜き出した単語は，以下のようなものである．

- 友好-快
  子供，恋人，友達
- 敵対-不快
  嘘つき，敵
- 主
  飼い主，社長，父
- 従
  赤ちゃん，生徒，弟

これらの名詞と動詞のすべての組み合わせ (682 文) に対してアンケートを作成した．被験者は，18 名の男女に web 上で行い,「喜び」,「悲しみ」,「怒り」,「恐怖」,「驚き」,「嫌悪」,「無感情」さらに，意味がないと感じられた場合は「不可」を選択してもらった．データの数値を (ある感情を選択した人数) ÷ (全体の人数=18 人) で算出し，名詞の性質ごとに平均を取った．その結果を表 3.15 に示す．

表 3.15: 動詞の分類アンケート結果

性質	動詞	喜び	怒り	驚き	悲しみ	恐怖	嫌悪	無感情	不可
友好	会う	0.37	0.00	0.02	0.00	0.00	0.02	0.43	0.17
敵対	会う	0.00	0.00	0.00	0.03	0.06	0.33	0.36	0.22
主	会う	0.26	0.00	0.02	0.00	0.00	0.00	0.56	0.17
従	会う	0.24	0.00	0.06	0.00	0.00	0.00	0.48	0.22
友好	上がる	0.19	0.00	0.02	0.00	0.00	0.00	0.37	0.43
敵対	上がる	0.03	0.08	0.03	0.00	0.06	0.11	0.25	0.44
主	上がる	0.11	0.02	0.02	0.00	0.00	0.00	0.41	0.44
従	上がる	0.15	0.00	0.06	0.00	0.00	0.00	0.31	0.48
友好	あせる	0.02	0.02	0.46	0.17	0.06	0.04	0.20	0.04
敵対	あせる	0.39	0.03	0.22	0.00	0.03	0.03	0.22	0.08
主	あせる	0.00	0.02	0.41	0.11	0.04	0.04	0.33	0.06

表 3.15: 動詞の分類アンケート結果

性質	動詞	喜び	怒り	驚き	悲しみ	恐怖	嫌悪	無感情	不可
従	あせる	0.00	0.02	0.33	0.09	0.04	0.04	0.41	0.07
友好	当たる	0.37	0.00	0.15	0.00	0.00	0.00	0.13	0.35
敵対	当たる	0.06	0.06	0.06	0.03	0.03	0.22	0.19	0.36
主	当たる	0.26	0.02	0.22	0.00	0.00	0.02	0.17	0.31
従	当たる	0.30	0.00	0.15	0.00	0.00	0.00	0.22	0.33
友好	集まる	0.50	0.00	0.06	0.00	0.02	0.04	0.35	0.04
敵対	集まる	0.00	0.03	0.03	0.03	0.31	0.39	0.22	0.00
主	集まる	0.13	0.02	0.15	0.00	0.04	0.00	0.54	0.13
従	集まる	0.35	0.00	0.04	0.00	0.02	0.00	0.44	0.15
友好	暴れる	0.00	0.41	0.22	0.04	0.22	0.09	0.02	0.00
敵対	暴れる	0.00	0.31	0.06	0.00	0.39	0.17	0.06	0.03
主	暴れる	0.00	0.30	0.24	0.00	0.39	0.02	0.04	0.02
従	暴れる	0.00	0.30	0.31	0.00	0.26	0.04	0.06	0.04
友好	謝る	0.26	0.04	0.31	0.20	0.00	0.02	0.13	0.04
敵対	謝る	0.36	0.06	0.42	0.00	0.00	0.11	0.03	0.03
主	謝る	0.06	0.04	0.56	0.13	0.00	0.04	0.17	0.02
従	謝る	0.17	0.06	0.41	0.11	0.00	0.00	0.20	0.06
友好	生きる	0.87	0.00	0.00	0.00	0.00	0.00	0.13	0.00
敵対	生きる	0.06	0.14	0.00	0.00	0.11	0.28	0.36	0.06
主	生きる	0.61	0.00	0.00	0.00	0.00	0.00	0.37	0.02
従	生きる	0.83	0.00	0.00	0.00	0.00	0.00	0.15	0.02
友好	いる	0.89	0.00	0.00	0.00	0.00	0.00	0.11	0.00
敵対	いる	0.00	0.11	0.00	0.06	0.22	0.44	0.14	0.03
主	いる	0.39	0.00	0.00	0.02	0.00	0.00	0.57	0.02
従	いる	0.44	0.00	0.00	0.00	0.00	0.00	0.56	0.00
友好	動く	0.24	0.00	0.04	0.02	0.00	0.00	0.69	0.02
敵対	動く	0.00	0.03	0.03	0.00	0.22	0.25	0.44	0.03
主	動く	0.11	0.00	0.02	0.02	0.00	0.00	0.83	0.02
従	動く	0.26	0.02	0.04	0.00	0.00	0.00	0.69	0.00
友好	移る	0.00	0.00	0.04	0.20	0.00	0.00	0.31	0.44
敵対	移る	0.00	0.00	0.00	0.06	0.06	0.25	0.39	0.25
主	移る	0.00	0.00	0.09	0.11	0.00	0.00	0.39	0.41
従	移る	0.06	0.00	0.07	0.11	0.00	0.00	0.48	0.28
友好	起きる	0.30	0.00	0.00	0.00	0.00	0.00	0.69	0.02
敵対	起きる	0.00	0.03	0.00	0.00	0.11	0.17	0.61	0.08
主	起きる	0.13	0.00	0.00	0.00	0.00	0.00	0.83	0.04

表 3.15: 動詞の分類アンケート結果

性質	動詞	喜び	怒り	驚き	悲しみ	恐怖	嫌悪	無感情	不可
従	起きる	0.28	0.00	0.02	0.00	0.00	0.00	0.67	0.04
友好	落ち着く	0.43	0.00	0.02	0.00	0.00	0.00	0.54	0.02
敵対	落ち着く	0.11	0.06	0.00	0.03	0.06	0.17	0.53	0.06
主	落ち着く	0.28	0.00	0.00	0.02	0.00	0.00	0.69	0.02
従	落ち着く	0.44	0.00	0.07	0.00	0.00	0.00	0.46	0.02
友好	落ちる	0.00	0.00	0.30	0.44	0.07	0.02	0.06	0.11
敵対	落ちる	0.64	0.00	0.03	0.00	0.00	0.00	0.17	0.17
主	落ちる	0.02	0.00	0.26	0.37	0.04	0.00	0.15	0.17
従	落ちる	0.00	0.00	0.30	0.35	0.19	0.00	0.07	0.09
友好	衰える	0.00	0.00	0.19	0.74	0.02	0.00	0.02	0.04
敵対	衰える	0.58	0.00	0.08	0.03	0.00	0.00	0.22	0.08
主	衰える	0.00	0.00	0.17	0.61	0.02	0.02	0.15	0.04
従	衰える	0.00	0.00	0.17	0.63	0.02	0.02	0.15	0.02
友好	溺れる	0.00	0.00	0.31	0.46	0.19	0.02	0.00	0.02
敵対	溺れる	0.42	0.00	0.17	0.08	0.06	0.03	0.25	0.00
主	溺れる	0.00	0.00	0.43	0.46	0.07	0.00	0.04	0.00
従	溺れる	0.00	0.02	0.28	0.57	0.13	0.00	0.00	0.00
友好	帰る	0.13	0.00	0.00	0.48	0.00	0.00	0.35	0.04
敵対	帰る	0.50	0.00	0.06	0.03	0.00	0.11	0.28	0.03
主	帰る	0.24	0.00	0.00	0.09	0.00	0.00	0.65	0.02
従	帰る	0.15	0.00	0.06	0.24	0.00	0.00	0.50	0.06
友好	稼ぐ	0.59	0.00	0.19	0.02	0.00	0.00	0.17	0.04
敵対	稼ぐ	0.06	0.19	0.00	0.03	0.03	0.47	0.19	0.03
主	稼ぐ	0.57	0.00	0.00	0.00	0.00	0.00	0.37	0.06
従	稼ぐ	0.41	0.00	0.35	0.02	0.00	0.00	0.15	0.07
友好	構える	0.06	0.00	0.15	0.02	0.04	0.04	0.57	0.13
敵対	構える	0.00	0.11	0.03	0.00	0.17	0.19	0.39	0.11
主	構える	0.06	0.00	0.06	0.00	0.15	0.02	0.61	0.11
従	構える	0.04	0.00	0.17	0.00	0.07	0.04	0.59	0.09
友好	変わる	0.09	0.00	0.54	0.20	0.00	0.00	0.13	0.04
敵対	変わる	0.08	0.00	0.39	0.03	0.14	0.11	0.19	0.06
主	変わる	0.07	0.00	0.48	0.20	0.04	0.02	0.17	0.02
従	変わる	0.11	0.00	0.59	0.06	0.02	0.00	0.17	0.06
友好	考える	0.15	0.00	0.07	0.04	0.00	0.02	0.70	0.02
敵対	考える	0.03	0.03	0.00	0.00	0.19	0.22	0.53	0.00
主	考える	0.06	0.00	0.04	0.00	0.00	0.00	0.91	0.00

表 3.15: 動詞の分類アンケート結果

性質	動詞	喜び	怒り	驚き	悲しみ	恐怖	嫌悪	無感情	不可
従	考える	0.17	0.00	0.13	0.00	0.00	0.00	0.63	0.07
友好	去る	0.00	0.00	0.00	0.72	0.00	0.00	0.24	0.04
敵対	去る	0.78	0.00	0.00	0.03	0.00	0.03	0.17	0.00
主	去る	0.00	0.00	0.02	0.50	0.04	0.00	0.41	0.04
従	去る	0.00	0.00	0.04	0.50	0.04	0.00	0.39	0.04
友好	狂う	0.00	0.06	0.26	0.17	0.52	0.00	0.00	0.00
敵対	狂う	0.17	0.11	0.11	0.00	0.33	0.17	0.08	0.03
主	狂う	0.00	0.09	0.20	0.06	0.61	0.00	0.04	0.00
従	狂う	0.02	0.07	0.30	0.07	0.48	0.04	0.02	0.00
友好	堪える	0.02	0.11	0.11	0.26	0.00	0.04	0.33	0.13
敵対	堪える	0.08	0.03	0.11	0.06	0.03	0.19	0.36	0.14
主	堪える	0.02	0.15	0.09	0.11	0.00	0.04	0.48	0.11
従	堪える	0.02	0.07	0.09	0.20	0.02	0.06	0.43	0.11
友好	転がる	0.07	0.00	0.26	0.17	0.00	0.00	0.44	0.06
敵対	転がる	0.19	0.00	0.14	0.03	0.03	0.11	0.42	0.08
主	転がる	0.02	0.00	0.39	0.09	0.02	0.00	0.35	0.13
従	転がる	0.15	0.00	0.26	0.13	0.02	0.00	0.33	0.11
友好	冴える	0.44	0.00	0.17	0.00	0.06	0.00	0.28	0.06
敵対	冴える	0.03	0.11	0.08	0.00	0.19	0.28	0.22	0.08
主	冴える	0.41	0.00	0.13	0.00	0.02	0.00	0.39	0.06
従	冴える	0.35	0.00	0.24	0.00	0.00	0.00	0.28	0.13
友好	覚める	0.33	0.00	0.02	0.06	0.00	0.00	0.41	0.19
敵対	覚める	0.00	0.03	0.00	0.00	0.14	0.19	0.44	0.19
主	覚める	0.20	0.00	0.02	0.00	0.02	0.00	0.56	0.20
従	覚める	0.30	0.00	0.00	0.00	0.00	0.00	0.56	0.15
友好	騒ぐ	0.19	0.30	0.17	0.00	0.04	0.19	0.13	0.00
敵対	騒ぐ	0.03	0.36	0.03	0.00	0.14	0.33	0.11	0.00
主	騒ぐ	0.06	0.19	0.31	0.00	0.04	0.28	0.11	0.02
従	騒ぐ	0.07	0.37	0.15	0.00	0.02	0.33	0.04	0.02
友好	死ぬ	0.00	0.00	0.09	0.89	0.02	0.00	0.00	0.00
敵対	死ぬ	0.50	0.00	0.06	0.28	0.03	0.00	0.11	0.03
主	死ぬ	0.00	0.00	0.09	0.87	0.02	0.00	0.02	0.00
従	死ぬ	0.00	0.00	0.06	0.94	0.00	0.00	0.00	0.00
友好	痺れる	0.00	0.00	0.39	0.19	0.04	0.06	0.19	0.15
敵対	痺れる	0.39	0.03	0.08	0.03	0.03	0.00	0.31	0.14
主	痺れる	0.02	0.00	0.37	0.13	0.04	0.02	0.26	0.17

表 3.15: 動詞の分類アンケート結果

性質	動詞	喜び	怒り	驚き	悲しみ	恐怖	嫌悪	無感情	不可
従	痺れる	0.00	0.00	0.30	0.24	0.11	0.04	0.13	0.19
友好	締まる	0.04	0.00	0.04	0.04	0.02	0.00	0.31	0.56
敵対	締まる	0.06	0.00	0.00	0.03	0.03	0.06	0.42	0.42
主	締まる	0.04	0.00	0.02	0.06	0.00	0.00	0.39	0.50
従	締まる	0.04	0.02	0.07	0.02	0.04	0.00	0.30	0.52
友好	退く	0.00	0.02	0.04	0.33	0.00	0.02	0.50	0.09
敵対	退く	0.64	0.03	0.03	0.00	0.03	0.00	0.22	0.06
主	退く	0.00	0.00	0.13	0.30	0.00	0.02	0.48	0.07
従	退く	0.02	0.00	0.09	0.15	0.00	0.04	0.61	0.09
友好	優れる	0.70	0.00	0.13	0.04	0.00	0.04	0.07	0.02
敵対	優れる	0.06	0.11	0.03	0.03	0.22	0.47	0.06	0.03
主	優れる	0.59	0.00	0.09	0.00	0.00	0.00	0.26	0.06
従	優れる	0.63	0.00	0.17	0.00	0.00	0.06	0.07	0.07
友好	滑る	0.06	0.00	0.41	0.19	0.00	0.02	0.26	0.07
敵対	滑る	0.31	0.00	0.11	0.03	0.00	0.00	0.36	0.19
主	滑る	0.00	0.00	0.26	0.13	0.02	0.02	0.46	0.11
従	滑る	0.06	0.00	0.31	0.20	0.06	0.02	0.28	0.07
友好	座る	0.17	0.00	0.00	0.00	0.00	0.00	0.81	0.02
敵対	座る	0.00	0.06	0.03	0.00	0.00	0.14	0.72	0.06
主	座る	0.04	0.00	0.00	0.00	0.04	0.00	0.91	0.02
従	座る	0.20	0.00	0.00	0.00	0.00	0.00	0.76	0.04
友好	ずれる	0.00	0.00	0.06	0.06	0.00	0.04	0.30	0.56
敵対	ずれる	0.03	0.03	0.08	0.03	0.06	0.03	0.22	0.53
主	ずれる	0.00	0.00	0.15	0.04	0.00	0.02	0.28	0.52
従	ずれる	0.02	0.00	0.07	0.04	0.00	0.02	0.33	0.52
友好	倒れる	0.00	0.00	0.33	0.59	0.02	0.04	0.02	0.00
敵対	倒れる	0.61	0.00	0.11	0.03	0.00	0.03	0.22	0.00
主	倒れる	0.00	0.00	0.46	0.41	0.04	0.02	0.06	0.02
従	倒れる	0.00	0.00	0.43	0.50	0.04	0.02	0.02	0.00
友好	助かる	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
敵対	助かる	0.11	0.28	0.03	0.11	0.00	0.31	0.17	0.00
主	助かる	0.91	0.00	0.00	0.00	0.00	0.00	0.04	0.06
従	助かる	0.98	0.00	0.00	0.00	0.00	0.00	0.02	0.00
友好	戦う	0.09	0.17	0.46	0.13	0.00	0.02	0.09	0.04
敵対	戦う	0.00	0.14	0.11	0.00	0.14	0.22	0.36	0.03
主	戦う	0.15	0.22	0.39	0.02	0.00	0.00	0.13	0.09

表 3.15: 動詞の分類アンケート結果

性質	動詞	喜び	怒り	驚き	悲しみ	恐怖	嫌悪	無感情	不可
従	戦う	0.04	0.11	0.50	0.04	0.04	0.00	0.19	0.09
友好	立つ	0.17	0.02	0.11	0.02	0.00	0.00	0.61	0.07
敵対	立つ	0.00	0.00	0.00	0.00	0.17	0.19	0.50	0.14
主	立つ	0.04	0.00	0.04	0.02	0.04	0.00	0.72	0.15
従	立つ	0.30	0.04	0.09	0.00	0.00	0.00	0.50	0.07
友好	黙る	0.00	0.00	0.20	0.15	0.02	0.13	0.48	0.02
敵対	黙る	0.36	0.03	0.17	0.00	0.11	0.11	0.22	0.00
主	黙る	0.00	0.02	0.09	0.04	0.17	0.11	0.54	0.04
従	黙る	0.07	0.02	0.09	0.07	0.02	0.07	0.63	0.02
友好	疲れる	0.00	0.00	0.02	0.46	0.00	0.07	0.39	0.06
敵対	疲れる	0.61	0.00	0.03	0.03	0.03	0.03	0.25	0.03
主	疲れる	0.00	0.00	0.02	0.31	0.00	0.09	0.50	0.07
従	疲れる	0.04	0.00	0.06	0.31	0.00	0.06	0.48	0.06
友好	尽きる	0.00	0.00	0.00	0.69	0.02	0.02	0.06	0.22
敵対	尽きる	0.50	0.00	0.00	0.03	0.00	0.00	0.22	0.25
主	尽きる	0.00	0.00	0.06	0.50	0.00	0.00	0.07	0.37
従	尽きる	0.00	0.00	0.04	0.48	0.04	0.00	0.06	0.39
友好	つぶれる	0.00	0.02	0.22	0.48	0.15	0.02	0.06	0.06
敵対	つぶれる	0.58	0.00	0.03	0.06	0.03	0.00	0.22	0.08
主	つぶれる	0.02	0.00	0.33	0.28	0.07	0.04	0.11	0.15
従	つぶれる	0.00	0.00	0.24	0.33	0.26	0.02	0.06	0.09
友好	つまずく	0.06	0.00	0.39	0.28	0.00	0.04	0.22	0.02
敵対	つまずく	0.44	0.00	0.06	0.06	0.00	0.00	0.42	0.03
主	つまずく	0.02	0.00	0.41	0.15	0.02	0.04	0.31	0.06
従	つまずく	0.04	0.00	0.39	0.24	0.04	0.00	0.28	0.02
友好	治る	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.04
敵対	治る	0.31	0.06	0.14	0.00	0.03	0.25	0.17	0.06
主	治る	0.85	0.00	0.00	0.00	0.00	0.00	0.11	0.04
従	治る	0.96	0.00	0.00	0.00	0.00	0.00	0.02	0.02
友好	泣く	0.00	0.02	0.17	0.70	0.00	0.09	0.02	0.00
敵対	泣く	0.25	0.06	0.22	0.08	0.00	0.11	0.28	0.00
主	泣く	0.00	0.00	0.30	0.61	0.00	0.02	0.06	0.02
従	泣く	0.02	0.06	0.13	0.69	0.00	0.07	0.04	0.00
友好	眠る	0.30	0.00	0.00	0.06	0.00	0.00	0.63	0.02
敵対	眠る	0.14	0.00	0.06	0.00	0.00	0.11	0.69	0.00
主	眠る	0.06	0.00	0.00	0.00	0.00	0.00	0.94	0.00

表 3.15: 動詞の分類アンケート結果

性質	動詞	喜び	怒り	驚き	悲しみ	恐怖	嫌悪	無感情	不可
従	眠る	0.26	0.02	0.00	0.04	0.00	0.02	0.65	0.02
友好	寝る	0.17	0.00	0.00	0.02	0.00	0.00	0.80	0.02
敵対	寝る	0.08	0.06	0.03	0.00	0.00	0.08	0.75	0.00
主	寝る	0.04	0.00	0.00	0.02	0.00	0.00	0.94	0.00
従	寝る	0.26	0.04	0.00	0.00	0.00	0.02	0.69	0.00
友好	這う	0.04	0.02	0.22	0.00	0.02	0.00	0.43	0.28
敵対	這う	0.00	0.00	0.08	0.03	0.22	0.08	0.42	0.17
主	這う	0.00	0.00	0.26	0.02	0.06	0.02	0.37	0.28
従	這う	0.09	0.04	0.20	0.02	0.02	0.02	0.39	0.22
友好	走る	0.19	0.00	0.15	0.00	0.00	0.02	0.65	0.00
敵対	走る	0.03	0.03	0.00	0.00	0.14	0.22	0.58	0.00
主	走る	0.07	0.02	0.20	0.00	0.00	0.00	0.65	0.06
従	走る	0.20	0.00	0.17	0.00	0.00	0.00	0.57	0.06
友好	働く	0.37	0.00	0.15	0.06	0.00	0.02	0.41	0.00
敵対	働く	0.06	0.06	0.11	0.00	0.11	0.22	0.44	0.00
主	働く	0.33	0.00	0.00	0.02	0.00	0.00	0.59	0.06
従	働く	0.28	0.00	0.28	0.00	0.00	0.00	0.33	0.11
友好	跳ねる	0.35	0.00	0.17	0.00	0.00	0.00	0.39	0.09
敵対	跳ねる	0.08	0.00	0.11	0.00	0.06	0.22	0.44	0.08
主	跳ねる	0.20	0.00	0.20	0.00	0.00	0.00	0.48	0.11
従	跳ねる	0.41	0.00	0.22	0.00	0.00	0.02	0.33	0.02
友好	ひらめく	0.69	0.00	0.11	0.00	0.00	0.00	0.19	0.02
敵対	ひらめく	0.11	0.08	0.03	0.00	0.28	0.25	0.25	0.00
主	ひらめく	0.56	0.00	0.13	0.00	0.00	0.00	0.31	0.00
従	ひらめく	0.52	0.00	0.19	0.00	0.00	0.00	0.22	0.07
友好	太る	0.02	0.02	0.35	0.24	0.00	0.17	0.20	0.00
敵対	太る	0.28	0.00	0.25	0.00	0.00	0.11	0.36	0.00
主	太る	0.02	0.02	0.31	0.06	0.00	0.11	0.46	0.02
従	太る	0.06	0.00	0.37	0.09	0.02	0.11	0.35	0.00
友好	震える	0.00	0.00	0.24	0.24	0.30	0.06	0.13	0.04
敵対	震える	0.22	0.00	0.19	0.00	0.17	0.08	0.33	0.00
主	震える	0.00	0.02	0.20	0.09	0.39	0.04	0.24	0.02
従	震える	0.00	0.00	0.35	0.11	0.35	0.00	0.15	0.04
友好	ふるまう	0.50	0.00	0.09	0.00	0.00	0.00	0.28	0.13
敵対	ふるまう	0.08	0.06	0.11	0.00	0.08	0.25	0.25	0.17
主	ふるまう	0.44	0.00	0.06	0.00	0.00	0.00	0.33	0.17

表 3.15: 動詞の分類アンケート結果

性質	動詞	喜び	怒り	驚き	悲しみ	恐怖	嫌悪	無感情	不可
従	ふるまう	0.35	0.00	0.13	0.00	0.00	0.00	0.30	0.22
友好	吠える	0.00	0.15	0.56	0.00	0.11	0.07	0.07	0.04
敵対	吠える	0.08	0.22	0.19	0.00	0.17	0.25	0.06	0.03
主	吠える	0.02	0.19	0.33	0.00	0.28	0.02	0.11	0.06
従	吠える	0.02	0.24	0.37	0.00	0.19	0.07	0.06	0.06
友好	迷う	0.07	0.02	0.26	0.22	0.02	0.06	0.33	0.02
敵対	迷う	0.44	0.06	0.14	0.03	0.00	0.03	0.28	0.03
主	迷う	0.00	0.04	0.15	0.17	0.09	0.09	0.39	0.07
従	迷う	0.00	0.00	0.19	0.19	0.00	0.07	0.43	0.13
友好	儲かる	0.67	0.00	0.17	0.02	0.02	0.00	0.07	0.06
敵対	儲かる	0.08	0.22	0.00	0.03	0.06	0.50	0.11	0.00
主	儲かる	0.81	0.00	0.02	0.00	0.00	0.00	0.17	0.00
従	儲かる	0.41	0.00	0.31	0.00	0.00	0.02	0.17	0.09
友好	痩せる	0.04	0.00	0.39	0.41	0.02	0.00	0.13	0.02
敵対	痩せる	0.31	0.00	0.14	0.00	0.03	0.06	0.39	0.08
主	痩せる	0.02	0.00	0.31	0.30	0.02	0.02	0.31	0.02
従	痩せる	0.04	0.00	0.31	0.33	0.02	0.04	0.22	0.04
友好	笑う	0.98	0.00	0.00	0.00	0.00	0.00	0.02	0.00
敵対	笑う	0.14	0.22	0.00	0.00	0.06	0.42	0.17	0.00
主	笑う	0.80	0.00	0.02	0.00	0.02	0.00	0.17	0.00
従	笑う	0.93	0.00	0.00	0.00	0.00	0.02	0.06	0.00

表3.15の動詞の中で，主・従，友好・敵対の名詞との組み合わせにより感情ありと判断された単語の割合をを表3.16に示す．表中の分散とは，一つの感情に決定できずに他の感情にばらついてしまい，感情が特定できなかった単語である．どの性質の名詞と組み合わせても無感情が最も大きくなっている．また，表3.7と表3.16を比較すると，形容詞の方が，動詞より有感情の割合が高いことがわかる．これは，形容詞の方が動詞よりも感情に影響を与えやすいということを示唆するデータとなっている．また，どの性質の名詞と組み合わさった場合でも，喜びが一番高い比率となった．

友好・敵対の関係を表3.17に示す．表3.17における数値は，以下のようにして算出した．まず，0.4以上の単語を“感情あり”と判断した．ただし，1つの単語において，2つ以上が0.4以上の場合は，相対的に高い方の感情とした．次に，敵対と友好の相互の比率を式(3.6)及び式(3.7)で算出した．表3.17における/の右側の数値を式(3.6)で

表 3.16: 友好・敵対-主・従の辞書中の感情の割合

	喜び	怒り	驚き	悲しみ	恐怖	嫌悪	無感情	不可	分散
友好	19%	2%	8%	19%	2%	0%	27%	6%	16%
敵対	21%	0%	2%	0%	0%	8%	24%	5%	42%
主	16%	0%	8%	10%	2%	0%	37%	6%	21%
従	15%	0%	5%	11%	2%	0%	35%	0%	32%

表し，左側の数値を式 (3.7) で表す.

$$\frac{\text{敵対と友好で感情が一致した数}}{\text{敵対の感情の数}} \tag{3.6}$$

$$\frac{\text{敵対と友好で感情が一致した数}}{\text{友好の感情の数}} \tag{3.7}$$

例えば,「敵対」の喜びの項目と「友好」の悲しみの項目は,「敵対」の喜びの項目に該当する単語が，13 個あり，そのうち友好において,「悲しみ」になった単語は，10 個ある．この場合，交差する点の数値はで，0.77 となる.

表 3.17: 友好・敵対の関係

		友好						
		喜び	怒り	驚き	悲しみ	恐怖	嫌悪	無感情
敵対	喜び	0	0	0	0.77/0.89	0	0	0.08/0.06
	怒り	0	0	0	0	0	0	0
	驚き	0	0	0	0	0	0	0
	悲しみ	0	0	0	0	0	0	0
	恐怖	0	0	0	0	0	0	0
	嫌悪	1.00/0.42	0	0	0	0	0	0
	無感情	0	0	0	0	0	0	0.87/0.81

表 3.17 より，友好の名詞と組み合わせて悲しみと判断された動詞は，高い割合で，敵対の名詞と組み合わせて喜びと判断される．逆に，敵対の名詞と組み合わせて喜びと判断された動詞についても，高い割合で，友好の名詞と組み合わせて悲しみと判断される．このように分類された動詞としては,「去る」,「落ちる」,「倒れる」,「疲れる」,「つぶれる」,「衰える」といった動詞がある.

次に，敵対の名詞と組み合わせて嫌悪と判断されている動詞は，すべて友好の名詞と組み合わせて喜びと判断されている．しかし，逆はあまり高い割合をしめてはいなかった．このような動詞としては,「儲かる」,「稼ぐ」,「いる」,「笑う」といった動詞がある.

最後に，敵対の名詞と組み合わせて無感情と判断されている動詞は，友好の名詞と組み合わせても無感情と判断さるものが多かった．つまり，これらの動詞は，前に係る友好・敵対に依存せずに無感情の動詞になるものである．このような動詞には,「寝る」,「座る」,「眠る」「起きる」,「走る」,「考える」,「立つ」,「働く」,「動く」,「覚める」,「転がる」,「這う」,「締める」といった動詞がある．

表3.18は，主・従の関係を示している．各数値の算出方法は，表3.17の場合と同様のであり，友好・敵対の関係を主・従の関係に変えて行った．

表 3.18: 主・従の関係

		主						
		喜び	怒り	驚き	悲しみ	恐怖	嫌悪	無感情
従	喜び	0.89/0.8	0	0	0	0	0	0.11/0.04
	怒り	0	0	0	0	0	0	0
	驚き	0	0	0.66/0.4	0	0	0	0
	悲しみ	0	0	0.14/0.2	0.85/1.0	0	0	0
	恐怖	0	0	0	0	1.0/1.0	0	0
	嫌悪	0	0	0	0	0	0	0
	無感情	0	0	0.05/0.2	0	0	0	0.86/0.83

全体的に主・従の関係において，高い相関という結果になった．どちらの名詞の性質でも同一の感情になった動詞を以下に示す．喜びになったものには「助かる」「治る」「笑う」「生きる」「優れる」「ひらめく」「稼ぐ」「儲かる」といった動詞があり，悲しみになったものには「死ぬ」「泣く」「衰える」「溺れる」「去る」「尽きる」が，驚きになったものとしては「変わる」「謝る」が，恐怖になったものには「狂う」が，無感情になったものには「座る」「寝る」「動く」「起きる」「眠る」「黙る」「考える」「退く」「構える」「走る」「覚める」「立つ」「帰る」「会う」「疲れる」「いる」「落ち着く」「集まる」「堪える」があった．

3.5 結論

本章では，入力された文章から感情を認識することを目的として，大局的な感情理解と局所的な感情理解の2つの方法について検討を行った．

大局的な感情理解では，Text tiling algorithm を用いることにより，場面分割を行う．この際の用いる刻み幅と窓幅のパラメータを実験で求め検証を行い，ある程度の成果を得た．

また，場面内の感情表現語の頻度が大きいものが，その場面における大局的感情であると考え，検証実験を行った．今回用いた文章に対する結果は，高い精度で上述のルールが正しいことを立証するものであった．

一方，局所的な感情分類を行うために，感情表現要素辞書を作成した．辞書の作成方法は，形容詞文については，形容詞の意味ごとの辞書に対して被験者実験を行うことによって感情を分類した．形容詞的な名詞文に対しても同様に被験者実験を行うことにより，感情分類を行った．動詞文については，組み合わせる名詞の性質によって動詞の感情が変化するために，名詞との組み合わせにより辞書を分類した．また，心理学的に重要とされる“対人関係”という概念を使用し，「人」を主語としたときに意味を持つ動詞を辞書化した．

第4章　動作制御モデルの構築

4.1　目的

現在のコンピュータグラフィックス (CG) 要素技術はゲームソフト業界，映画業界に留まらず広範囲に浸透してきている．その理由として，計算機における画像表現能力の向上と価格破壊があげられる．その結果，個人レベルにおいても CG による映像コンテンツ制作は身近なものになると予想される．

しかし，CG による映像コンテンツの中でも，主要な位置にあるキャラクタアニメーションにおいては，まだ個人レベルの制作は困難なままである．何故ならば，ゲームソフト業界や映画業界などの CG 制作の現場で用いられているキャラクタアニメーション制作手法を個人レベルの制作に適用する場合，様々な問題が浮上してくるからだ．

例えば，従来からキャラクタアニメーション制作に利用されてきたモーションキャプチャは機材の設置場所の確保やコストの面から個人で利用できるものではない．また，個人で利用できるキャラクタアニメーション制作手法として．キーフレームアニメーション [32]，数値入力やプログラミングによるアニメーションなどが考えられるが，それらの手法は利用にあたり熟練した技能が必要であり，直感的にアニメーションを制作する事は困難である．

直感的にアニメーションを制作可能なシステムを構築する為には，まずあらかじめ動作に応じてその拘束条件と自由度を規定しなければならない．そして，それを基に様々な動作を生成・制御する為の制御パラメータを導出し，その制御パラメータを用いて自然かつ多様な動作を生成・制御可能なモデルを構築する必要がある．

動作生成・制御におけるモデル構築には，過去の研究では三角関数による関数近似を用いたもの [33] がある．この研究では計測した動作データ (各関節の関節角度の時系列データ) を離散コサイン変換 (DCT) により関数化し，その関数のパラメータ修飾により，1 つの動作から種々の動作を生成している．また，関数近似の動作だけでは状況に応じた動作ができないことから，一部の関節のみに逆運動学を用いることにより，状況に応じた動作の生成を可能としている．

同じく関数近似を用いたものに [34],[35] がある．この研究では，歩行動作データに対し関数近似による周波数解析を行うことで，歩行動作を基準となる基準歩行動作と動作の特徴を表す特徴量成分に分離し，それぞれをデータベースに登録している．歩行動作の生成は基準歩行動作と特徴量成分を足し合わせることにより行っている．また，特徴量成分の重みを変えることにより種々の歩行動作の生成を可能としている．

関数近似以外の手法としては，人間の筋肉を考慮した解剖学的に正当なモデル「ア

クションユニット」を用いたもの [36], [37] がある．この研究では，解剖学に従って各筋肉を正当な位置に配置したアクションユニットを用い，筋肉の収縮運動によって動作を生成している．

また，ラバノテーションと呼ばれる人物動作の記述法を用いて人物動作を記号によって表現する手法 [38] もある．人の動きの生成には，その関節の多さから多大な時間と労力を必要とするが，関節角度で動作を制御するのではなく，言葉や記号を用いて表現し，これを関節角度に変換できれば制作コストを削減することができる．ラバノテーション (Labanotation) はそのための最も一般的な動作の記述法であり，人の動作を体の部位で分割し，それをさらに時間で分割した基本動作を記号として体系化したものである．動作の記述は staff と呼ばれる楽譜のような 2 次元平面を記号で埋めていく作業となる．

いずれの研究も，動作における人間の各可動部位の運動をそれぞれ独立したものとして考えており，可動部位相互の連動性に言及したものはない．人間の動作は個々の可動部位を連動させて制御されているものであることから，その連動性をルール化し，そのルールの基で動作を生成・制御する必要性があると考えられる．そこで本研究では，可動部位の連動性に着目して動作生成・制御モデルの構築を行う．

具体的には，モーションキャプチャ装置で収集した人物動作データを基に，各部位の空間的及び時間的連動性を考慮して，その連動性の高い部位をまとめて動作生成・制御可能なモデルを構築する．次に，物理的な特徴を表す特徴量成分を実測動作データから抽出し，それに動作を制御するパラメータとしての重みを掛け合わせる事で多様な動作を生成・制御可能なモデルの構築を行う．なお本研究では，対象人物動作として，最も基本的で且つ多様性を保持する動作である歩行動作を取り扱う．

4.2 歩行動作制御モデル

4.2.1 目標システム

構築するシステム概念図を図 4.1 に示す.「普通に歩く」を初期状態とし，それに対して様々な指示を与える事により，自らの頭に描いているイメージに対応した歩行動作，目標動作を作成する．システムの要求定義は，イメージしている動作を精密且つ簡易的に作成可能な事である．

そこで，少ない指示で多様な歩行動作を精密に生成・制御可能なモデル (歩行動作制御モデル) を構築する必要がある．

4.2.2 歩行動作制御モデルの定義

歩行動作制御モデルとは，関節構造モデルと動作制御モデルの 2 つのモデルからなる．

関節構造モデルとは，歩行動作において個々の関節の動作がどのように連動しているのかをグループとして構造化したものである．このモデルにより，制御する部位を

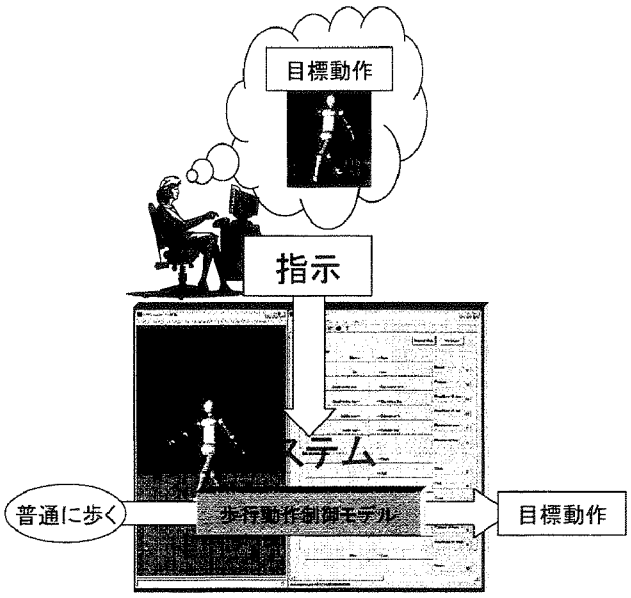


図 4.1: システム概念図

少なくし，連動性を基にしたリアルな歩行動作を生成・制御する事が可能となる．

動作制御モデルとは，物理的な特徴を表す複数の特徴量成分を保持し，それらを関節構造モデルに基づいて様々に付加する事で，多様な歩行動作を制御可能にしたモデルである．

歩行動作を生成・制御する際には，動作制御モデルが保持する特徴量成分を選択し，関節構造モデルにおける対応制御部位に，その特徴量成分に対して重みを掛けて付加する事になる．

4.2.3 歩行動作制御モデルの構築

対象動作データ

歩行動作には人それぞれの個性が反映されるが，モーションキャプチャ装置によって収録した多人数の動作データを用いて動作制御機構をモデル化すれば，動作者の個性を吸収した動作を生成・制御する事が可能となる．

1. 収録歩行動作

   モーションキャプチャ装置で収録した歩行動作は以下に示す 19 種類の歩行動作である．この 19 種類の歩行動作をそれぞれ 20 人分収録した．

   - 速く歩く
   - 遅く歩く
   - 俯いて歩く

- 胸を張って歩く
- 腕を大きく振って歩く
- 細かく歩く
- 大股で歩く
- 細かく歩く
- 腕を外向きにして歩く
- 腕を内向きにして歩く
- 蟹股で歩く
- 内股で歩く
- 縦に揺れながら歩く
- 横に揺れながら歩く
- 捻りながら歩く
- 肘を曲げながら歩く
- 頭を前に落として歩く
- 上を向きながら歩く
- 腿を大きく振りながら歩く

2. 動作データ形式

モーションキャプチャ装置で収録した人物動作データはBVH(BioVision Hierarchical data) 形式となっている．BVH 形式のファイルフォーマットは，モーションキャプチャデータの収録・販売を行っている Biovision 社によって定義されたものである．

BVH ファイルは，オブジェクト構造データとフレームデータの 2 つのデータから構成されている．

オブジェクト構造データは，19 の関節パーツの連結として定義されており，この 19 関節の連結を階層構造を用いて記述している．このオブジェクトはオブジェクト初期姿勢と呼ばれ，本研究で用いたオブジェクト (人体モデル) を図式化すると図 4.2 のようになる．オブジェクトは，図 4.3 に示すように，オブジェクトを構成するパーツを記号「{」「}」により再帰的に表現することで定義されている．

オブジェクト構造データは "HIERARCHY" というキーワードで開始される．その行に続く "ROOT Hips" はオブジェクトを定義する階層構造の一番上位にある関節を示している．人間を表すオブジェクトの場合，腰を表現するパーツを "ROOT Hips" として定義する．また，その他の関節は "JOINT 関節名" と表現される．

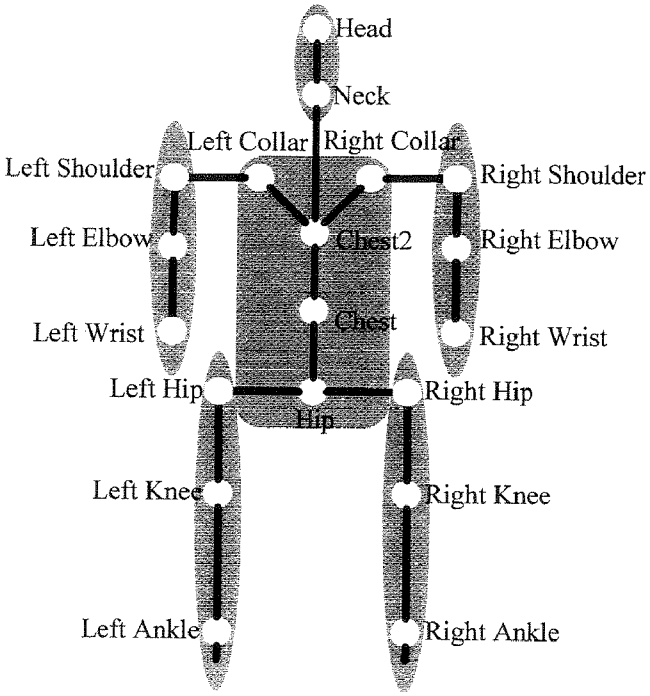


図 4.2: BVH の階層構造

次にキーワード "OFFSET" であるが，この後に続く 3 つの数字は "ROOT Hips" からの位置を示しており，それぞれ X 軸，Y 軸，Z 軸の順番で定義されている．

その下に続くキーワード "CHANNELS" には各関節が持つことのできるパラメータの数が定義されている．"ROOT Hips" は回転情報のほかオブジェクト全体の位置を示す情報を保持している為，3 次元の位置情報と 3 軸方向の回転情報の合計 6 つのパラメータが定義されている．"JOINT" は，"OFFSET" に定義されている "ROOT Hips" との相対位置からその位置が決定できる為，"JOINT" に繋がる関節の 3 軸方向の回転情報のみが定義される．従って，"ROOT Hips" は "CHANNELS 6 Xposition Yposition Zposition Zrotation Xrotation Yrotation" の様に定義され，"JOINT" は "CHANNELS 3 Zrotation Xrotation Yrotation" の様に定義される．

最後にオブジェクトの先端 (例えば手の先や足の先) は "End Site" として "ROOT Hips" からの位置を定義してある．

フレームデータは，フレーム毎の "ROOT Hips" の位置データと "JOINT" 及び "ROOT Hips" の回転データから構成されている．各関節における Zrotation, Xrotation, Yrotation の回転軸は，図 4.4 のように設定されている．

フレームデータは "MOTION" というキーワードで開始され，図 4.3 のオブジェクト構造データに続き，図 4.5 の様に記述されている．

"MOTION" の次に続く "Frames:" はデータ全体の長さを示すパラメータである．

```
HIERARCHY
ROOT Hips
{
  OFFSET 0.0000 0.0000 0.0000
  CHANNELS 6 Xposition Yposition Zposition Zrotation Xrotation Yrotation
  JOINT LeftHip
  {
    OFFSET 3.4300 -0.0000 -0.0000
    CHANNELS 3 Zrotation Xrotation Yrotation
    JOINT LeftKnee
    {
      OFFSET 0.0000 -18.4700 0.0000
      CHANNELS 3 Zrotation Xrotation Yrotation
      JOINT LeftAnkle
      {
        OFFSET -0.0000 -17.9500 0.0000
        CHANNELS 3 Zrotation Xrotation Yrotation
        End Site
        {
            OFFSET -0.0000 -3.1200 0.0000
        }
      }
    }
  }
  JOINT RightHip

                                    ⋮
```

図 4.3: BVH ファイル (オブジェクト構造データ)

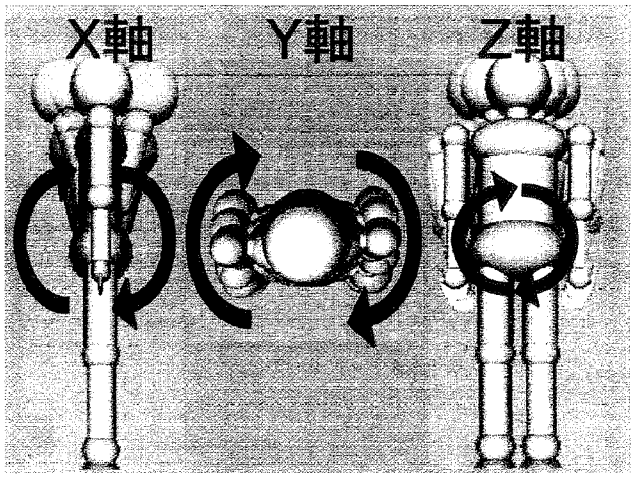


図 4.4: 3 回転軸

```
                                    ⋮

        CHANNELS 3 Zrotation Xrotation Yrotation
        JOINT Head
        {
          OFFSET -0.0000 5.1900 0.0000
          CHANNELS 3 Zrotation Xrotation Yrotation
          End Site
          {
              OFFSET -0.0000 4.1400 0.0000
          }
        }
      }
    }
  }
}
MOTION
Frames: 200
Frame Time: 0.0333333
```

12.330000	−7.790000	0.430000	−7.260000	−26.570000	−8.930000	⋯
0.000000	0.000000	0.000000	−0.100000	0.120000	−4.370000	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

図 4.5: BVH ファイル (フレームデータ)

図4.5のBVHデータは200フレーム記録されている例である．次に続く“Frame Time:”は1フレームあたりの秒数を示すパラメータである．図4.5の例では，1フレームあたりの秒数は“0.0333333”すなわち30分の1秒であり，このBVHデータは30fpsのモーションを記録していることになる．

この下に続くデータ列には，“ROOT Hips”の位置情報及び各関節の回転情報が行方向に時系列順に記録される．このデータ列は，オブジェクト階層データに記述した“JOINT”順に“CHANNELS”で指定したパラメータが記述される．

関節構造モデルの構築

前項で述べたように，人体モデルは19関節で定義されているため，分析対象データは19関節×3軸方向＝57関節動作から構成される．本項では，重回帰分析を用いて連動性を基に，この57関節動作を構造化する方法について述べる．

1. スペクトル情報への集約

   歩行動作における関節動作は全て周期的な運動をするため，連続時間周期信号として捉えることが可能である．よって，各関節動作信号$\Theta_m(t)$は式(4.1)のようにフーリエ級数で近似でき，

$$\Theta_m(t) = \sum_{k=-\infty}^{\infty} C_{mk} e^{i\frac{2\pi k}{T}t} \tag{4.1}$$

   各係数は式(4.2)のフーリエ級数展開により求められる．

$$C_{mk} = \int_{-\infty}^{\infty} \Theta_m(t) e^{-i\frac{2\pi k}{T}t} dt \tag{4.2}$$

   ここで，mは関節動作番号$(0 \leq m \leq 56)$を，kはスペクトルの高調波次数を表している．よって，$|C_{mk}|$は関節動作番号mの関節動作信号の第k高調波のスペクトルを表す．

   本研究で使用するモーションキャプチャデータは離散時間信号であるため，スペクトルは式(4.3)に示す離散フーリエ変換を用いて算出される．

$$C_{mk} = \sum_{n=0}^{N-1} \Theta_{mn} e^{-i\frac{2\pi nk}{N}} \tag{4.3}$$

   また，このフーリエ係数から式(4.4)に示す逆離散フーリエ変換を用いて，関節動作データが算出される．

$$\Theta_{mn} = \frac{1}{N} \sum_{k=0}^{N-1} C_{mk} e^{i\frac{2\pi nk}{N}} \tag{4.4}$$

過去の研究より，人物動作における関節動作は全て第3高調波までのスペクトル情報で十分再現可能であることが確認されている [34]．図 4.6 に，右膝の前後運動における第 3 高調波までの情報によって生成した動作データの波形と元動作データの波形を示す．図 4.6 より，本研究で使用するデータに対しても，第 3 高調波までの情報でほぼ元動作データが再現できていることがわかる．

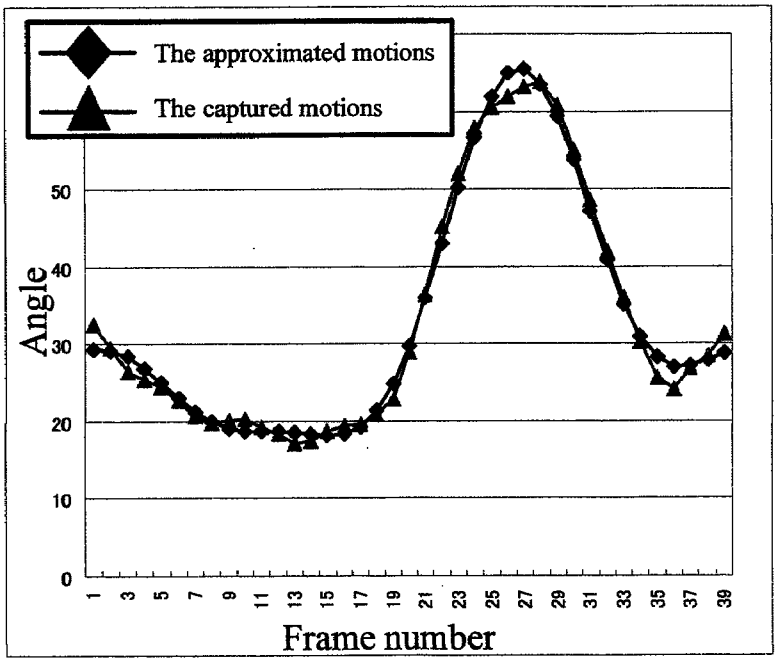


図 4.6: 第 3 高調波までを用いて生成した動作データ

時間領域において各関節動作を制御する場合，1 周期のフレーム数に相当する次元数のパラメータを制御する必要があるが，第 3 高調波までのスペクトルに情報を集約することで，制御する情報を簡略化することができる．そこで本研究では，全関節動作の時系列データを第 3 高調波までのスペクトル情報に集約しモデルを構築する．

2. 連動性のルールセットの導出

   関節動作間の連動性を明らかにするために，重回帰分析を用いて連動性のルールセットを導出する．まず，他の関節動作と連動性が高いと推測される関節動作をキー関節動作とする．次に，説明変数にキー関節動作のスペクトル情報を，目的変数にキー関節動作以外の関節動作のスペクトル情報を使用して，前者のデータから後者のデータを生成する関数群を重回帰分析によって導出する (図 4.7)．この関数群を関節動作間の連動性のルールセットとする．

3. 関節動作の構造化

   LeftHip，RightHip，LeftShoulder，RightShoulder の関節動作をキー関節動作 I として，他の関節動作との連動性のルールセットを導出した．次に，この連動性

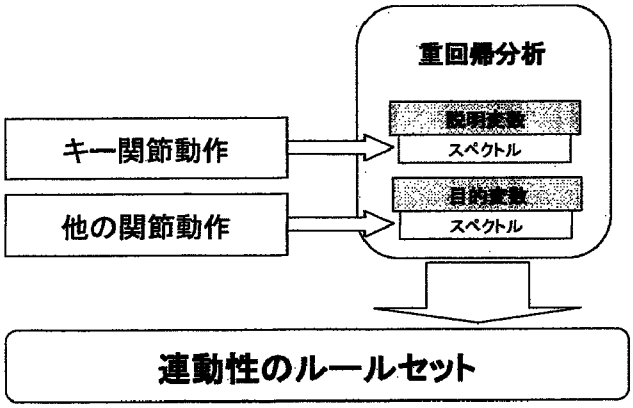


図 4.7: 連動性のルールセットの導出法

のルールセットの導出に使用していない歩行動作データのキー関節動作Iのスペクトル情報から，連動性のルールセットによりキー関節動作以外の関節動作データを導出する．この動作データと元動作データとの相関関係から，各関節ごとの動作の再現度がわかる．15人分の収録歩行動作に対して，関節動作ごとの相関係数の平均値を再現度とし，キー関節動作Iとの連動性の評価を行った．

表4.1に，分析結果の1つである右脚部位における再現度(相関係数の平均)を示す．表4.1から，再現度の高い関節動作と再現度の低い関節動作があることがわかる．

表 4.1: 再現度表 - 右脚部位

関節動作名	再現度
RightKnee-Z	0.8553
RightKnee-X	0.9410
RightKnee-Y	0.7048
RightAnkle-Z	0.6801
RightAnkle-X	0.8541
RightAnkle-Y	0.3531

ここで，再現度が0.5以上の関節動作を従属関節動作Iと設定する．従属関節動作Iは，連動性のルールセットから生成可能であることから，キー関節動作Iとの連動性が高いと推測される．つまり，従属関節動作Iはキー関節動作Iの変化に伴って規則的に変化するものとして扱うことができる．

また，再現度が0.5より小さい関節動作を独立関節動作Iと設定する．独立関節動作Iは，連動性のルールセットから生成困難であることから，キー関節動作Iとの連動性が低いと推測される．つまり，独立関節動作Iはキー関節動作Iの変化に伴って規則的に変化しないと考えられる．

次に，独立関節動作Iの中で，また新たにキー関節動作IIを設定し，同様の手法で関節動作の分類を行う．これにより，独立関節動作Iを，キー関節動作II，従属関節動作II及び独立関節動作IIに分類する．

これまでに分類した関節動作を図4.8に示す．全関節動作57動作の中で，歩行動作において連動して制御される関節動作は40動作(キー関節動作I・II，従属関節動作I・II)あり，独立に変化する関節動作は17動作(独立関節動作II)あった．

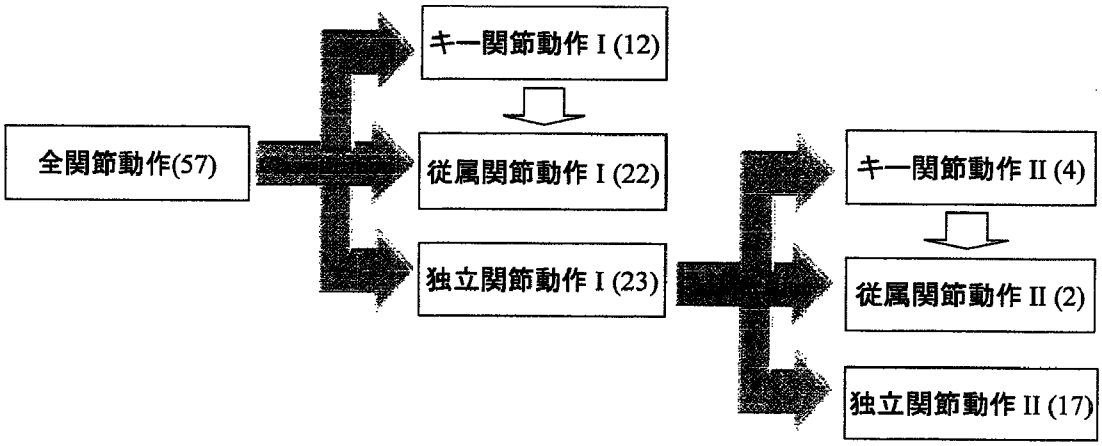


図 4.8: 関節動作の分類図

次に，各関節動作を連動して制御される関節群に分類し，歩行動作における関節動作の構造を明確化した．分類された結果を図4.9に示す．

関節動作は大きく分けて，互いに高い連動性を保持しながら運動している関節動作群と，そうでない関節動作群(グループ5)の2種類に分類される．そして，連動性の高い関節動作は更に4つのグループ(グループ1, 2, 3, 4)に分類される．

動作制御モデルの構築

1. 特徴量成分の抽出

歩行動作を制御する際に重要な動作制御項目として，「速さ」，「姿勢」，「腕の振り」，「脚の振り」，「腕の回転」，「脚の回転」，「縦揺れ」，「横揺れ」，「捻り」，「肘の曲げ」，「頭の姿勢」，「腿の振り」という12の項目が考えられる．そこで，それらの項目に対応した物理的な特徴を保持している19の歩行動作(モーションキャプチャ収録歩行動作)から，その特徴を表す特徴量成分を抽出する．表4.2に動作制御項目と抽出する特徴量成分を示す．

特徴量成分を抽出する手法としては，Unuma[34]やKraus[35]の手法が適用できる．これらは「楽しく歩く」などの感情を持った歩行動作データと「普通に歩く」の歩行動作データのスペクトルの差分を算出する事で，感情の成分を数量化したものである．本研究では，物理的な特徴を表す特徴量成分を抽出する為にこの手

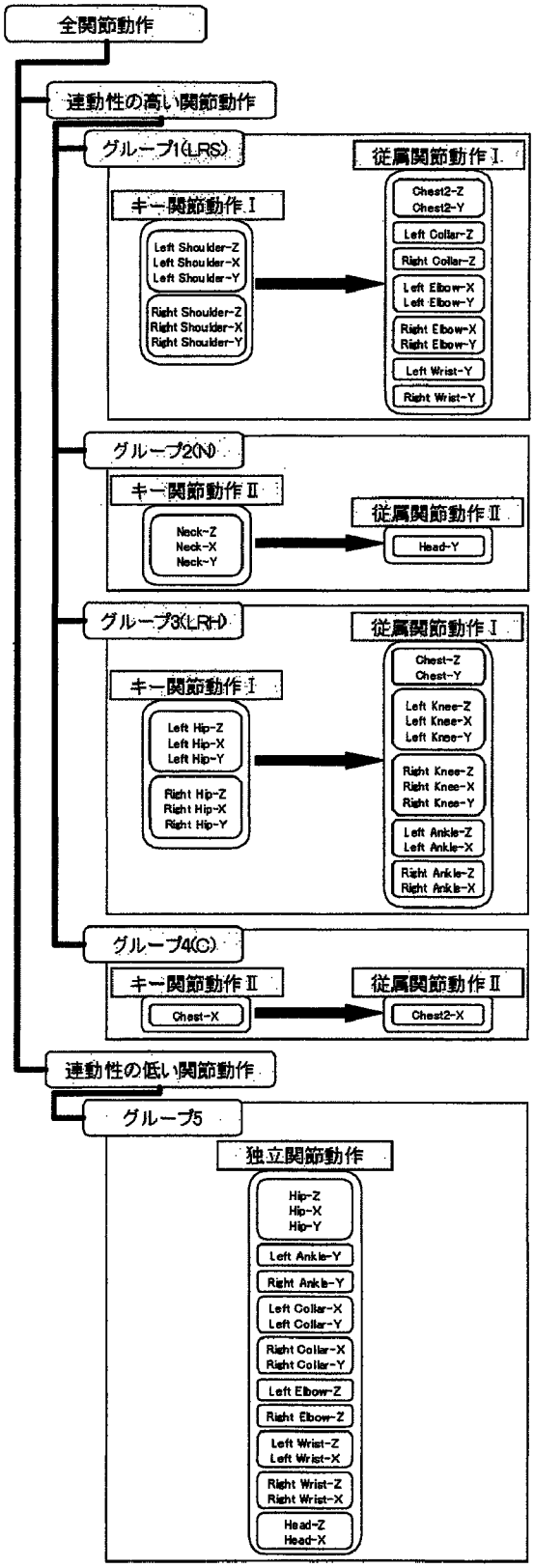


図 4.9: 関節動作グループ構造

表 4.2: 動作制御項目と特徴量成分

動作制御項目	特徴量成分 (抽出対象歩行動作)
速さ	速く歩く 遅く歩く
姿勢	俯いて歩く 胸を張って歩く
腕の振り	腕を大きく振って歩く 細かく歩く
脚の振り	大股で歩く 細かく歩く
腕の回転	腕を外向きにして歩く 腕を内向きにして歩く
脚の回転	蟹股で歩く 内股で歩く
縦揺れ	縦に揺れながら歩く
横揺れ	横に揺れながら歩く
捻り	捻りながら歩く
肘の曲げ	肘を曲げながら歩く
頭の姿勢	頭を前に落として歩く 上を向きながら歩く
腿の振り	腿を大きく振りながら歩く

法を用いる．ただし，抽出した特徴量成分に多様性を付加する為に，対象歩行動作データの分散を考慮する．

以下に抽出方法を述べる．まず，基準歩行動作「普通に歩く」のスペクトルデータ $\mu_{mk}^{(S)}$ を，モーションキャプチャで収録した20人分の「普通に歩く」のスペクトルデータを平均する事で算出する．また，その分散 $\left\{\sigma_{mk}^{(S)}\right\}^2$ を算出する．ここで，m は関節動作番号 $(0 \leq m \leq 56)$ を，k はスペクトルの高調波次数 $(0 \leq k \leq 3)$ を表す．

同様に，物理的な特徴を保持している19の歩行動作のスペクトルデータ $\mu_{mkl}^{(P)}$ を，モーションキャプチャで収録した20人分の歩行動作のスペクトルデータの平均から算出する．また，その分散 $\left\{\sigma_{mkl}^{(P)}\right\}^2$ を算出する．ここで，l は特徴量成分の番号 $(0 \leq l \leq 18)$ である．

次に，特徴量成分を式 (4.5) により算出する．

$$\delta_{mkl} \equiv \mu_{mkl}^{(P)} - \mu_{mk}^{(S)} \tag{4.5}$$

また，動作制御の多様化を図る為に，式 (4.5) で算出した特徴量成分に，実測データの分散からなる変動 λ_{mkl} を加える．これは，正規分布 $N\left(\mu_{mkl}^{(P)}, \sigma_{mkl}^{(P)}\right)$，$N\left(\mu_{mk}^{(S)}, \sigma_{mk}^{(S)}\right)$ に乱数を与えて算出される $\mu_{mkl}^{(P)\prime}$，$\mu_{mk}^{(S)\prime}$ を用いて，式 (4.6) により算出する．

$$\delta_{mkl} + \lambda_{mkl} \equiv \mu_{mkl}^{(P)\prime} - \mu_{mk}^{(S)\prime} \tag{4.6}$$

2. 動作制御パラメータを用いた動作制御

動作を制御する際には，基準歩行動作「普通に歩く」のスペクトルデータに対し，特徴量成分に分散による変動を加えて重みを掛け合わせたものを付加する．

よって，この重み (動作制御パラメータ) を変化させる事で，多様な歩行動作を生成・制御する事が可能になる．動作制御パラメータは動作制御項目毎に設定する．なお，2つの特徴量成分を保持する動作制御項目においては，パラメータを $+$ に移動させた場合には表4.2の上の特徴量成分 $(\delta_{mkr}^{+}, \lambda_{mkr}^{+})$ を，$-$ に移動させた場合は下の特徴量成分 $(\delta_{mkr}^{-}, \lambda_{mkr}^{-})$ を用いる (表4.3参照)．ここで，r は動作制御項目の番号 $(0 \leq r \leq 11)$ を表す．

したがって，動作制御によって生成される歩行動作の各関節動作データは式 (4.7) によって算出される．

$$\Theta_{mn} = \sum_{k=0}^{3} \left[\mu_{mk}^{(S)} + \sum_{r=0}^{11} \left\{ w_r (\delta_{mkr}^{+or-} + \lambda_{mkr}^{+or-}) \right\} \right] e^{i\frac{2\pi nk}{N}} \tag{4.7}$$

ここで，w_r は動作制御パラメータを表す．

表 4.3: 動作制御項目と特徴量成分の式

動作制御項目	特徴量成分 （抽出対象歩行動作）	式
速さ	速く歩く	$\delta^{+}_{mk0}, \lambda^{+}_{mk0}$
	遅く歩く	$\delta^{-}_{mk0}, \lambda^{-}_{mk0}$
姿勢	俯いて歩く	$\delta^{+}_{mk1}, \lambda^{+}_{mk1}$
	胸を張って歩く	$\delta^{-}_{mk1}, \lambda^{-}_{mk1}$
腕の振り	腕を大きく振って歩く	$\delta^{+}_{mk2}, \lambda^{+}_{mk2}$
	細かく歩く	$\delta^{-}_{mk2}, \lambda^{-}_{mk2}$
脚の振り	大股で歩く	$\delta^{+}_{mk3}, \lambda^{+}_{mk3}$
	細かく歩く	$\delta^{-}_{mk3}, \lambda^{-}_{mk3}$
腕の回転	腕を外向きにして歩く	$\delta^{+}_{mk4}, \lambda^{+}_{mk4}$
	腕を内向きにして歩く	$\delta^{-}_{mk4}, \lambda^{-}_{mk4}$
脚の回転	蟹股で歩く	$\delta^{+}_{mk5}, \lambda^{+}_{mk5}$
	内股で歩く	$\delta^{-}_{mk5}, \lambda^{-}_{mk5}$
縦揺れ	縦に揺れながら歩く	$\delta^{+}_{mk6}, \lambda^{+}_{mk6}$
横揺れ	横に揺れながら歩く	$\delta^{+}_{mk7}, \lambda^{+}_{mk7}$
捻り	捻りながら歩く	$\delta^{+}_{mk8}, \lambda^{+}_{mk8}$
肘の曲げ	肘を曲げながら歩く	$\delta^{+}_{mk9}, \lambda^{+}_{mk9}$
頭の姿勢	頭を前に落として歩く	$\delta^{+}_{mk10}, \lambda^{+}_{mk10}$
	上を向きながら歩く	$\delta^{-}_{mk10}, \lambda^{-}_{mk10}$
腿の振り	腿を大きく振りながら歩く	$\delta^{+}_{mk11}, \lambda^{+}_{mk11}$

ただし，関節構造モデルにより，歩行動作における関節動作は連動性を基に幾つかのグループに分類される事から，特徴量成分の付加を，その特徴量成分に影響する関節動作グループのみに対して行う．さらに，動作制御項目に応じて，付加する特徴量成分のスペクトルデータの高調波成分 ($f0, f1, f2, f3$) を取捨選択する．これにより，動作制御項目間の独立性を保ち，より精密な歩行動作を生成・制御することが可能となる．例えば,「脚の振り」の動作制御項目は，脚の部位の動きに影響しているので，脚の部位に多くの関節動作が見られる関節動作グループ 3(LRH) のみに，また，スペクトル成分の内波形に影響する ($f1, f2, f3$) 成分のみに特徴量成分を付加する事になる (図 4.10 参照).

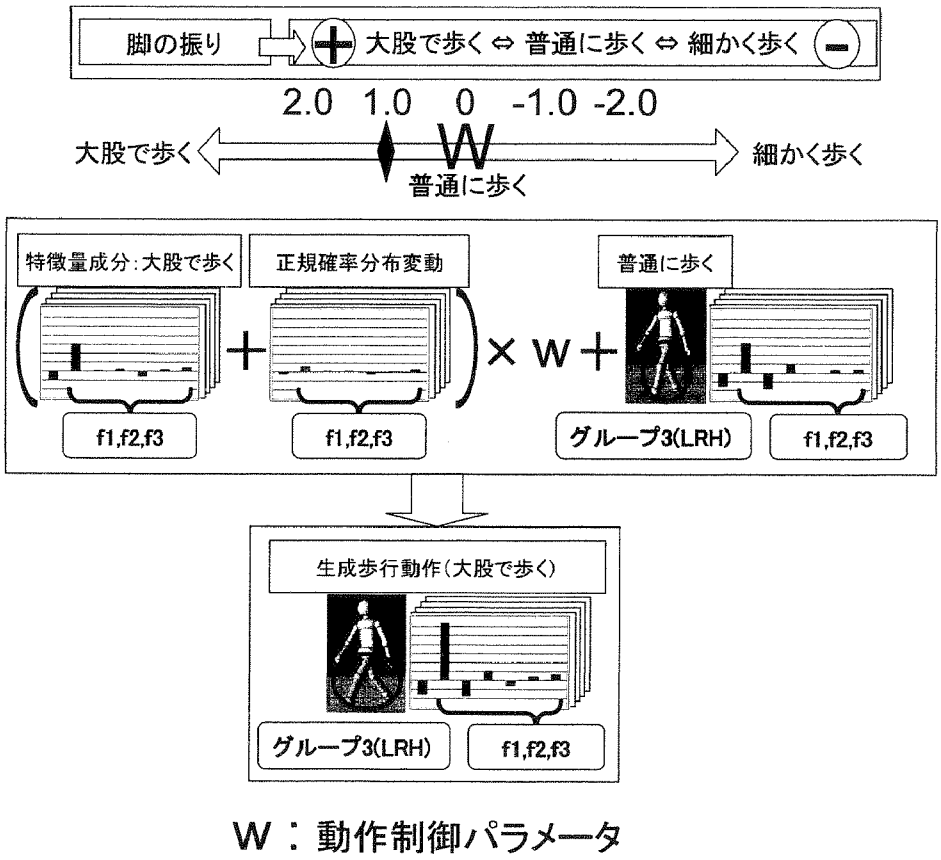


図 4.10: 動作制御の概念図

表 4.4 に動作制御項目とそれに対応する関節動作グループ及びスペクトル成分を示す．なお，グループ 5(独立関節動作) は他の関節動作と連動しない事から，どの特徴量成分を付加する際にも無視する．

図 4.11 に本モデルの構成図を示す．図 4.11 に示すように，本モデルの入力は，特徴量成分に掛け合わせる重み (動作制御パラメータ) と，特徴量成分に変動を付加する為の乱数であり，出力が生成動作の関節動作データとなる．

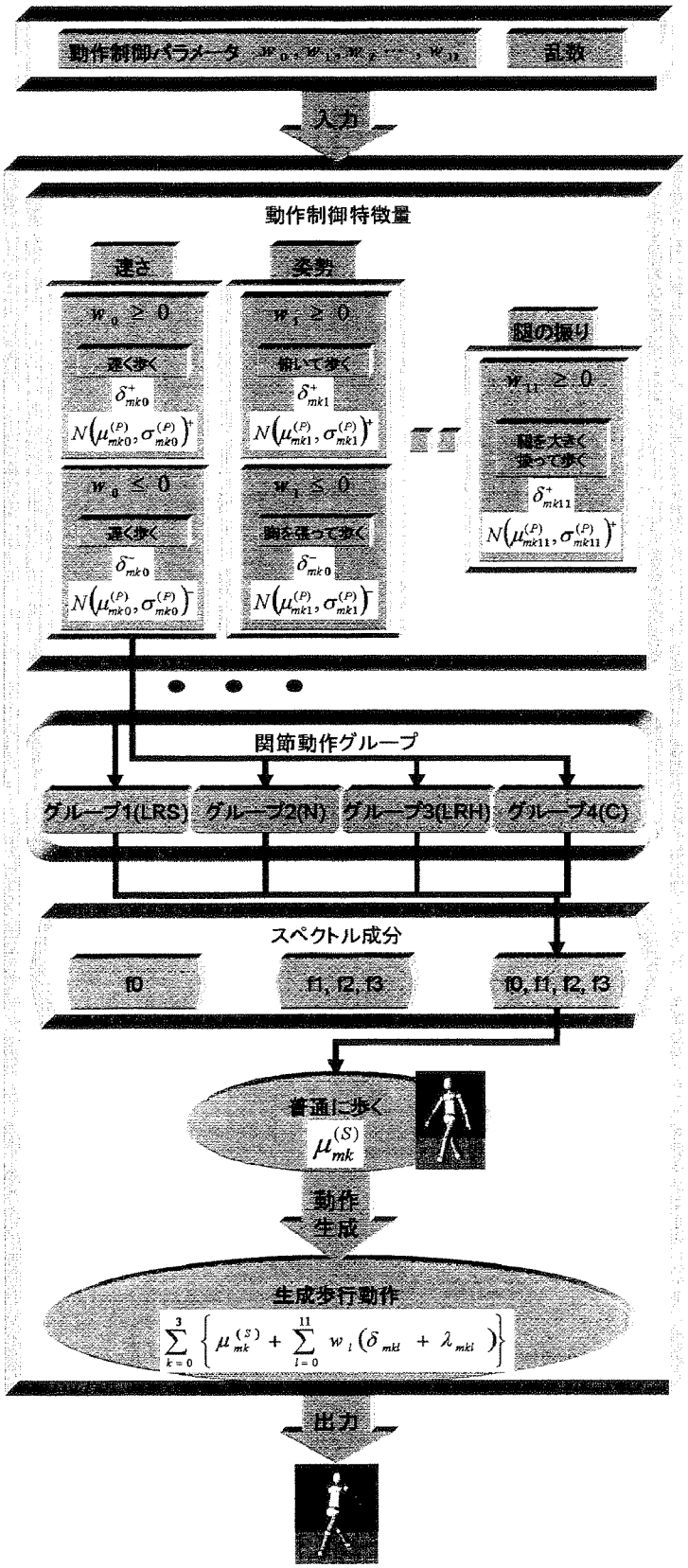


図 4.11: 歩行動作制御モデルの構成

表 4.4: 動作制御項目と関節動作グループ及びスペクトル成分

動作制御項目	関節動作グループ	スペクトル成分
速さ	グループ 1, 2, 3, 4	$f0, f1, f2, f3$
姿勢	グループ 1, 2, 3, 4	$f0$
腕の振り	グループ 1, 2	$f1, f2, f3$
脚の振り	グループ 3	$f1, f2, f3$
腕の回転	グループ 1	$f0$
脚の回転	グループ 3	$f0$
縦揺れ	グループ 2, 4	$f1, f2, f3$
横揺れ	グループ 1, 2, 3, 4	$f1, f2, f3$
捻り	グループ 1, 2, 3, 4	$f1, f2, f3$
肘の曲げ	グループ 1	$f0$
頭の姿勢	グループ 2	$f0$
腿の振り	グループ 3	$f0, f1, f2, f3$

4.3 モデル評価実験

本節では，本研究で構築したモデル評価実験について説明する．

4.3.1 評価実験手法

本モデルの精度は,「生成動作の多様性と精密性及び簡易性」によって評価できる．そこでまず，モーションキャプチャ装置で収録した 8 つの物理的な特徴を保持している歩行動作と，3 つの感情に関する特徴を保持している歩行動作 7 人分のデータを目標動作とし，目標動作とのスペクトルデータの誤差二乗和が最も小さくなるような歩行動作をモデルによってそれぞれ生成する．次に，式 (4.8) で表されるスペクトルデータの正規化誤差を用いて，モデルの精度の評価を行う．

$$\frac{\sum_{m=0}^{56}\sum_{k=0}^{3}\left[\mu_{mk}^{(G)}-\mu_{mk}^{(T)}\right]^2}{\sum_{m=0}^{56}\sum_{k=0}^{3}\left[\mu_{mk}^{(T)}\right]^2} \tag{4.8}$$

ここで，$\mu_{mk}^{(G)}$ は生成動作のスペクトルデータであり，$\mu_{mk}^{(T)}$ は目標動作のスペクトルデータである．この正規化誤差の値を各目標動作毎に平均し，それを評価実験結果とする．この正規化誤差の値が小さければ小さい程本モデルの精度は高いことになる．図 4.12 に評価実験の流れを示す．なお，図中の番号は実験の処理の順番を表す．

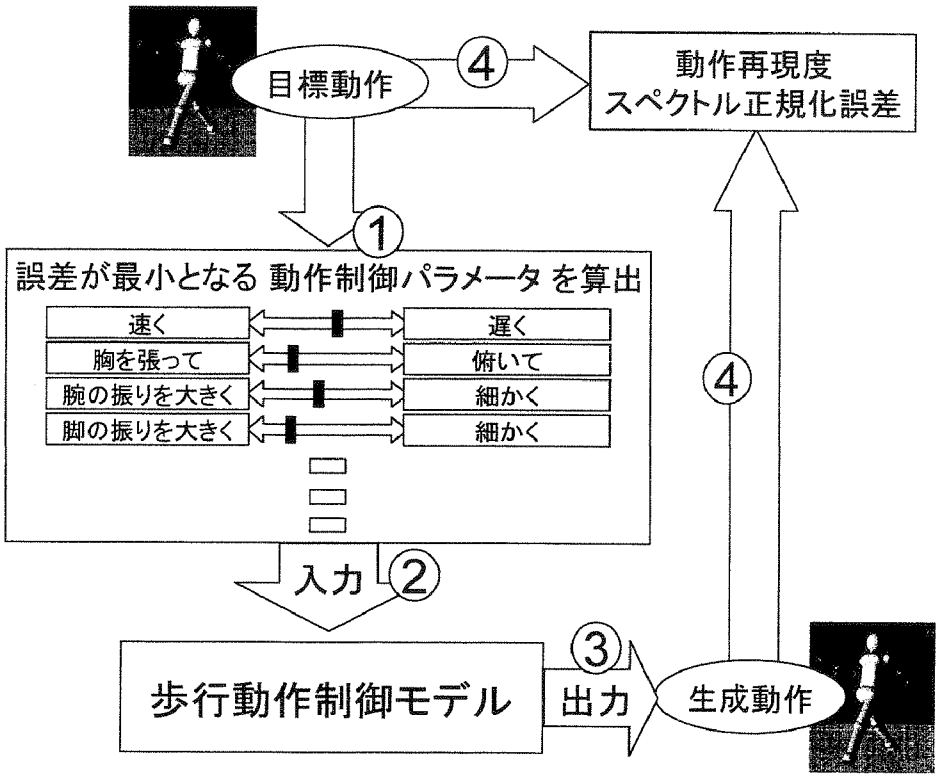


図 4.12: モデル評価実験図

4.3.2 評価実験結果

表 4.5, 4.6 にモデル評価実験結果を示す.

4.3.3 考察

評価実験により，物理歩行動作については本モデルでの動作再現性が高い事が解る．一方，感情歩行動作については動作再現性が比較的低い．ここから，本モデルが定量的な側面から判断すると，モデル構築に用いていない歩行動作 (感情歩行動作) に対して十分な動作再現性を満たしていない事が解る．これは，動作制御項目の種類が少ないという事に起因すると考えられる.

本実験では，主観的に (見た目で) どの程度の動作が再現できるのかという定性的判断ができない．そこで，4.4 節で，構築したモデルの定性的評価を実装システムを用いて行う.

表 4.5: モデル評価実験結果–物理歩行動作

歩行動作	正規化誤差
普通に歩く	0.2777
遅く歩く	0.2514
速く歩く	0.1967
腕を大きく振って歩く	0.2711
大股で歩く	0.2004
細かく歩く	0.2748
胸を張って歩く	0.2078
俯いて歩く	0.2688
平均	0.2436

表 4.6: モデル評価実験結果–感情歩行動作

歩行動作	正規化誤差
怒って歩く	0.3419
喜んで歩く	0.4962
悲しく歩く	0.2534
平均	0.3638

4.4 システム評価実験

構築した歩行動作制御モデルの定性的評価実験を，モデルを基に実装したシステムを用いて行う．この実験により，モデルの定性的評価だけでなく，システムとして実装した場合のインタフェースの評価も行う．

4.4.1 実装システム

構築した歩行動作制御モデルを基に歩行動作制御システムを実装した．システムはVC++で実装されている．インタフェースは図4.13に示すように，動作提示部と動作制御部からなり，動作制御部は12の動作制御パラメータと1周期のフレーム数を制御可能な13の動作制御スライダからなる．

4.4.2 評価実験手法

実験手法は，被験者に感情歩行動作(怒って歩く，喜んで歩く，悲しく歩く)を頭の中でイメージしてもらい，そのイメージ動作を目標動作として，その目標動作に近づけるよう，システムを用いて歩行動作を生成してもらう．次に，生成した歩行動作を頭の中でイメージした目標動作にどれだけ似せる事が出来たかという満足度としての動作再現度を，5段階評価のアンケート(5=非常に似ている，4=似ている，3=どちらでもない，2=似ていない，1=全く似ていない)で答えてもらう．また，実験を通してシステムの不満点・改良点をインタビューする．

図4.14に評価実験の流れを示す．なお，図中の番号は実験の処理の順番を表す．

4.4.3 評価実験結果

5人の被験者に行ったシステム評価実験の結果を表4.7に示す．

表4.7: システム評価実験結果

歩行動作	動作再現度
怒って歩く	4.2
喜んで歩く	4.6
悲しく歩く	4.6
平均	4.3

また，システムの不満点，改良点のインタビュー結果の主なものは,「腰の捻りを大きくしたい」,「姿勢の前屈みを大きくしたい」,「肘を下に曲げたい」などの動作制御

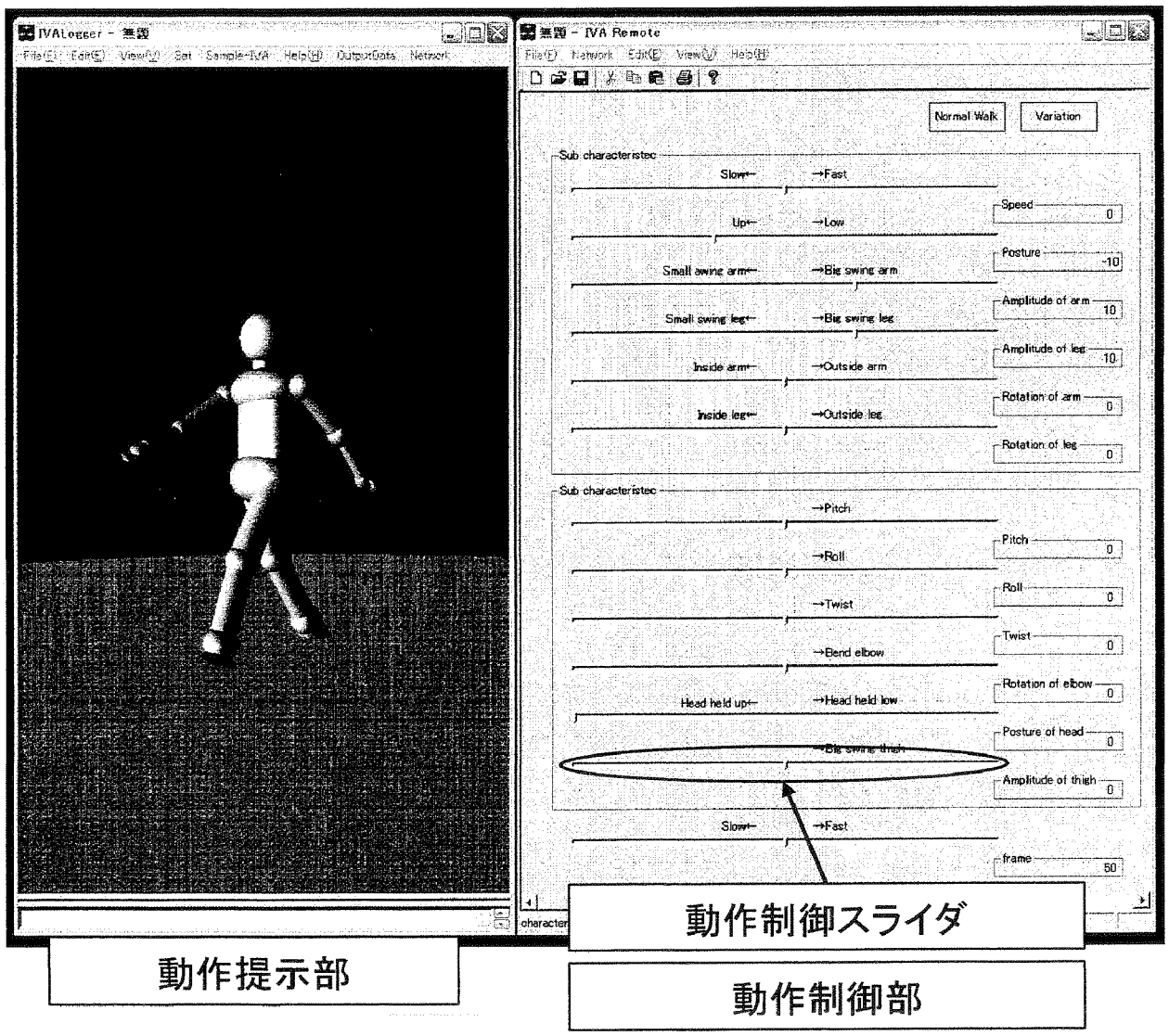

IVALogger - 無題
File(F) Edit(E) View(V) Set Sample-IVA Help(H) OutputData Network
無題 - IVA Remote
File(F) Network Edit(E) View(V) Help(H)
Normal Walk
Variation
Sub characteristec
Slow
Fast
Speed
0
Up
Low
Posture
-10
Small swing arm
Big swing arm
Amplitude of arm
10
Small swing leg
Big swing leg
Amplitude of leg
10
Inside arm
Outside arm
Rotation of arm
0
Inside leg
Outside leg
Rotation of leg
0
Sub characteristec
Pitch
Pitch
0
Roll
Roll
0
Twist
Twist
0
Bend elbow
Rotation of elbow
0
Head held up
Head held low
Posture of head
0
Big swing thigh
Amplitude of thigh
0
Slow
Fast
frame
50
動作制御スライダ
動作提示部
動作制御部

図 4.13: 実装システム

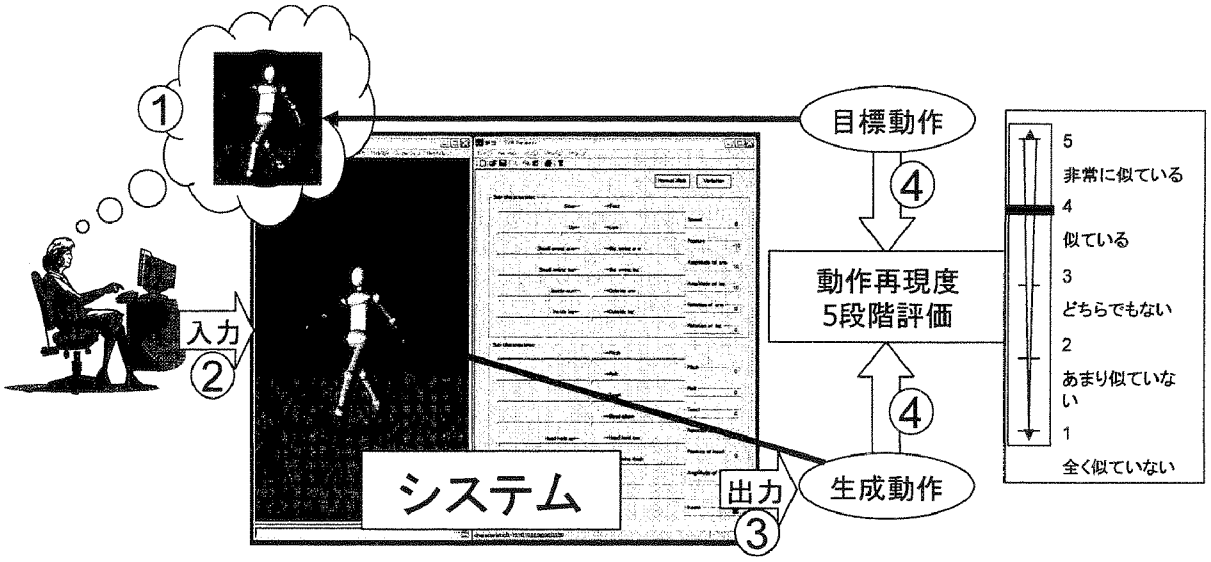


図 4.14: システム評価実験図

機構に関するものや「スライダが体のどの部位を動かすのかが解り難い」,「1 つ前の動作に戻れる機能がほしい」などのインタフェースに関するものなどであった.

4.4.4 考察

評価実験結果によると，どの歩行動作においても 5 段階評価指標の 4(似ている) より高く，本システムの定性的動作再現度が高い事が解る．つまり，本システムを用いれば，見た目ではほぼイメージ通りの歩行動作を生成する事が可能である事が解る．

またアンケート結果により，動作制御項目の種類数が，動作再現度の程度に深く関係している事が解る．つまり，動作制御項目を増やせば増やすほど，より多様な歩行動作を生成する事が可能となり，本モデル・システムの精度を表す精密性・多様性は向上すると考えられる．しかしその反面，制御しなければならないパラメータが増加する事によって，簡単に目標動作を生成する事が困難となり，本モデル・システムのもう 1 つの精度を表す簡易性は低下してしまう．被験者の中には，13 のスライダ全てを使用しない人もおり，13 ものスライダが必要ではないということも考えられる．

よって，本モデル・システムの相反する評価尺度，精密性・多様性，簡易性を考慮しながら，システム全体の精度を上げていく必要がある．その為には，多様な歩行動作を生成するのに必要な動作制御項目をより適切なものにしなければならない．アンケート結果によれば，腕の動きに対してより細かい指示を望む被験者が多く，今後は腕の細かい動きを表す動作制御項目をシステムに加える必要があると考えられる．

4.5 結論

本研究では，少ない指示で多様な歩行動作を精密に生成・制御可能な歩行動作制御モデルの構築を試みた．

まず，歩行動作において，全関節動作は連動性を基にいくつかの独立なグループに分類される事が 重回帰分析によって確認された．

次に，12 の動作制御項目に対応した 19 の特徴量成分を抽出し，これらの強さを制御する事で多様な歩行動作を生成・制御可能なモデルを構築した．

モデル評価実験及びモデルを実装したシステム評価実験により，定量的な側面からは，物理的な特徴を保持する歩行動作は十分な動作再現する事が可能であり，感情に関する特徴を保持する歩行動作については十分な動作再現を図る事が困難である事が判明した．一方，定性的な側面からは，感情に関する特徴を保持する歩行動作は十分な動作再現する事が可能である事が判明した．

以上の事から，本モデル・システムの有効性が示されたと考える．

第5章　感情動作の表出

本章では，CG キャラクタの動作を感情によって制御する方法について説明する．

感情動作を言語による指示で生成するには，動作と感情語との関係を記述した感情表出モデルを構築する必要がある．5.1 節では，感情表出モデルの構築と言語による指示により動作を生成・制御可能なディジタルアクタシステムの構築について述べる．

また，感情動作をスライダを用いた直感的なインタフェースで制御するためには，スライダの 1 目盛り毎に感情が均等に変化しなければならない．5.2 節では，心理的尺度を導入し，感情の変化と動作制御特徴量との間の関係を明らかにすることで，感情動作を均等かつ連続的に変化させることが可能なインタフェースの構築について述べる．

5.1　感情表出モデルの構築

5.1.1　目的

本節では，3DCG キャラクタアニメーション制作を目的としたディジタルアクタシステムを提案する．ディジタルアクタシステムとは，コンピュータ内の 3DCG 擬人化キャラクタ (アクタ) に対してユーザが言語で指示を与えることで，動作制御を行うシステムである．言語を用いた指示により，煩わしい数値プログラミング等が回避でき，監督が俳優に指示を与えるような方法で動作制御が可能となるため，容易で直感的なインタフェースを実現することができる．

言語を用いた指示から動作を生成・制御するためには，言語 (概念)—動作間の関係を明らかにしなければならない．言語による指示は,「大きく手を振って」といった物理的な変動を要求するものから「楽しそうに」といった感情に関係する変動を要求するものまで，様々ななものが存在する．本研究では，主に感情に関する変動を要求する動作指示 (感情語) により動作制御を行うことを目的とするため，感情語—動作間の関係を表現する「感情表出モデル」の構築を行う．

本研究では，感情を，喜び・悲しみ・怒り・驚きといったいくつかのカテゴリで表現されるものとしては扱わず，より多様なものとして捉える．したがって，いくつかのカテゴリの感情動作を収録して分析することによりモデルを構築するという手法はとらず，以下に示す方法でモデル構築用データの収集を行う．まず，被験者に目的動作を想像してもらい，それを元に自由に感情語による指示を行ってもらう．次に，第 4 章で示したシステムを使用して，被験者に想像した通りの動作を生成してもらう．得られた感情語と生成された動作の対応関係から，多次元動作空間中に感情語を表現す

る．このように被験者に自由な動作指示をしてもらうことにより多様性を保証することができる．さらに，この多次元空間から適切な動作を表出するためには，空間中に分布する多様な感情語相互の関係を明らかにする必要がある．本研究では，それらを木構造として表現することで感情表出モデルを構築する．

また，多様性を持たせてモデルを構築したとしても，未知語による指示が出現することは避けられない．そのため，未知語が入力された場合に新たに感情を学習し，感情表出モデルを更新する感情学習の機能についても検討を行う．

最後に，感情表出モデルに従い動作を制御する感情表出部と，未知語が入力された場合に感情表出モデルを更新する感情学習部を実装したディジタルアクタシステム (図 5.1) を構築する．

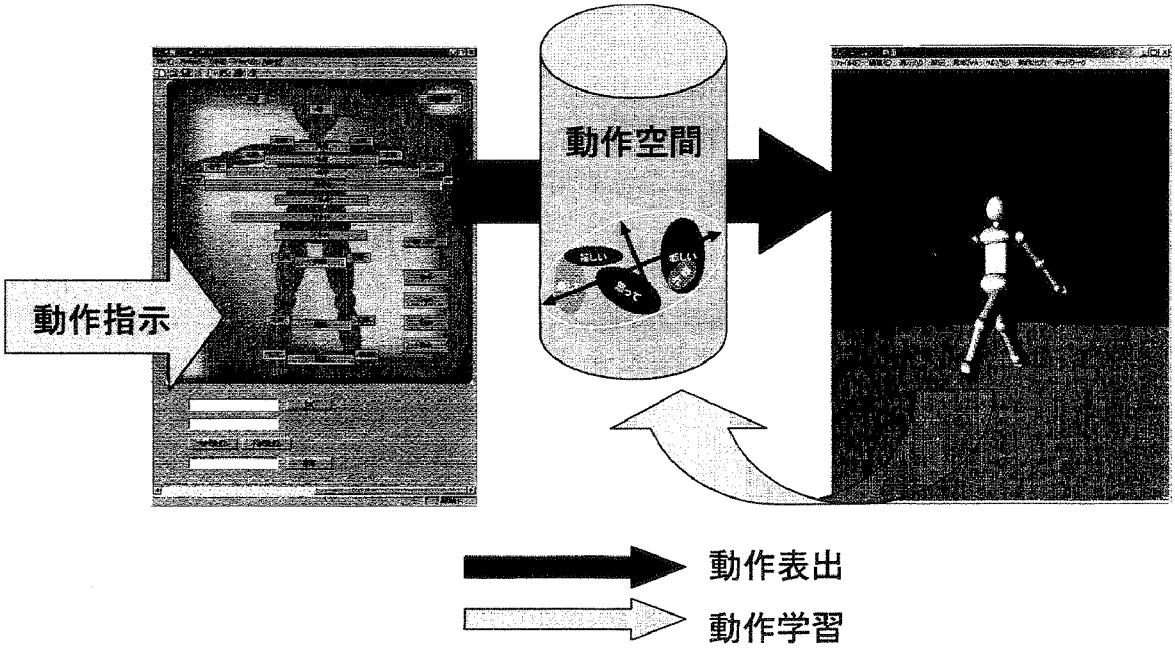


図 5.1: 感情表出モデルイメージ

5.1.2 データ収集

言語を用いた指示により動作を生成・制御するためには，感情語—動作間の関係を記述した感情表出モデルを構築しなければならない．したがって，感情表出モデルの構築には，様々な感情語に対応した動作データを収集する必要がある．直接的な方法としては，感情語を指定し，それに従って被験者に実際に動作をしてもらうことで動作データを収集する方法が考えられるが，指定した感情以外の動作データの収集が困難であるという欠点がある．そこで本研究では，被験者が目的動作を想像し，それに適した感情語を指定し，第 4 章で示したシステムを用いて想像した通りの動作を生成するという方法で，間接的にデータ収集を行う．このように被験者に自由な動作指示

を提示してもらうことにより感情の多様性を保証することができる.

本研究で行った動作収集の手順を以下に示す.

1. 状況提示

   全く何も指定されない状態で目的動作を被験者に想像してもらうことは困難であると考え，被験者の想像力が高まるような状況を予め用意し提示する．これは，台本に与えられた状況から言語による指示により動作を生成・制御するディジタルアクタシステムの仕組みにも対応している．また，本研究では歩行動作のみを対象にするため，状況は全て歩行動作に適用可能なものを選んでいる.

2. 動作想像

   被験者に1で与えられた状況に対して目的動作を想像してもらう.

3. 感情選択

   被験者に2で想像した目的動作を予め設定した基本4感情のいずれかに当てはめてもらう．基本4感情には，過去に心理学の研究から得られた8感情の内，歩行動作に付加し易い「歓喜」「悲嘆」「激怒」「恐怖・警戒」を選択した．このデータは，本研究で構築する感情表出モデルにおいて，従来の基本感情がどのように分布するかを考察するために使用するもので，感情表出モデルの構築には直接使用しない.

4. 感情語の付与

   被験者に2で想像した目的動作に対して感情語を付与してもらう.

5. 動作生成

   被験者に2で想像した目的動作を第4章で示したシステムにより生成してもらう．生成された動作データはBVHファイル形式で保存される.

以上の過程を図5.2に示す．被験者は男子大学生7人を対象とし，各60状況についてデータ収集を行なった．今回の実験で特に多く出現した感情語は「びくびく」「とぼとぼ」「肩を落として」「ルンルン」「いらいらして」「うきうき」「がっくり」等であり，これらは，歩行動作に対してイメージの沸き易い感情 (概念・語彙) であると考えられる．収集した感情語のうち，出現頻度が高かったものを表5.1に示す.

5.1.3 感情表出モデルの構築

感情表出モデルの構築方法

5.1.2項で収集したデータを元に，多次元動作空間に感情語を表現することができる．各感情語に対応した動作を生成するには，多次元動作空間中の1点に対応する動作データを生成すればよい．しかし，適切に動作を生成・制御するためには，各感情語間の関

図 5.2: データの収集方法

表 5.1: 収集した感情 (概念) 語

出現数	感情 (概念) 語
16	とぼとぼ　ビクビクしながら
14	ルンルン
13	いらいらして　肩を落として
11	むかつきながら
10	うきうきして
9	がっくりしながら
7	がっかりして　逃げるように
6	怒って　胸を張って　ワクワク
5	うなだれながら　おそるおそる　おどおど こそこそ　慎重に
4	うつむいて　ドキドキしながら 早足で　へこんで

係を明らかにする必要がある．本研究では，動作空間における距離関係に着目し，クラスター分析により各感情語を木構造として表現する．

使用する動作データは，5.1.2 項で収集した 420 動作である．各動作データから離散フーリエ変換によりパワースペクトルを求め，第 3 高調波までを特徴量として抽出する．また，全 57 関節動作のうち，4.2.3 項で示した独立関節動作を除いた計 40 関節動作を使用する．したがって特徴量は，1 動作につきパワースペクトル 4× 関節動作数 40=160 次元からなる．

クラスター分析には Ward 法を使用し，距離の近い動作データをつなぎ続け最終的に 2 分木を構成する．また，距離尺度には標準化ユークリッド距離を用いる．式 (5.1)

は動作データ x_{ki} と動作データ x_{kj} の標準化ユークリッド距離を表している．

$$D_{ij} = \sum_{k=1}^{160} (x_{ki} - x_{kj})^2 / \sigma_k^2 \tag{5.1}$$

ここで，σ_k^2 は動作データ x_k の分散である．

感情表出モデルの構築結果

本研究では，計 18 個のクラスターになるように 2 分木を構成した．結果を図 5.3 に示す．ツリーの高さは各クラスターの非類似度に対応している．クラスターには左端のものから順に A〜P と名付ける．図中の数字は，そのクラスター内 (又はツリーの下位) に存在する動作数である．また，斜字体の記号はそれぞれ Joy・Grief・Rage・Fear の頭文字であり，基本感情の「歓喜」「悲嘆」「激怒」「恐怖・警戒」に対応している．各記号は，そのクラスター内において過半数を占める基本感情を示しており，3 分の 1 以上半数未満の場合には括弧書きで示している．

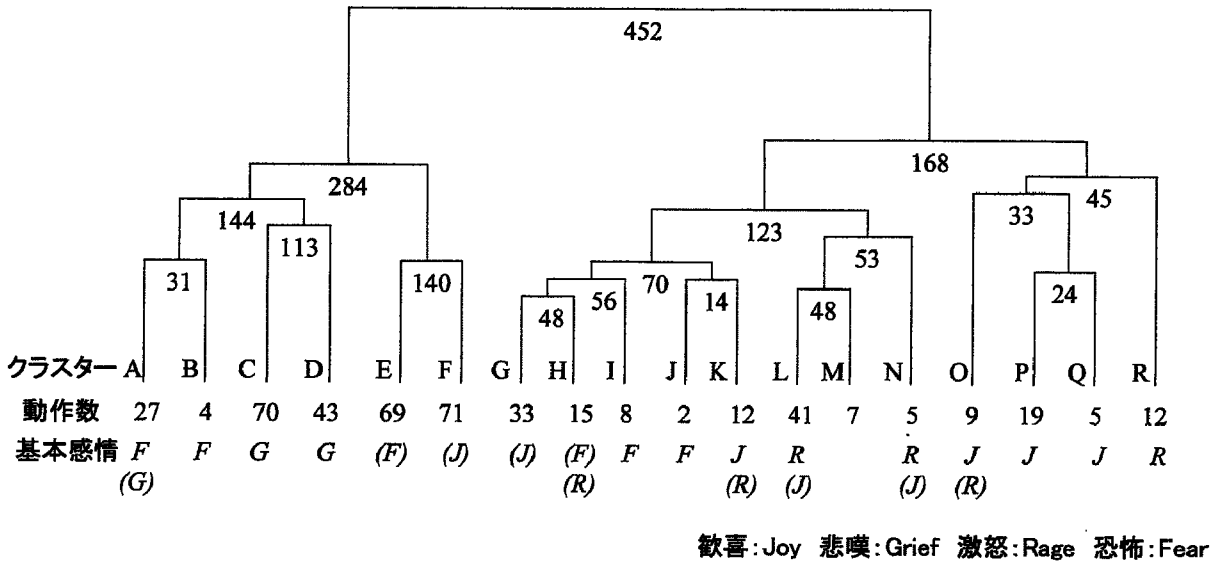


図 5.3: クラスター分析結果

1. 基本感情の分布

   各クラスター間における基本感情の割合を表 5.2 に示す．例えば，クラスター A には 27 動作が存在し，その内「悲嘆」に属する動作が 40.7%(11 個) 含まれているということを示している．

   図 5.3，表 5.2 から，まず,「歓喜」の感情を持つ動作が多く含まれるクラスターは K, O，P, Q を中心に多数存在することがわかる．したがって,「歓喜」の感情

表 5.2: 各クラスター間における基本感情の割合

クラスター	歓喜 [%](個)	悲嘆 [%](個)	激怒 [%](個)	恐怖・警戒 [%](個)	合計 [%](個)
A	0.0 (0)	40.7 (11)	0.0 (0)	59.3 (16)	100.0 (27)
B	0.0 (0)	25.0 (1)	0.0 (0)	75.0 (3)	100.0 (4)
C	2.9 (2)	68.6 (48)	7.1 (5)	21.4 (15)	100.0 (70)
D	16.3 (7)	55.8 (24)	11.6 (5)	16.3 (7)	100.0 (43)
E	34.0 (8)	30.9 (14)	13.8 (15)	21.3 (32)	100.0 (69)
F	39.5 (28)	22.5 (16)	14.1 (10)	23.9 (17)	100.0 (71)
G	39.4 (13)	6.1 (2)	24.2 (8)	30.3 (10)	100.0 (33)
H	20.0 (3)	0.0 (0)	40.0 (6)	40.0 (6)	100.0 (15)
I	12.5 (1)	12.5 (1)	25.0 (2)	50.0 (4)	100.0 (8)
J	0.0 (0)	0.0 (0)	0.0 (0)	100.0 (2)	100.0 (2)
K	58.3 (7)	0.0 (0)	41.7 (5)	0.0 (0)	100.0 (12)
L	39.0 (16)	2.4 (1)	53.7 (22)	4.9 (2)	100.0 (41)
M	14.2 (1)	28.6 (2)	28.6 (2)	28.6 (2)	100.0 (7)
N	40.0 (2)	0.0 (0)	60.0 (3)	0.0 (0)	100.0 (5)
O	66.7 (6)	0.0 (0)	33.3 (3)	0.0 (0)	100.0 (9)
P	100.0 (19)	0.0 (0)	0.0 (0)	0.0 (0)	100.0 (19)
Q	100.0 (5)	0.0 (0)	0.0 (0)	0.0 (0)	100.0 (5)
R	16.7 (2)	0.0 (0)	75.0 (9)	8.3 (1)	100.0 (12)

を持つ動作には多くのバリエーションが存在することが確認できる．しかし，全てのクラスターに満遍なく存在するのではなく，各クラスターにおける存在確率に偏りがあることがわかる．

次に「悲嘆」についてはC, Dを中心に密集して存在し，さらに「悲嘆」の感情を持つ動作のほとんどは最初の分岐よりも左に存在していることがわかる．したがって,「悲嘆」の感情を持つ動作は「歓喜」ほど多くのバリエーションは持たないと考えられる．

「激怒」の感情を持つ動作はL，N, Rを中心にいくつか存在し，ばらつきも大きい．よって,「激怒」の感情を持つ動作は非常にバリエーションが大きいと考えられる．

最後に「恐怖・警戒」の感情を持つ動作が多く含まれるクラスターはA, B, I, Jを中心にいくつか存在する．また「恐怖・警戒」は「悲嘆」の近辺に多く存在する傾向があり，動作自体も「悲嘆」と近いと推測される．また，動作の分散は「悲嘆」よりもやや広く，多少のバリエーションがあると考えられる．

総合的に見ると，ある程度の傾向は確認できるが,「激怒」の感情を持つ動作については明らかな傾向は現れていない．また，クラスターE，Fについては含まれる動作が多いにも関わらず，基本感情における傾向は見られなかった．

2. 感情語の分布

各感情語における所属クラスターの割合を表5.3に示す．ただし，この表は同一感情語が特に多く出現したもののうち上位9種(6個以上出現)のみに対するもので，クラスター欄には各感情語について2個以上出現したクラスターのみを記述している．例えば,「びくびく」という感情語を持つ動作は，全ての「びくびく」の動作の内，25.0%がクラスターFに，18.8%がクラスターC及びEのそれぞれに，12.5%がクラスターHに所属しているということである．

各感情語の内,「とぼとぼ」「うきうき」には，特定ないしは近隣のクラスターに集中する傾向が見られた．しかし,「びくびく」「いらいら」については，非常にバラツキが大きく，傾向はほとんど確認できなかった．以上より，感情語単位で見た場合，人によるイメージの違いが大きい感情語と小さい感情語が存在することがわかる．

表5.3: 各感情語における所属クラスターの割合

感情(概念)語	クラスター(割合[%])
とぼとぼ	C(62.5) D(18.8) E(12.5)
びくびくしながら	F(25.0) C,E(18.8) H(12.5)
ルンルン	P(28.6) F,L(21.4)
いらいらして	L(23.1) E,F,J,H(15.4)
肩を落として	C(30.8) D,F(23.1)
むかつきながら	F(27.3) L,R(18.2)
うきうきして	L(30.0) D,P(20.0)
がっくりしながら	D(44.4) C(33.3) E(22.2)
がっかりして	C(42.9) F(28.6)
逃げるように	E,F,G(28.6)

5.1.4 ディジタルアクタシステムの実装

本項では，5.1.3項で示した感情表出モデルを用いて動作を表出するディジタルアクタシステムについて述べる．本研究で提案するシステムは，木構造で表現されている感情表出モデルから目的動作を生成・制御する感情表出部と，未知語が入力された場合等に新たに感情語を学習し感情表出モデルを更新する感情学習部から成る．

感情表出部

図5.4に感情表出部の機構を示す．

提案システムの第1の目的は，ユーザの言語(感情語)による指示から目的動作を早く生成することである．したがって，木構造で表現されている感情表出モデルから目

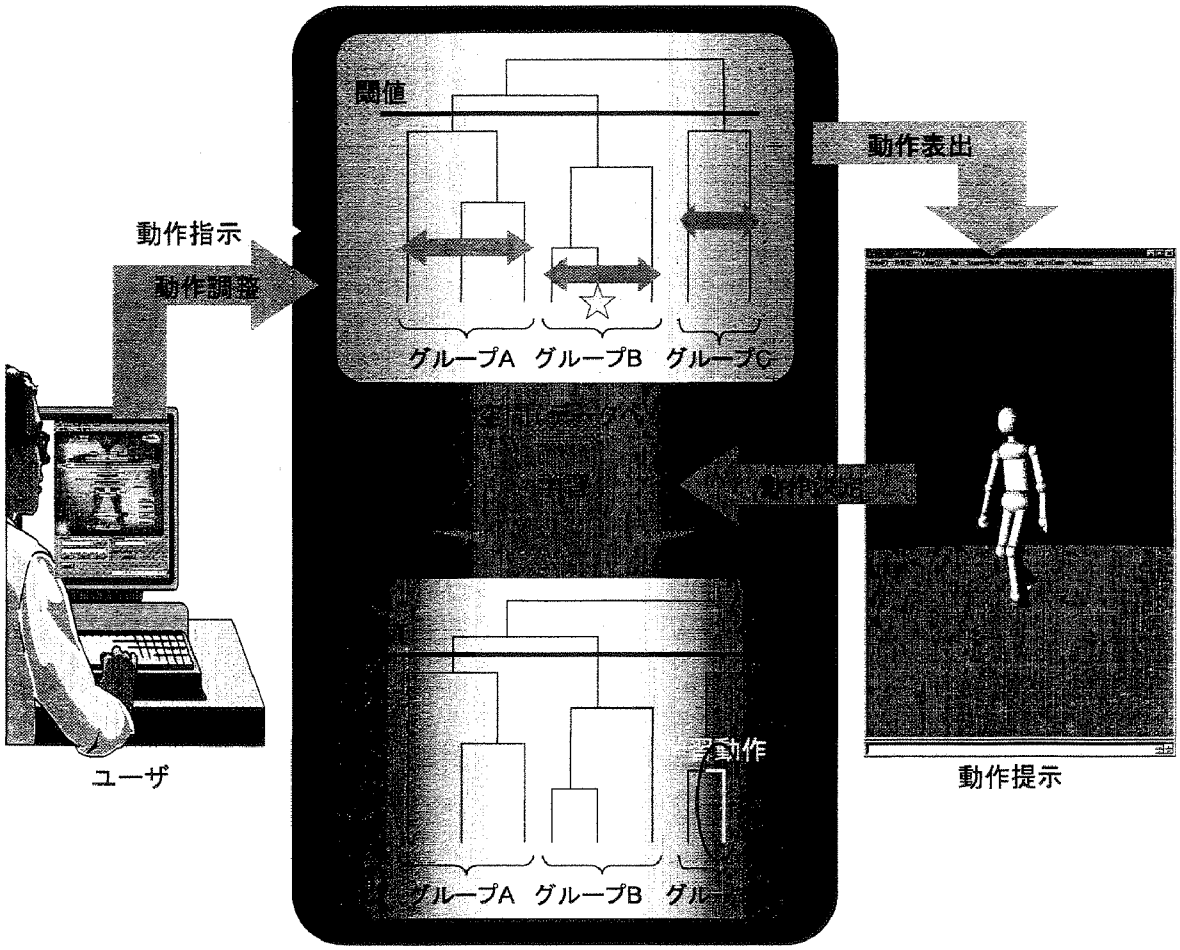


図 5.4: 感情表出部の機構

的動作を効率良く探索する必要がある．提案システムでは，入力された語が既に学習済みの語であった場合は，その語に対応する動作を表出する．ここで，対応する動作が複数存在し，それらが同一のクラスター内にある場合には，代表的な動作を1つ表出する．逆に，それらが複数のクラスターに分散して存在する場合には，各クラスターから代表的な動作をユーザに提示し選択してもらう．

提案システムの第2の目的は，目的動作を適切に生成することであるが，以上の方法で生成した動作は，予め学習されている動作そのものであるため，ユーザが意図したとおりの動作とは異なる場合がある．そこで，提案システムでは，表出された動作が満足であるか否かをユーザに判断してもらい，満足であれば動作生成を終了し，満足でなければ動作制御を行う．具体的には，提示した動作が存在しているクラスター内の動作において，分散の大きい特徴量成分を抽出し，その特徴量成分に対応する軸方向に変動を加えることで動作制御を行う．これをユーザに提示し，満足な動作が生成されるまで繰り返す．さらに不満が続く場合には，同一クラスター内では目的動作が生成不可能であると判断し，木構造において1つ上の階層に移動し，隣接するクラスター内から，同様の方法で新たに動作を生成・制御する．これを繰り返すことにより，感情表出モデルで表現される階層構造や包含関係を有効に利用し，より早く適切な動作を生成することが可能となる．

また，これを繰り返しても満足な動作が生成されない場合は，第4章で示したシステムを使用して目的動作を生成してもらう．

感情学習部

本研究で提案する感情表出モデルは，5.1.2項で述べたように，感情語の多様性を保証する形で収集されたデータを元に構築されている．しかし，多様性を持たせてモデルを構築したとしても，未知語による指示が出現することは避けられない．このため，未知語が入力された場合に新たに感情語を学習し，感情表出モデルを更新する必要がある．

提案システムでは，ユーザから未知語による指示があった場合，既知語が入力されるまでユーザに語の変更を依頼する．既知語が入力された場合の動作生成は感情表出部に従い，最終的に生成された動作を最初に入力された未知語と対応させて学習を行う．また，既知語が入力された場合でも，感情表出部で動作制御を行い新しい動作が生成された場合には，既知語と生成動作を対応させて学習を行うこととなる．

学習方法は，既に学習されているデータに新たに収集されたデータを加え，改めてクラスター分析を行い，感情表出モデルを更新するという方法である．これを繰り返すことで，洗練された感情表出モデルを構築することが可能となる．

実装システム

感情表出部及び感情学習部から成るディジタルアクタシステムを実装した．本システムはVisual C++を使用し，大きく2つのプログラムから成り立つ．一方は擬人化キャラクタの内部構造，動作指示が与えられた際の処理機構，アニメーションの表出の部分を記述したシステム部であり，もう一方はユーザの動作指示入力を受け，適切な命令をシステム部に送るインタフェース部である．

感情表出部のインタフェースは図5.5のようになる．ユーザがテキストボックスに命令を入力すると，その結果が下のテキストボックスに表示される．図5.5は,「わくわく」と入力し「3個発見しました」と表示される例である．ユーザは「OK」ボタンを繰り返し押すことで結果の各動作を表示させることができる．

提示された動作がユーザの意図したとおりの動作かどうかを判断するために「満足」「不満」ボタンが用意されている.「不満」ボタンが押されると，提示された動作に対して動作制御が行われ，その結果生成された新しい動作が表示される．最終的にユーザの意図したとおりの動作が表示された場合には,「満足」ボタンを押すことにより動作の表出は終了し，必要であれば感情表出モデルが更新される．

また，ユーザが未知語を入力した場合には「Not Found!」と表示され，他の語に言い換えて入力することとなる．このとき，ユーザが入力した最初の未知語(図5.5の例では「嬉しそうに」)は感情学習部のテキストボックスに表示され，最終的に適切な動作が生成された後,「学習」ボタンを押すことで感情表出モデルを更新することができる．

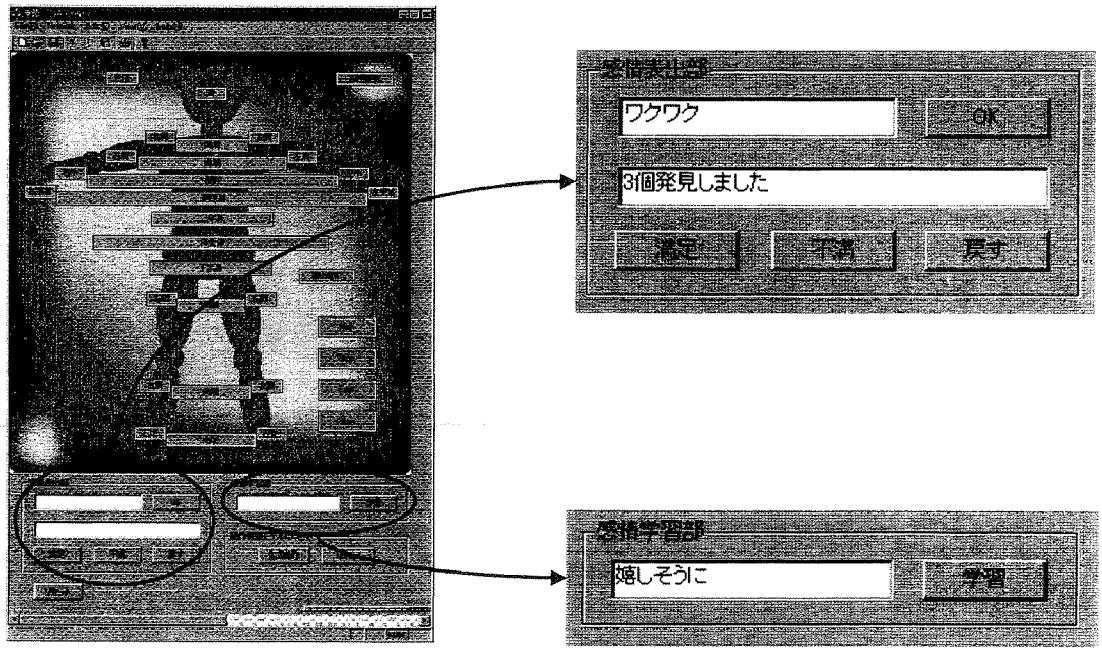


図 5.5: 作成インタフェース

5.1.5 評価実験

本項では，5.1.4 項で実装したシステムにより適切な動作が表出可能か否かを評価するための実験について説明する．

評価実験手法

評価実験の方法を以下に示す．まず被験者に目的動作を想像してもらい，実装システムを用いて目的動作を生成してもらう．次に，図 5.6 に示すように，生成した動作と感情表出モデルの各クラスターから選出した代表動作とを順次比較し，どちらが目的動作に近いかを評価してもらう．これにより，膨大な探索空間を全て探索する必要なく適切な動作が表出されているかどうか，あるいは，動作制御により不適切な動作が生成されていないかを評価することができる．

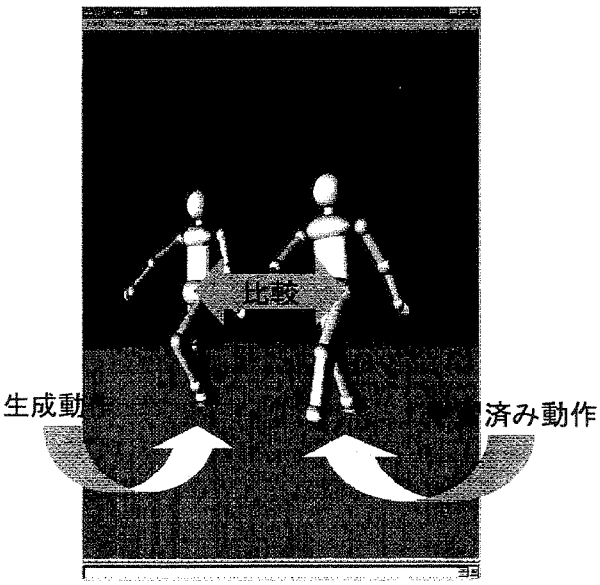


図 5.6: 評価実験イメージ

評価結果

5 名の男子大学生に各 3 つの動作を生成してもらい評価を行った．各クラスターから選出した比較対象動作は 43 動作とした．その結果，生成動作より目的動作に近かった動作は 5 動作であり，式 (5.2) に示す表出誤り率は約 0.8%であった．

$$\text{表出誤り率}\,(\%) = \frac{\text{目的動作に近いと感じた比較対象動作の数}}{\text{比較対象動作の数}} \times 100 \qquad (5.2)$$

また，システムについての感想をインタビュー形式で行なったところ,「学習済み感情語彙の数が少なく，指示入力の際に言い換えてもなかなか学習動作が出てこない」,「入力の際にもう少し親切であれば (入力文字の頭文字で学習済み動作データを探索する等) さらに使用しやすい」,「不満ボタンを押す際に，見た目には段々遠くの動作を探索しているように感じられないことがある」等のコメントが得られた.

5.1.6 結論

本研究では，言語を用いた指示により，感情動作を生成・制御可能なディジタルアクタシステムの構築を目的とし，感情語—動作間の関係を木構造として表現した感情表出モデルの構築を行った.

また，感情表出モデルから目的動作を生成・制御する感情表出部と，未知語が入力された場合等に新たに感情語を学習し感情表出モデルを更新する感情学習部から成るディジタルアクタシステムを実装した.

さらに，実装システムを用いて，適切な動作が表出可能かどうかの評価実験を行った結果，表出誤り率は約 0.8%であった.

以上の事から，提案モデル・システムの有効性が示されたと考える.

5.2 心理的尺度による感情動作制御

5.2.1 目的

本研究の目的は，感情動作を簡単かつ直感的なインタフェースで生成できるシステムを構築することである．本研究では，その1つとして，スライダの操作によって感情動作を制御するインタフェースを提案する．尚本研究では，最も基本的であり多様性を保持した動作である歩行動作を対象とする．

従来研究により,「大きく歩く」のような物理的な変動を伴った歩行動作に対して，内挿・外挿を用いることで動作を生成することが可能であることが明らかになっている[33],[39],[40]．また，感情に起因する変動を伴った歩行動作を内挿・外挿を用いることによって制御した前例があるが[41]，単純に物理的変化を均等に変化させたものであるため，感情の変化が均等に変化しているか否かは明らかではない．感情付き歩行動作を物理歩行動作と同様に内挿・外挿によるスライダによって制御させるためには，スライダの1目盛り毎に均等に感情を変化させる必要がある．そこで本研究では，心理的尺度という指標を用い，被験者実験により，感情の変化に対応する動作の変動量を明らかにする．

また，得られた心理的尺度により，図5.7に示すような,「普通に歩く」動作から「楽しく歩く」・「怒って歩く」・「悲しく歩く」動作を均等かつ連続的に変化させる3つのスライダを持つインタフェースを作成する．

5.2.2 心理的尺度

心理的尺度の定義

本研究では，歩行動作において表れやすい感情として,「楽しく歩く」・「怒って歩く」・「悲しく歩く」の3つを対象とする．心理的尺度とは，感情動作それぞれに存在し,「普通に歩く」動作から各感情動作の感情の強さを均等に変化させる心理的な評価尺度のことをいう．例えば,「普通に歩く」動作から「楽しく歩く」動作を均等に変化させる心理的尺度を「楽しいにおける心理的尺度」と呼ぶことにする．

心理的尺度の獲得方法

まず,「普通に歩く」動作と「楽しく歩く」・「怒って歩く」・「悲しく歩く」動作のそれぞれから，内挿・外挿によって歩行動作データを生成する．次に，得られた歩行動作を使用して心理的評価を行う被験者実験をし，その評価値を分析することによって心理的尺度を獲得する．

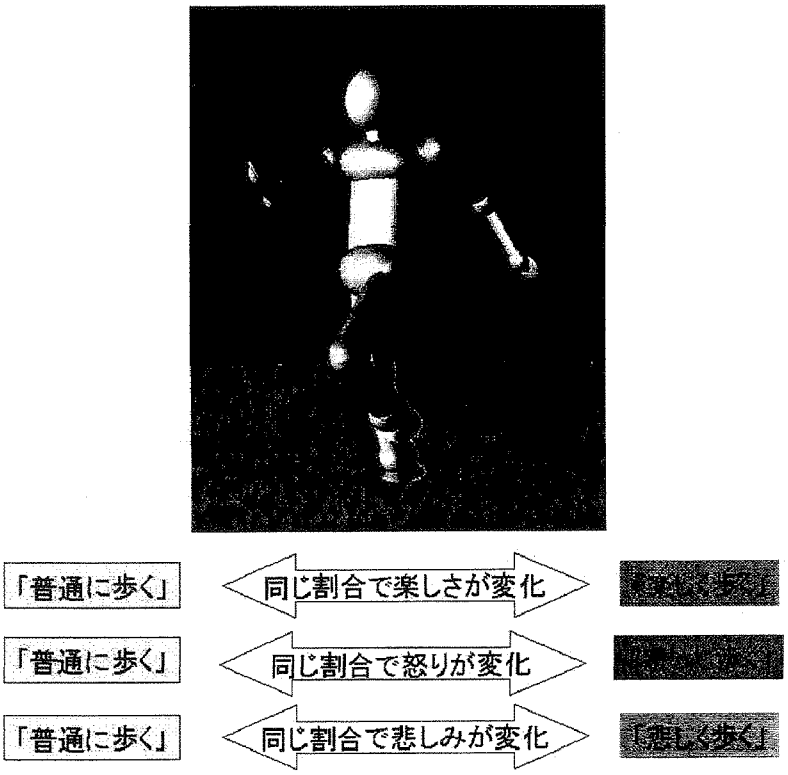


図 5.7: インタフェースのイメージ

5.2.3 実験

本項では，心理的尺度を獲得するための実験について説明する．

使用データ

まず，モーションキャプチャ装置によって 10 人分の「普通に歩く」動作,「楽しく歩く」動作,「怒って歩く」動作,「悲しく歩く」動作を収録し，BVH データ形式で保存する．

歩行動作における関節動作は全て周期的な運動をするため，連続時間周期信号として捉えることが可能である．よって，各関節動作信号 $\Theta_m(t)$ は式 (5.3) のようにフーリエ級数で近似でき，

$$\Theta_m(t) = \sum_{k=-\infty}^{\infty} C_{mk} e^{i\frac{2\pi k}{T}t} \tag{5.3}$$

各係数は式 (5.4) のフーリエ級数展開により求められる．

$$C_{mk} = \int_{-\infty}^{\infty} \Theta_m(t) e^{-i\frac{2\pi k}{T}t} dt \tag{5.4}$$

ここで，m は関節動作番号 $(0 \leq m \leq 56)$ を，k はスペクトルの高調波次数を表している．よって，$|C_{mk}|$ は関節動作番号 m の関節動作信号の第 k 高調波のスペクトルを表す．

本研究で使用するモーションキャプチャデータは離散時間信号であるため，スペクトルは式 (5.5) に示す離散フーリエ変換を用いて算出される．

$$C_{mk} = \sum_{n=0}^{N-1} \Theta_{mn} e^{-i\frac{2\pi nk}{N}} \tag{5.5}$$

また，このフーリエ係数から式 (5.6) に示す逆離散フーリエ変換を用いて，関節動作データが算出される．

$$\Theta_{mn} = \frac{1}{N} \sum_{k=0}^{N-1} C_{mk} e^{i\frac{2\pi nk}{N}} \tag{5.6}$$

過去の研究より，人物動作における関節動作は全て第 3 高調波までのスペクトル情報で十分動作を再現可能であることが示されている [41]．そこで本実験では，第 3 高調波までのスペクトル情報を使用する．

また，歩行データごとに被験者の個性などが含まれているため，それらを排除する必要がある．そこで，それぞれの歩行動作 10 人分のデータを平均し，そのそれぞれを「普通に歩く」動作,「楽しく歩く」動作,「怒って歩く」動作,「悲しく歩く」動作の代表とする．

「普通に歩く」動作のスペクトルデータの平均を $\bar{C}_{mk}$，感情付き歩行動作のスペクトルデータの平均を $\bar{C}'_{mk}$ とすると，その物理的な変化に対応した特徴量 ω_{mk} は式 (5.7) より求めることができる．

$$\omega_{mk} = \bar{C}'_{mk} - \bar{C}_{mk} \tag{5.7}$$

さらに，式 (5.8) によって，16 段階の内挿・外挿動作データ A_α を作成する．

$$A_\alpha = C_{mk} + \alpha\omega_{mk} \tag{5.8}$$

ここで，$\alpha = 0.2l(l = 0, 1, \cdots, 15)$ である．

実験方法

実験の様子を図 5.8 に示す．

まず，作成した「普通に歩く」動作から「楽しく歩く」動作を内挿・外挿した 16 段階の歩行動作を被験者にランダムに提示し,「普通に歩く」を 0 点，被験者がイメージする最も「楽しく歩く」動作を 100 点として，それぞれの動作に点数をつけてもらう．ただし，被験者がイメージする最も「楽しく歩く」動作よりもさらに楽しいと感じる歩きがあれば，100 点以上の点数をつけても構わないものとする．

次に，被験者がつけた最高点を100点として0〜100点になるように正規化し，それを被験者の数で平均を取り，さらに0〜100点に正規化する．これを「楽しいにおける心理的尺度」とする．

「怒って歩く」動作,「悲しく歩く」動作に対しても同様に行う．

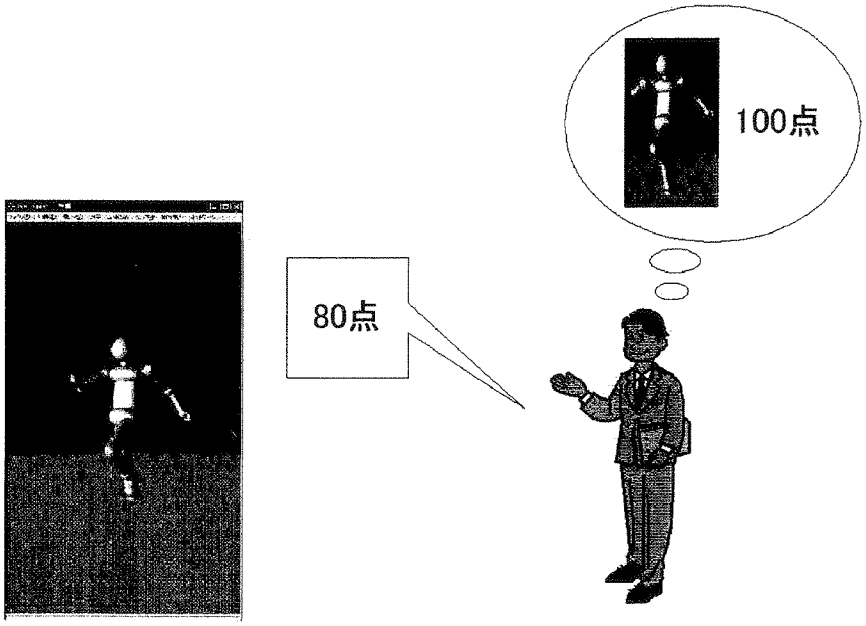


図 5.8: 実験風景イメージ

実験結果

1. 「楽しく歩く」動作の結果
   図5.9に「楽しく歩く」動作の結果を示す．図より，$0.0 \leq \alpha \leq 2.4$で全体的に増加し，それ以降は減少に転じていることがわかる．また，$0.0 \leq \alpha \leq 0.6$ではほとんど変化が見られず，$0.6 \leq \alpha \leq 1.4$の区間では振動していることが確認できる．

2. 「怒って歩く」動作の結果
   図5.10に「怒って歩く」動作の結果を示す．図より，$0.0 \leq \alpha \leq 2.6$で全体的に増加し，それ以降は減少に転じることがわかる．また，$0.0 \leq \alpha \leq 0.4$，$0.8 \leq \alpha \leq 1.2$ではほとんど変化が見られず，$1.2 \leq \alpha \leq 1.8$の区間でははは振動していることが確認できる．

3. 「悲しく歩く」動作の結果
   図5.11に「悲しく歩く」動作の結果を示す．図より，$0.0 \leq \alpha \leq 2.2$で全体的に増加し，それ以降は減少に転じることがわかる．また，$0.0 \leq \alpha \leq 0.2$ではほとん

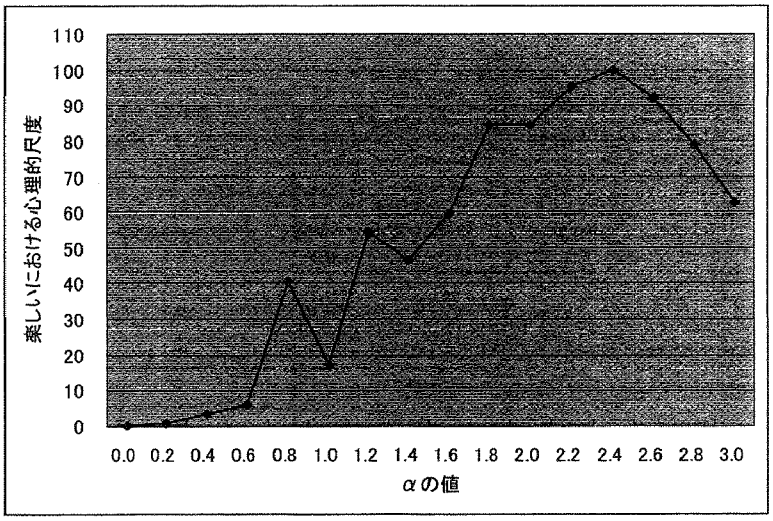


図 5.9: 「楽しく歩く」動作の結果

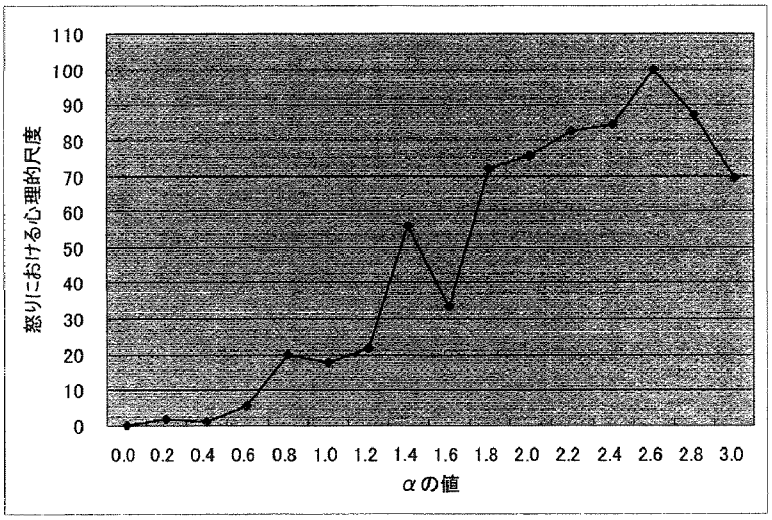


図 5.10: 「怒って歩く」動作の結果

ど変化が見られず，$0.8 \leq \alpha \leq 1.4$ の区間でははは振動していることが確認できる．

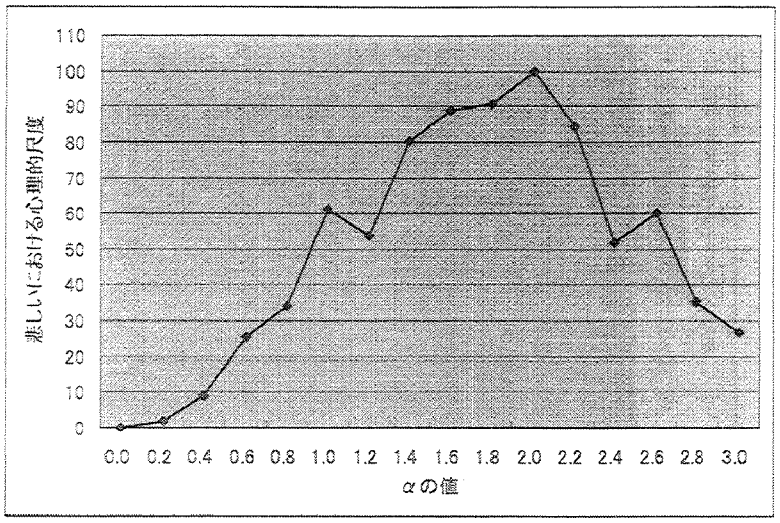


図 5.11: 「悲しく歩く」動作の結果

考察

「楽しく歩く」動作,「怒って歩く」動作,「悲しく歩く」動作の全てにおいて，α の値が増加するにしたがって，心理的尺度が初めは増加し，あるところで減少に転じるという傾向が見られた．心理的尺度が増加する部分だけを用い，さらにその部分で振動している部分は平均を取ることによって，1 目盛り毎に感情の強さを変化させるスライダが作成可能であると考えられる．

11 段階に均等に感情の強さを制御可能な心理的尺度と，それに対応する内挿・外挿による重み α の値の関係を表 5.4-5.6 に示す．実際の動作と違い，感情付き歩行動作のアニメーションでは α の値が 1.0 以上となる外挿，つまり誇張した動作が重要だと言える．

5.2.4 実装システム

実験によって獲得した「楽しいにおける心理的尺度」・「怒りにおける心理的尺度」・「悲しいにおける心理的尺度」のそれぞれをスライダにより 0 点から 100 点まで 10 点毎に変化させることで,「普通に歩く」動作から「楽しく歩く」動作・「怒って歩く」動作・「悲しく歩く」動作を均等に変化させるシステムを実装した．実装システムの外観を図 5.12 に示す．

表 5.4: 「楽しいにおける心理的尺度」と α の値

楽しいにおける心理的尺度	α の値
0	0
10	0.6239
20	0.8911
30	0.8996
40	0.9080
50	1.3131
60	1.6013
70	1.6820
80	1.7627
90	2.1007
100	2.4

表 5.5: 「怒りにおける心理的尺度」と α の値

怒りにおける心理的尺度	α の値
0	0
10	0.6624
20	1.1196
30	1.2485
40	1.4943
50	1.2718
60	1.7376
70	1.7897
80	2.1274
90	2.4710
100	2.6

表 5.6: 「悲しいにおける心理的尺度」と α の値

悲しいにおける心理的尺度	α の値
0	0
10	0.4132
20	0.5343
30	0.7043
40	0.8431
50	0.9174
60	1.0897
70	1.3228
80	1.3988
90	1.7232
100	2.0

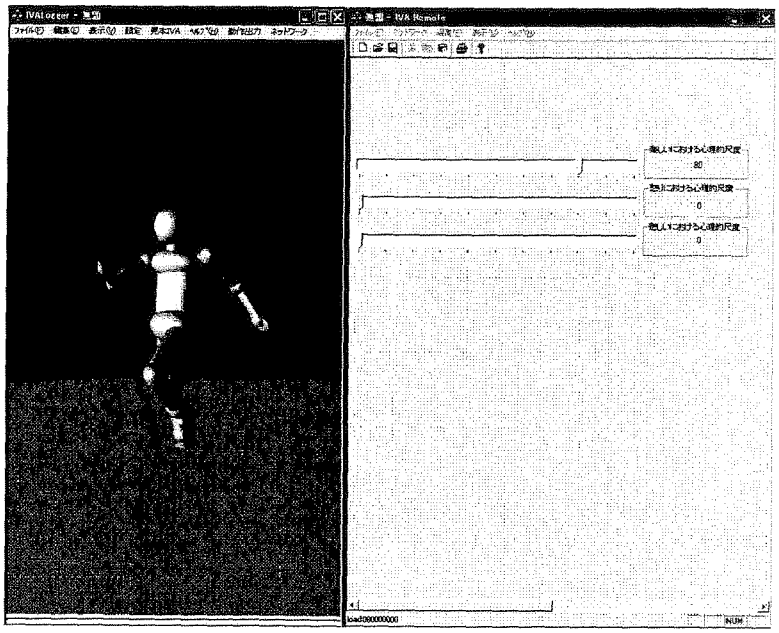


図 5.12: 完成したシステム

5.2.5 結論

本研究では,「普通に歩く」動作から「楽しく歩く」動作・「怒って歩く」動作・「悲しく歩く」動作を均等かつ連続的に変化させるスライダを持つインタフェースの作成を試みた．感情の強さを判断する指標となる心理的尺度を定義し，実験によって「楽しいにおける心理的尺度」・「怒りにおける心理的尺度」・「悲しいにおける心理的尺度」を獲得した．それによって，物理歩行動作と同様に単純に内挿・外挿で歩行動作を変化させるのではなく，それぞれの心理的尺度を段階的に変化させるスライダを作成し，ユーザが容易にかつ直感的に操作できるシステムを構築した．

第6章　感情表現における映像と音声の役割

本研究の目的は，コンピュータアニメーションにおいて，CG キャラクタの音声や動作を適切に制御することによって感情表現を可能にするマルチモーダル感情表出モデルを構築することである．本章では，複数のモダリティを使用して感情表現を行った場合，人間がそれらをどのように知覚するかを明らかにする．具体的には，音声や映像という物理的信号を組み合わせ，音色や顔表情を制御した場合，人間が知覚する感情にどういう影響を与えるかを実験により明らかにする．

6.1　目的

人間は外界の状態に対する評価を行うために，いくつかのチャネルを有している．つまり，人間の外界とのコミュニケーションは，言語というチャネルを通じた言語的コミュニケーションと，言語以外のチャネルを介した非言語的コミュニケーションから成り立っている．すべてのチャネルから入ってくるものに対してさまざまな色付けを行うのが「感情」という評価機構である．そのため,「感情」は人間同士のコミュニケーションに大きな影響力を持つと考えられる．また人間同士に限らず，アニメーション内の CG キャラクタにとっても，必要な評価機構であると考えられる．

CG アニメーションによる映像が増えている今，専門家だけではなく，一般の人にもより容易に CG キャラクタに感情を表出させる技術が求められ始めている．

人間が日常的に行っている感情の表出を CG キャラクタに行わせる為には，まず人間同士が日頃どのような基準で相手の表出した感情を判断しているのかを知らなければならない.「画像」や「音声」といった物理信号と，人間がそこから知覚する「感情」との対応関係を明らかにすることで，CG キャラクタによる容易な感情表出も可能になると考えられる．

従来より，感情の判断基準であると考えられる表情，音声個々の研究は行われているが [42]-[44]，本章では「顔表情 (画像)」と「声色 (音声)」の関係に着目し，感情判断におけるこれら 2 つの，貢献度の違いや相互作用を検討する．

本章では，以下 2 点を分析の観点とし，研究を進めることにした．

観点 1: 聞き手は発話者が意図したとおりに感情のレベルを認識できるか.

観点 2: ひとつのモダリティー（画像または音声）を固定した場合，聞き手はもう一方の

モダリティー（音声または画像）による感情のレベルを発話者が意図したとおりに認識できるか.

これらを検討するにあたり，まず感情表現収録実験を行い，次に主観評価実験を行った.

1. 感情表現収録実験

   発話者に感情を込めて文章を読んでもらい，顔画像と音声を同時収録した．採用した感情は，平静，喜び，怒り，悲しみの4種類で，喜び・怒り・悲しみの3感情については，各3段階ずつ（小・中・大）のレベル付で表現してもらった.

2. 主観評価実験

   収録した画像と音声を，Adobe Premiere 6.5を用いて組替え合成を行った．収録した，平静と，喜び・怒り・悲しみの3感情についてはレベル（小・中・大）付のデータを組替え合成に用いた．平静と喜び（小・中・大），平静と怒り（小・中・大），平静と悲しみ（小・中・大）という，平静を含む同一感情内で画像と音声の全組み合わせを合成した.

   組換え合成を行ったデータを評価者に見せ，それぞれどのような感情で，どの程度の感情の度合いが含まれていると考えるか主観評価してもらった.

6.2 感情表現収録実験

感情評価における,「表情 (画像)」と「声色 (音声)」の，貢献度の違いや相互作用を調べるにあたり，感情表現収録実験を行った．実験において以下のようなデータを採取した.

デジタルビデオカメラにより，顔画像と音声を同時収録した．採用した感情は，平静，喜び，怒り，悲しみの4種類で，喜び・怒り・悲しみの3感情については，各3段階ずつ（小・中・大）のレベル付で表現してもらった．発話内容は，条件として以下を含むものとし，AからCまでの3文とした.

- 条件
  - 文意が感情判断に影響を与えないように発話文文意は中立
  - 発話者が感情表現を行いやすいように，全無核の文は採用しない
  - 疑問文，平叙文の両方を含む
  - 頻繁に会話に登場する発話文を含む
- 発話内容

  A: 「バイトの間にメールが届いた.」
  （UAP+UAP+UAP+AAP:4文節）

B: 「犬が仰向けになるのを見ましたか？」
(AAP+UAP+AAP+AAP:4 文節)

C: 「ちょっと待ってください.」
(AAP+AAP+AAP:3 文節)

AAP: Accented Accentual Phrase（有核句）
UAP: Unaccented Accentual Phrase（無核句）

収録は，演劇経験のある桐朋学園大学芸術短期大学の生徒，女性 1 名男性 3 名に依頼した.

6.3 主観評価実験

感情評価における,「表情 (画像)」と「声色 (音声)」の，貢献度の違いや相互作用を調べるにあたり，収録したデータの画像と音声の組替え合成をしたデータを評価者に見てもらう主観評価実験を行った.

6.3.1 画像と音声の組替え合成方法

収録した画像と音声を，Adobe Premiere 6.5（映像編集ソフト：図 6.1）を用いて組替え合成を行った.

収録した，平静と，喜び・怒り・悲しみの 3 感情についてはレベル（小・中・大）付のデータを組替え合成に用いた．平静と喜び（小・中・大），平静と怒り（小・中・大），平静と悲しみ（小・中・大）という，平静を含む同一感情内で画像と音声の全組み合わせを合成した．作成したデータの組み合わせパターンを表 6.1 に示す．作成したデータは，1 発話者 1 文あたり 45 セット，文 A から文 C の合計は 135 セットである．時間長は画像のものを音声に揃えた.

表 6.1: 作成データの組み合わせパターン

	画像											
音声		なし(音声のみ)	平静	怒り・大	怒り・中	怒り・小	喜び・大	喜び・中	喜び・小	悲しみ・大	悲しみ・中	悲しみ・小
	なし(画像のみ)		1画像	23画像	22画像	21画像	8画像	7画像	6画像	38画像	37画像	36画像
	平静	1音声	1	23-1	22-1	21-1	8-1	7-1	6-1	38-1	37-1	36-1
	怒り・大	23音声	1-23		22-23	21-23						
	怒り・中	22音声	1-22	23-22		21-22						
	怒り・小	21音声	1-21	23-21	22-21							
	喜び・大	8音声	1-8					7-8	6-8			
	喜び・中	7音声	1-7				8-7		6-7			
	喜び・小	6音声	1-6				8-6	7-6				
	悲しみ・大	38音声	1-38								37-38	36-38
	悲しみ・中	37音声	1-37							38-37		36-37
	悲しみ・小	36音声	1-36							38-36	37-36	

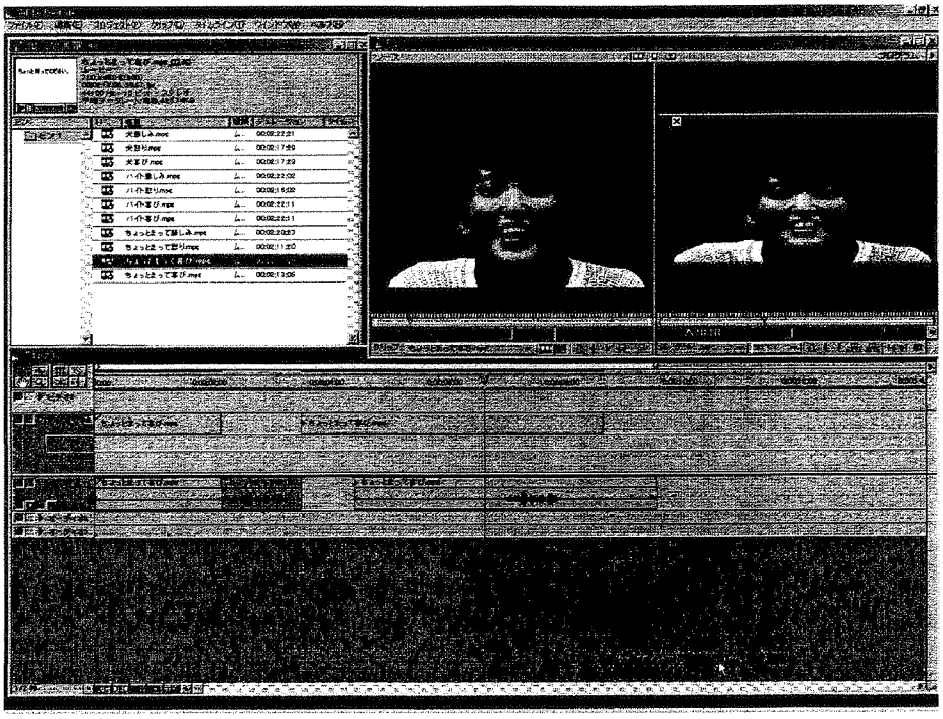


図 6.1: Adobe Premiere 6.5 使用例

6.3.2 実験内容

観点1のための実験には，発話者が演じたそのままの元データを用いた．データはランダムに呈示し，各データは繰り返し視聴してよいこととした．

観点2のための実験には，平静と喜び（小・中・大），平静と怒り（小・中・大），平静と悲しみ（小・中・大）という，平静を含む同一感情内で画像と音声の全組み合わせを合成したデータを用いた．データはランダムに呈示した．観点2では，データ同士の比較ではなく，各データそのものの評価を得るために，データの繰り返しは行わないこととした．その代わりに各発話者の元データを初めに呈示し，評価する際の指標としてもらった．

観点1,2ともに，各データに対して図6.2のような評定スケールを用いて，評価者に評価してもらった．評定スケール作成にあたり，[45] を参考にした．

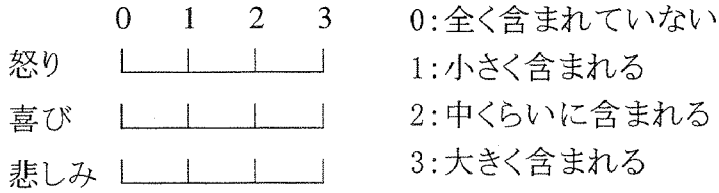


図 6.2: 評定スケール

6.3.3 観点1 評価結果

発話者4名分のデータを3名の評価者に主観評価してもらった．発話者の意図した感情のレベルに対する認識の一致率を表6.2に示す．

表 6.2: 感情レベルの一致率

発話者	一致率 [%]
A	98.7
B	96.3
C	98.8
D	92.6

いずれの発話者に対しても高い認識の一致率を得ることが出来た．これにより今回用いた4名の発話者のデータのうち，一致率の特に高い発話者A，B，Cの3名分のデータを，観点2に活用することにした．

6.3.4 観点2 評価結果

発話者A，B，Cの3名分の組替え合成済みデータを10名の評価者（a〜j）に主観評価してもらった．

全体

評価結果を感情ごとに，音声レベルを固定し画像レベルを変えた結果と，画像レベルを固定し音声レベルを変えた結果を以下に示す．評価者がつけた評価値の平均を図6.3〜図6.5に，評価者がつけた評価値と発話者の意図したレベル（1，2，3）との相関係数を表6.3に示す．ここでは，評価者，発話者，発話内容の区別は行なっていない．

1. 評価値平均による結果

   図中の数値は，評価者がつけた評価値の平均である．

   図6.5の，画像レベルを[小]に固定し，それぞれ音声レベル[中][大]を合成した[小・大][小・中]（[画像レベル・音声レベル]），画像レベルを[平静]に固定し，それぞれ音声レベル[中][大]を合成した[平静・大][平静・中]に対する評価は音声レベルの変化に対応していない．しかし，画像レベルを[中][大]に固定した場合や，図6.3，図6.4においては，合成した音声レベルにほぼ比例した評価を得ている．

   画像レベルを固定した場合と比較し，音声レベルを固定したデータへの評価は，合成した画像レベルに比例していない．図6.4は特にその傾向が顕著である．図

6.4において，音声レベルを[大]に固定した[中・大][大・大]に対する評価は画像レベルの変化に対応していない．

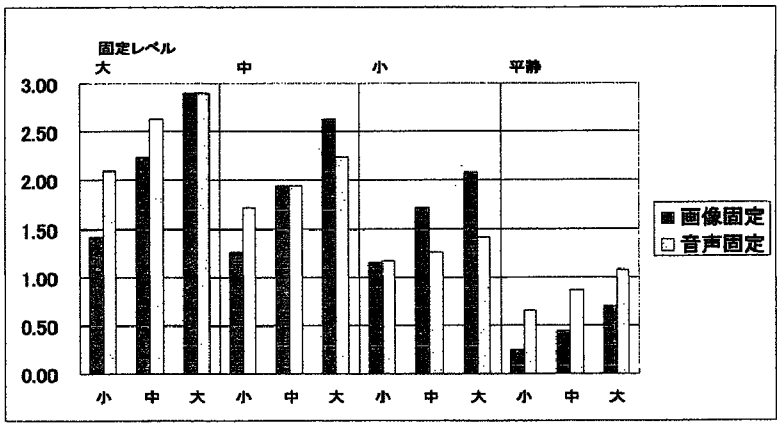


図 6.3: 「喜び」に対する評価値平均による結果

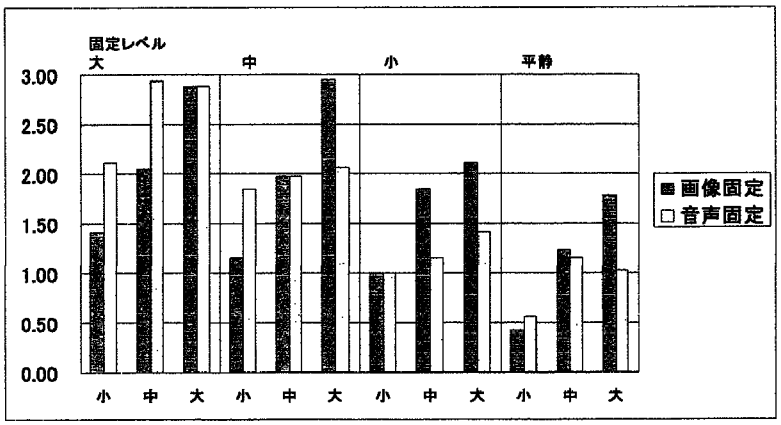


図 6.4: 「怒り」に対する評価値平均による結果

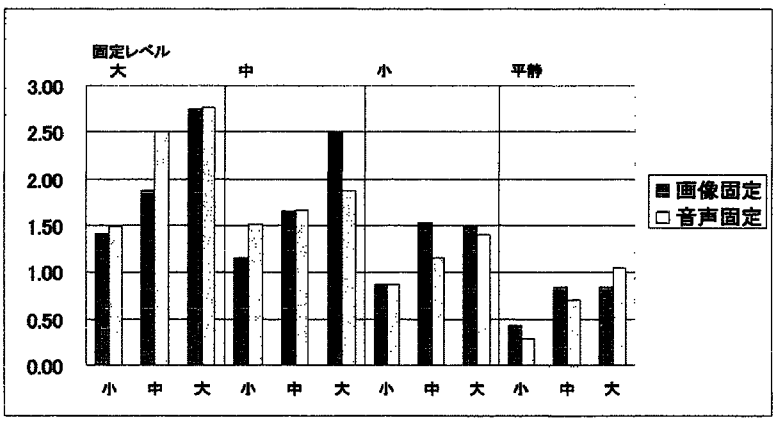


図 6.5: 「悲しみ」に対する評価値平均による結果

2. 相関係数による結果

表 6.3 の左側には，音声レベルを固定し，合成した画像レベル（1，2，3）と評価者がつけた評価値との相関係数が示されており，右側には，画像レベルを固定し，合成した音声レベル（1，2，3）と評価者がつけた評価値との相関係数が示されている．

表 6.3 より，以下のことが確認された．

- 表右側の音声レベルとの相関のほうが，表左側の画像レベルとの相関に比べ，大きくなっている．
- 画像固定の「怒り」では固定した画像レベルの違いに関係なく，音声レベルと相関がある．
- 画像固定の「喜び」「悲しみ」では，固定した画像レベルが大きいほど，合成した音声レベルとの相関が大きくなっている．固定した画像のレベルが小さいと，音声レベルとの相関はない．
- 音声固定において，音声レベルを「3」に固定したもののみ，合成した画像との相関がある．

評価者別

評価者による傾向を見るために，評価者別の結果を表 6.4～表 6.13 に示す．ここでは，発話内容，発話者の区別は行なっていない．

表左側には，音声レベルを固定し，合成した画像レベル（1，2，3）と評価者がつけた評価値との相関係数が示されている．表右側には，画像レベルを固定し，合成した音声レベル（1，2，3）と評価者がつけた評価値との相関係数が示されている．

表 6.3: 全体評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.29	0	0.31
	1	0.23	1	0.57
	2	0.39	2	0.76
	3	0.53	3	0.79
怒り	0	0.33	0	0.65
	1	0.29	1	0.63
	2	0.23	2	0.86
	3	0.52	3	0.80
悲しみ	0	0.48	0	0.26
	1	0.37	1	0.36
	2	0.26	2	0.67
	3	0.62	3	0.75

上述の評価者，発話者，発話内容の区別なしの全体の結果と同じ傾向が見られる．例外として，表6.5, 表6.8, 表6.13の「悲しみ」では，表左側の画像レベルとの相関ほうが，表右側の音声レベルとの相関に比べ，大きくなっている．

表 6.4: 評価者 a による評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.00	0	0.08
	1	0.29	1	0.48
	2	0.43	2	0.63
	3	0.60	3	0.69
怒り	0	0.00	0	0.78
	1	0.35	1	0.63
	2	0.00	2	0.67
	3	0.48	3	0.73
悲しみ	0	0.35	0	0.35
	1	0.35	1	0.23
	2	0.29	2	0.71
	3	0.78	3	0.81

発話者別

発話者による傾向を見るために，発話者別の結果を表6.14〜表6.16に示す．ここでは，発話内容，評価者の区別は行なっていない．

表左側には，音声レベルを固定し，合成した画像レベル（1，2，3）と評価者がつけた評価値との相関係数が示されている．表右側には，画像レベルを固定し，合成した音声レベル（1，2，3）と評価者がつけた評価値との相関係数が示されている．

評価者，発話者，発話内容の区別なしの全体の結果と同じ傾向が見られる．

表 6.5: 評価者 b による評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.31	0	0.24
	1	0.00	1	0.00
	2	0.53	2	0.49
	3	0.78	3	0.85
怒り	0	0.57	0	0.54
	1	-0.27	1	0.33
	2	0.00	2	0.89
	3	0.57	3	0.86
悲しみ	0	0.65	0	0.22
	1	0.72	1	0.23
	2	0.43	2	0.42
	3	0.56	3	0.47

表 6.6: 評価者 c による評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.55	0	0.29
	1	0.00	1	0.73
	2	0.11	2	0.84
	3	0.48	3	0.90
怒り	0	0.27	0	0.80
	1	0.26	1	0.79
	2	0.00	2	0.95
	3	0.44	3	0.91
悲しみ	0	0.40	0	0.08
	1	-0.29	1	0.36
	2	0.16	2	0.72
	3	0.62	3	0.95

表 6.7: 評価者 d による評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.35	0	0.29
	1	0.00	1	0.68
	2	0.24	2	0.89
	3	0.60	3	0.97
怒り	0	0.55	0	0.69
	1	0.00	1	0.65
	2	0.35	2	0.94
	3	0.73	3	0.95
悲しみ	0	0.75	0	0.73
	1	0.29	1	0.68
	2	0.00	2	0.83
	3	0.80	3	0.88

表 6.8: 評価者 e による評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.56	0	0.60
	1	0.81	1	0.68
	2	0.54	2	0.83
	3	0.68	3	0.74
怒り	0	0.71	0	0.84
	1	0.62	1	0.51
	2	0.55	2	0.92
	3	0.74	3	0.82
悲しみ	0	0.83	0	0.44
	1	0.78	1	0.56
	2	0.61	2	0.68
	3	0.57	3	0.62

表 6.9: 評価者 f による評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.12	0	0.48
	1	0.29	1	0.77
	2	0.29	2	0.89
	3	0.51	3	0.95
怒り	0	0.17	0	0.47
	1	0.58	1	0.75
	2	0.35	2	0.97
	3	0.60	3	0.88
悲しみ	0	0.61	0	0.47
	1	0.00	1	0.49
	2	0.55	2	0.92
	3	0.79	3	1.00

表 6.10: 評価者 g による評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.35	0	0.00
	1	-0.09	1	0.69
	2	0.52	2	0.85
	3	0.35	3	0.84
怒り	0	0.35	0	0.66
	1	0.55	1	0.89
	2	0.24	2	0.92
	3	0.43	3	0.84
悲しみ	0	0.44	0	0.30
	1	0.65	1	0.47
	2	-0.33	2	0.88
	3	0.73	3	0.77

表 6.11: 評価者 h による評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.30	0	0.40
	1	0.35	1	0.65
	2	0.27	2	0.75
	3	0.51	3	0.77
怒り	0	0.27	0	0.76
	1	0.46	1	0.65
	2	0.38	2	0.82
	3	0.64	3	0.77
悲しみ	0	0.56	0	0.22
	1	0.34	1	0.28
	2	0.25	2	0.57
	3	0.80	3	0.79

表 6.12: 評価者 i による評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.15	0	0.50
	1	0.30	1	0.54
	2	0.50	2	0.80
	3	0.58	3	0.84
怒り	0	0.33	0	0.62
	1	0.24	1	0.78
	2	0.19	2	0.80
	3	0.09	3	0.69
悲しみ	0	0.08	0	0.35
	1	0.19	1	0.26
	2	0.29	2	0.67
	3	0.47	3	0.64

表 6.13: 評価者 j による評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.38	0	0.50
	1	0.28	1	0.55
	2	0.75	2	0.79
	3	0.32	3	0.61
怒り	0	0.36	0	0.74
	1	0.33	1	0.60
	2	0.47	2	0.87
	3	0.58	3	0.82
悲しみ	0	0.56	0	0.08
	1	0.50	1	0.35
	2	0.54	2	0.62
	3	0.58	3	0.66

表 6.14: 発話者 A に対する評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.33	0	0.44
	1	0.14	1	0.70
	2	0.36	2	0.82
	3	0.33	3	0.80
怒り	0	0.20	0	0.73
	1	0.18	1	0.75
	2	0.10	2	0.81
	3	0.45	3	0.78
悲しみ	0	0.37	0	0.29
	1	0.28	1	0.33
	2	0.31	2	0.58
	3	0.65	3	0.74

表 6.15: 発話者 B に対する評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.11	0	0.30
	1	0.31	1	0.57
	2	0.36	2	0.68
	3	0.56	3	0.74
怒り	0	0.44	0	0.69
	1	0.41	1	0.65
	2	0.29	2	0.88
	3	0.50	3	0.82
悲しみ	0	0.42	0	0.34
	1	0.33	1	0.55
	2	0.19	2	0.82
	3	0.46	3	0.76

表 6.16: 発話者 C に対する評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.42	0	0.19
	1	0.25	1	0.49
	2	0.46	2	0.78
	3	0.71	3	0.85
怒り	0	0.41	0	0.57
	1	0.29	1	0.53
	2	0.31	2	0.89
	3	0.63	3	0.82
悲しみ	0	0.63	0	0.17
	1	0.52	1	0.24
	2	0.27	2	0.68
	3	0.75	3	0.75

発話内容別

発話内容による傾向を見るために，発話内容別の結果を表6.17〜表6.19に示す．ここでは，発話者，評価者の区別は行なっていない．

表左側には，音声レベルを固定し，合成した画像レベル（1，2，3）と評価者がつけた評価値との相関係数が示されている．表右側には，画像レベルを固定し，合成した音声レベル（1，2，3）と評価者がつけた評価値との相関係数が示されている．

評価者，発話者，発話内容の区別なしの全体の結果と同じ傾向が見られる．また，表6.17の音声固定では，他の表6.18, 表6.19に比べて合成した画像レベルとの相関係数が高くなっている．

表 6.17: 「バイトの間にメールが届いた.」に対する評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.41	0	0.44
	1	0.25	1	0.59
	2	0.42	2	0.76
	3	0.61	3	0.81
怒り	0	0.40	0	0.70
	1	0.34	1	0.74
	2	0.16	2	0.89
	3	0.43	3	0.73
悲しみ	0	0.51	0	0.29
	1	0.54	1	0.46
	2	0.36	2	0.61
	3	0.58	3	0.70

表 6.18: 「犬が仰向けになるのを見ましたか？」に対する評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.35	0	0.28
	1	0.22	1	0.53
	2	0.24	2	0.73
	3	0.52	3	0.82
怒り	0	0.39	0	0.68
	1	0.24	1	0.55
	2	0.22	2	0.88
	3	0.57	3	0.83
悲しみ	0	0.44	0	0.22
	1	0.37	1	0.38
	2	0.33	2	0.68
	3	0.62	3	0.78

表 6.19: 「ちょっと待ってください.」に対する評価結果

	音声固定		画像固定	
	固定感情レベル	相関係数	固定感情レベル	相関係数
喜び	0	0.11	0	0.24
	1	0.20	1	0.62
	2	0.52	2	0.79
	3	0.45	3	0.75
怒り	0	0.20	0	0.56
	1	0.30	1	0.64
	2	0.32	2	0.84
	3	0.55	3	0.85
悲しみ	0	0.50	0	0.29
	1	0.18	1	0.26
	2	0.06	2	0.72
	3	0.67	3	0.77

6.4 結論

音声レベルを固定した場合と画像レベルを固定した場合では，6.3.4 項の様な相違が生じた．このような結果がでた要因として，以下の点が考えられる．

- 発話者が表情による感情の度合いの変化を音声に比べてうまく表現できていない．
- 一般的に感情の度合いの変化は表情ではなく声色の変化で表現される．
- 聞き手は，発話者の表情を見ることでどのような感情であるかは認識できるが，その感情の度合いに関しては，声色を重視して認識している．
- 音声，画像いずれかに含まれる感情の度合いが大きい場合には，もう一方に対する判断が容易に行える．

これらのことより，聞き手は感情を判断する基準として，表情より音声による感情の表現をより重視していることがわかる．

また，さらに細かく見た場合，以下のことが確認できる．

- 声色に「怒り」が少しでも含まれていれば，表情が怒っていなくても，「相手は怒っている」と判断してしまう．
- 「悲しみ」では，声色よりも表情を重視する人も若干いる．

さらに，発話内容別の結果において，無核句を多く含む発話内容の場合は，有核句を多く含む発話内容に比べて，画像との相関が大きくなっていたことから，以下のことが考えられる．

- 無核句を多く含む文章は，発話者が音声による感情の度合いをうまく表現できていない．
- 聞き手は，有核部分に表れる音響的な特徴を頼りに感情の度合いを認識している．

以上より，CG キャラクタの画像と音声を作成する際には，上記のような例外を除いて，音声作成により重点をおいて行うべきであると考えられる．

第7章　あとがき

以上の各章で述べてきたように，本研究の目的は，映像コンテンツ中のCGキャラクタの動作と発話時の韻律を適切に制御することで，動作と音声に感情を付与することのできるマルチモーダル感情表出モデルを構築することである．本研究では，これを達成する上で重要な諸問題に取り組んだ．

各章で述べた成果は，いずれも今後，さらに環境を整備することにより十分実用化の可能性の高いものである．

以下，本研究で得られた主要な知見をまとめる．

- 噺家3名について基本周波数・パワー・発話速度をパラメータとし，発話全体・文頭のアクセント句・文末のアクセント句それぞれに対して，喜び・驚き・怒り・平静・恐れ・あきれの6つの感情の判別を行った．その結果より以下の知見が得られた．
  - データ全体を2群に分別する際には発話単位での分析の方が有効であるが，各感情へ細かく分ける際には文末のアクセント句に注目する方が比較的良い結果を得られている．つまり，判別結果を総合的に見ると感情表現は発話の最後の部分に比較的現れやすいのではないかと推測できる．
  - 2群へのグルーピングをする場合はしない場合に比べると各感情データ内の判別率は多少の向上が見られる．
  - 各感情相互の誤判別の起こりやすさは，平静・恐れ，喜び・驚き，驚き・怒り間に多いようである．つまり，これらは特徴が類似しているものと思われる．
- 「笑い声」の音声について，規則合成方式による音声合成に必要な音響特徴を分析した．その結果「笑い声」に関して以下のことがわかった．
  - 笑い声の発声において，音声パワーの時間変化は下り調子の変化をするか，またはほとんど変化をしないかであり，多くは全体的な単調減少の時間変化をすることがわかった．
  - 「笑い声」と「通常発話」の/ha/の音についてピッチ周波数を比較したところ，「笑い声」は「通常発話」よりも全体的に50〜100Hzほどピッチ周波数が高く，話者によっては裏声がよく現れ，400Hz付近の音声になることがあるとわかった．また，主観評価実験の結果から，ピッチ周波数の低い

ものは「笑い声」として認識され難く，ピッチ周波数が高いことが笑い声らしさのための重大な要因であることがわかった．このことから，ピッチ周波数は「笑い声」の笑い声らしさに対する最も基礎的なパラメータであるといえる．

- 「笑い声」は声道からの過剰な呼気の流出により無声化が強まるが，主観評価実験の結果によれば，そのために生じる無声音を聴取者が感じることは無いといえる．無声化の影響は，主に音色に影響していると考えられる．
- 「笑い声」と「通常発話」の/ha/の音について，有声部である後続母音の音素/a/のスペクトル包絡の概形を比較したところ,「笑い声」の場合は無声音や雑音のように高調波構造が曖昧になっていることがわかった．
- 定量的な分析により,「笑い声」の音声と「通常発話」の音声では，音素/a/の音素持続時間に大きな差異があることがわかった．主観評価実験の結果により，この差が「笑い声」と「通常発話」の聞こえ方の違いに対する重大な要因になっていることがわかった．音素タイミングは「笑い声」の笑い声らしさに対する基礎的なパラメータの一つであるといえる．

- 入力された文章から感情を認識することを目的として，大局的な感情理解と局所的な感情理解の2つの方法について検討を行った．大局的な感情理解では，Text tiling algorithm を用いることにより場面分割を行い，各場面における感情表現語の頻度が大きいものが，その場面における大局的感情であると考え，検証実験を行った．その結果，高い精度で上述のルールが正しいことが確認された．また，局所的な感情分類を行うために，被験者実験により文型毎の感情表現要素辞書の作成を行った．

- 少ない指示で多様な歩行動作を精密に生成・制御可能な歩行動作制御モデルの構築を行った．まず，歩行動作において，全関節動作は連動性を基にいくつかの独立なグループに分類される事が 重回帰分析によって確認された．次に，12の動作制御項目に対応した19の特徴量成分を抽出し，これらの強さを制御する事で多様な歩行動作を生成・制御可能なモデルを構築した．モデル評価実験及びモデルを実装したシステム評価実験により，定量的な側面からは，物理的な特徴を保持する歩行動作は十分な動作再現する事が可能であり，感情に関する特徴を保持する歩行動作については十分な動作再現を図る事が困難である事が判明した．一方，定性的な側面からは，感情に関する特徴を保持する歩行動作は十分な動作再現する事が可能である事が判明した．

- 言語を用いた指示により，感情動作を生成・制御可能なディジタルアクタシステムの構築を目的とし，感情語—動作間の関係を木構造として表現した感情表出モデルの構築を行った．また，感情表出モデルから目的動作を生成・制御する感情表出部と，未知語が入力された場合等に新たに感情語を学習し感情表出モデルを更新する感情学習部から成るディジタルアクタシステムを実装した．さらに，実

装システムを用いて評価実験を行った結果，表出誤り率は約0.8%であり，適切な動作が表出可能であることが確認された．

- 「普通に歩く」動作から「楽しく歩く」動作・「怒って歩く」動作・「悲しく歩く」動作を均等かつ連続的に変化させるスライダを持つインタフェースの作成を目的として，感情の強さを判断する指標となる心理的尺度を導入し，実験によって「楽しいにおける心理的尺度」・「怒りにおける心理的尺度」・「悲しいにおける心理的尺度」を獲得した．次に，それぞれの心理的尺度を段階的に変化させるスライダにより，ユーザが容易にかつ直感的に操作できるシステムを実装した．

- 映像や音声といった物理信号と，人間がそこから知覚する感情との対応関係を明らかにするために，映像と音声を組み合わせ，それぞれに感情表現を付与した場合，人間がそれをどのように知覚するかについて分析を行った．その結果以下のような知見が得られた．

  - 発話者が表情による感情の度合いの変化を音声に比べてうまく表現できていない．
  - 一般的に感情の度合いの変化は表情ではなく声色の変化で表現される．
  - 聞き手は，発話者の表情を見ることでどのような感情であるかは認識できるが，その感情の度合いに関しては，声色を重視して認識している．
  - 音声，画像いずれかに含まれる感情の度合いが大きい場合には，もう一方に対する判断が容易に行える．

  以上より，聞き手は感情を判断する基準として，表情より音声による感情の表現をより重視していることが確認された．

これらの成果は，いずれも今後の発展が期待できるものであり，継続的に研究・評価を続けていく予定である．

参考文献

[1] 川波弘道，広瀬啓吉，“態度・感情音声における韻律的特徴量の考察”，信学技報，SP97-67，pp.73-76，1997

[2] 藤井洋之，重永実，“感情音声の機械による感情の判別の試み”，音講論，2-P-21，pp.343-344，1996

[3] 上床弘幸，小林豊，新美康永，“音声の感情表現の分析とモデル化”，信学技報，SP92-131，pp.65-72，1993

[4] 平賀裕，斎藤善行，森島繁生，原島博，“音声に含まれる感情抽出の一検討”，信学技報，HC93-66，pp.1-8，1994

[5] 広瀬啓吉，“音声情報処理におけるパラ・非言語情報”，音講論，1-10-10，pp.243-246，2002

[6] 武田昌一，海老義人，鈴木修平，舟木克年，“落語音声の韻律的特徴の解析”，音講論，1-P-4，pp.289-290，1996

[7] 門谷信愛希，阿曽弘具，鈴木基之，牧野正三，“音声に含まれる感情の判別に関する検討”，情処論，SLP34-8，pp.43-48，2000

[8] 菊池英明，五十嵐陽介，米山聖子，前川喜久雄，“X-JToBI リファレンスマニュアル var.1.3”，2002

[9] 門谷信愛希，阿曽弘具，鈴木基之，牧野正三，“音声に含まれる感情の判別に関する検討”，信学技報，SP2000-82，2000

[10] 武田昌一，西澤良博，大山 玄，“「怒り」の音声の特徴分析に関する1考察”，信学技報，SP2000-164，2001

[11] 飯田朱美，ニック・キャンベル，安村通晃，“感情表現が可能な合成音声の作成と評価”，情報処理学会論文誌，vol40，No2，1999

[12] 大原遼，柏岡秀紀，ニック・キャンベル，“対話音声の笑い声や笑い方についての分析”，日本音響学会講演論文集，2004

[13] Marti Hearst, "Multi-Paragraph Segmentation of Expository Text", Proceedings of the 32nd Annual Meeting of Association for Computational Linguistics, pp.9-pp.16, 1994

[14] 望月源，奥村学，岩山真，"語彙的結束性に基づく語彙的連鎖の計算"，JAIST Technical Memorandum，IS-TM-2000-002，2000

[15] 長尾真，"自然言語処理"，岩波書店，1996

[16] http://chasen.naist.jp/hiki/ChaSen/

[17] 国立国語研究所 (編集)，"分類語彙表"，大日本図書出版，2002

[18] shulsberg.H, "The Description of Facial Expressions in Terms of Two Dimension", Jounal of ExperimentalPscology, 44, pp.229-237

[19] http://chasen.org/ taku/software/cabocha/

[20] 情報処理振興事業協会，"IPALexicon of the Japanese Language for computers"，1999

[21] 瀬谷正敏，"対人関係の真理"，培風館，1977

[22] http://www.digital-lib.nttdocomo.co.jp/

[23] http://homepage1.nifty.com/ohanasi/

[24] http://www.geocities.jp/medeta/

[25] http://www.h5.dion.ne.jp/ jupiter/

[26] http://s-1670.hp.infoseek.co.jp/

[27] ttp://www.aozora.gr.jp/

[28] http://www.chiiki-dukuri-hyakka.or.jp/

[29] http://www.fujipan.co.jp/minwa/mi_index.html

[30] http://www.chiiki-dukuri-hyakka.or.jp/home.html

[31] http://www.enjoy.ne.jp/ toyohama/index.htm

[32] C.Rose, B.Guenter, B.Bodenheimer, M.F.Cohen, "Efficient generation of motion transitions using spacetime constraints", SIGGRAPH 96, pp.147-154, 1996

[33] 稲田和彦，佐藤潤一，中川雅通，浜田正弘，前原文雄，"関数近似と逆運動学を用いた人物動作の生成方法"，画像電子学会研究予稿，94-01-02，1994

[34] M.Unuma, R.Takeuchi, "Generation of Human Walking Motion with Emotion for Computer Animation", IEICE Journal, J76-D-II(8), pp.1822-1831, 1993

[35] M.Kraus, "Human motion and emotion parameterization", CESCG 2004, 2004.

[36] T Komura, Y Shinagawa, and T L Kunii, "Muscle-based feed-forward controller of human body", Computer Graphics Forum, 16(3), pp.165-176, 1997

[37] T Komura, Y Shinagawa, and T L Kunii, "Creating and retargetting motion by the musculoskeletal human body model", The Visual Computer, 5, pp.254-270, 2000

[38] Y.Yoshida, Y.Matsuoka, and K.Hachimura, "Processing of Human Body Movement based on Labanotation -LabanReader and LabanEditor-", IPSJ/SIG Computers and Humanities, 38(6), pp.61-68, 1998

[39] 鵜沼宗利，武内良三，"感情を伝え合うコミュニケーションを行うインタフェースエージェント"，情報処理学会研究報告，2004-HI-107，pp33-40，2004

[40] 西田諭司，杉浦彰彦，"アクションユニットを用いた人間の動作の分析・作成"，豊橋技術科学大学

[41] 鵜沼宗利，武内良三，"コンピュータアニメーションにおける感情を伴った人間の歩行動作の生成方法"，電子情報通信学会論文誌，D-II，Vol.J76-D-II，No.8，pp.1822

[42] 森山剛，小沢慎治，"ファジー制御を用いた音声における情緒性評価法"，電子情報通信学会論文誌，J82-D-II，No.10，pp.1710-1719，1999

[43] 森山剛，斎藤英雄，小沢慎治，"音声における感情表現語と感情表現パラメータの対応付け"，電子情報通信学会論文誌，J82-D-II，No.4，pp.1710-1719，1999

[44] 濱治世，鈴木直人，濱保久，"感情心理学への招待：感情・情緒へのアプローチ"，サイエンス社

[45] 鎌原雅彦，宮下一博，大野木裕明，中澤潤，"心理学マニュアル質問紙法"，北大路書房，2000

[特 集] 「音声言語インタフェースの実用化と音声言語対話への展開」

音声対話研究の現状と動向

Trend of Spoken Dialogue Research

小林 哲則 早稲田大学理工学部
Tetsunori Kobayashi School of Science and Engineering, Waseda University.
koba@elec.waseda.ac.jp

Keywords: spoken dialogue system, architecture, voice XML, group conversation.

1. はじめに

音声は，入力にあたって大きな装置を必要とせず，また目や手を使う必要もないことから，携帯型端末におけるデータ入力や，ほかの用途に目・手を拘束される場合のデータ入力に適していると言われてきた．おりしも携帯電話とカーナビゲーションの急速な普及にあいまって，音声対話型の情報アクセス技術に対する期待は急速に高まっている．このような中にあって，音声対話技術はどういう状態にあって，どういう方向に向かっているのか，今後どうあるべきかなどについて，私見を交えて述べてみたい．

音声対話と一口で言ってもさまざまである．対話参加者の構成，対話の目的，対話時に利用する情報など種々の切り口で音声対話を分類することが可能である．まず，対話参加者の構成からは，一対一の対話とグループでの対話に分類できる．対話の目的からは，一方が他方に仕事を依頼する指令型対話と，対話を継続することそのものが目的となる相互交流型の対話に分類できる．多く見られる情報取得を目的とした対話は，情報検索・提示を依頼するものであるから，指令型対話として見なすことができる．また，対話時に利用できるモダリティの種類から，音声情報のみの利用が許されるユニモーダル対話と，画像も含めた情報交換が許されるマルチモーダル対話に分類できる．

本稿では，これらのうち，最も一般に研究・開発が進められている，一対一・指令型対話の現状について概観した後，ほかの型の対話研究の動向について述べ，最後に音声対話システムが広く受け入れられるために今後解くべき問題について考えてみたい．

2. 一対一・指令型対話の現状

従来の対話研究は，一対一の指令型対話を扱ったものが圧倒的に多い．一般に人間-機械の音声対話というと，この形態が想起される．旅行計画の支援問題を扱うDARPA Communicator プロジェクト [Communicator, Levin 00b, Rudnicky 00]，列車の運行案内を行うARISE [Os 99] など欧米の音声対話プロジェクトや，ロボット対話研究の多く [Asoh 99, Hashimoto 97] は，すべてこの対話形態と見てよい．この種の音声対話システムの歴史は古く，情報検索技術あるいは音声認識技術の進歩に相応した進展は見られるもの，システムの振舞いを決める対話技術の本質的な部分は，80年代に発展した理論研究 [Cohen 90, Kobsa 89] の枠組みを大きく超えるものではない．要素技術の進歩は目を見張るものがあるから，実システムを見れば時代の違いは明らかなのではあるが，システムの振舞いにおける特徴を言葉で表現すると，多かれ少なかれどれも昔から言い古されてきた表現に落ち着いてしまうのが現実である．

最近は，対話システムにどのような振舞いをさせるかという中身の問題よりは，それをどう実装するか，どう記述するか，その記述が静的か動的か，といった問題が議論されている．

このような観点での一つの話題は，対話システムを構成するモジュールの構成方式である．現在最もポピュラーな構成は，DARPA Communicator プロジェクトで採用されたGalaxy アーキテクチャ [Polifroni 00] である．Galaxy アーキテクチャは，MIT LCS で開発された対話システムアーキテクチャで，音声認識，音声合成，言語処理など音声対話システムを構成する各要素技術を達成するモジュールをサーバとして配置し，それらの資源をクライアントプログラムから利用することで対話を達成する枠組みである．モジュールごとの独立性を高め，モジュールの入替えを容易にすることで，対話システムの枠組みで，各個別モジュールの評価ができるようになっている．このことによって，対話システムの要素技術単体の開発にのみ興味がある組織でも，容易に対話システムを構成することが可能となり，対話研究に参入しやすい環境を実現している．

Galaxy アーキテクチャは，比較的単純なシーケンスに従って機能モジュールを利用するシステムの構築に向いている．このため，情報の流れをあらかじめ規定できるユニモーダル型の音声対話システムを構築するのには適するが，複雑なシーケンスで機能モジュールが動作す

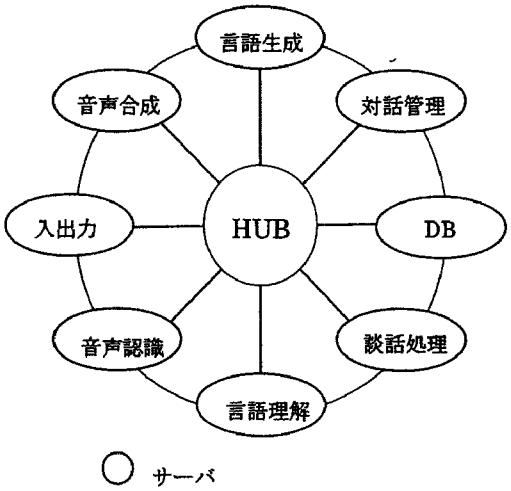


図1 Galaxy アーキテクチャ

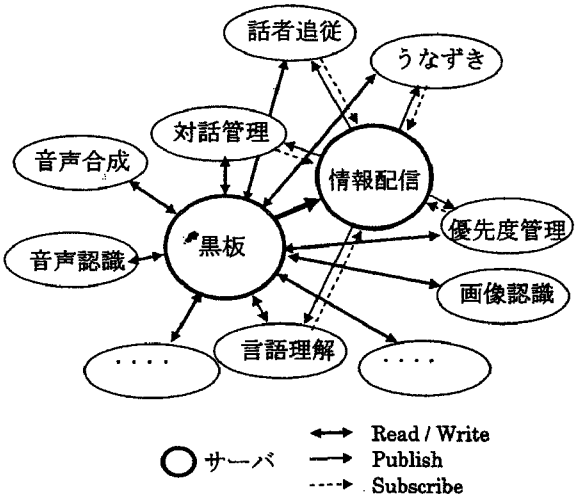


図2 ROBITA アーキテクチャ

ることが要求されるマルチモーダルな対話システムを構成するのは困難である．マルチモーダル型では，画像や音声など複数のモダリティに対応した処理モジュールが並行的に動作し，時に応じて，イベント同期的に，あるいは非同期的に情報を交信する必要がある．

このようなことから，情報共有システムをサーバにおき，各機能モジュールをクライアントとしてモジュール間の協調動作を行わせる方式が提案されている [松坂 01a]．この形態は基本的には黒板システムとなるが，通常の黒板システムではイベントに同期した情報伝送には向かない．文献 [松坂 01a] の ROBITA で採用されたアーキテクチャではこの点に考慮し，情報項目ごとにあらかじめ登録したモジュールには，データが瞬時転送される機能を付加することによって，複数のクライアントプログラムが，並行して動作しながら，イベントに同期した振舞いをすることを可能にしている．この機能によって，マルチモーダル対話システムの開発が非常に容易になっている．

一対一・指令型対話の構成技術に関するもう一つの話題は，対話の流れの記述の問題である．今日音声ポータルなど，Web へのアクセスが音声対話の一つの重要な応用であり，この目的からは Web との親和性の高い対話システムの記述が求められる．Voice XML は，対話システムの振舞いを記述する標準的な言語を目指していることから，現在多くの対話システムが，Voice XML を用いて記述されている [荒木 01, 鯨井 00, VOIZI, V ポータル]．Voice XML は対話システムを一つのコンテンツとして配信可能にすることから，携帯端末や，カーナビ用途には期待されている．

しかしながら，現状における Voice XML は音声対話記述のための基本的な用件を満たしているとは考え難い．Voice XML は音声によるメニュー選択によって，状態遷移を行わせながら対話を進行させるものである．このため，利用者はシステムの状態というものを意識して発話することを強要される．しかし，音声対話においては，多くの場合，利用者はシステムの状態を知ることができない．結果として，その状態において何を話せばよいのかを理解できず，円滑に対話を進めることができない．

Voice XML はまた，マルチモーダルな対話の扱いにも難がある．状態の遷移を決めるのが音声による発話に限られる．カーナビの用途では，運転に集中すべきところでは対話を一時中断させるなど，車の走行状態に応じて対話システムの振舞いを変化させることが望まれるが，このような動作を Voice XML 標準の機能できれいに記述することは困難である．このため，自動車走行電子技術協会にネットワーク型音声利用技術委員会が設置され，車載音声認識器の対話記述のための Voice XML の機能拡張が検討されている [自走協 01]．

Voice XML に対抗して，マイクロソフトが中心となって進めている SALT (Speech Application Language Tags) の動向も目が離せない．Voice XML が比較的単純な音声応答の記述を目指しているのに対し，SALT は，音声とペン入力を含むマルチモーダルな入力を許容したうえで，音声を利用した新しいサービスを Web 上で行うことを目指している．

対話制御に関しては，上で紹介した Voice XML による静的記述のほかに，動的対話制御 [菊池 00] や対話制御戦略の自動学習 [Levin 01a] の問題が興味深いが，これらについては，本特集別稿 [中野 02] に詳しいので，そちらを参照されたい．

3. 対話研究の新しい芽

3·1 相互交流型対話への展開

多くの指令型対話システムの開発が進められているのに対し，相互交流型の対話システムは，ほとんど例がない．現状は，人対人の対話を観察して，対話のモデルを構成している段階といえる．さまざまな対話コーパスが開発され，種々の分析が進んでいるが [堀内 99, 白井 97],

システムとなると，その基礎となる音声認識システムの基本性能からして問題が多い．指令型対話における音声発話は比較的整った読上げ口調でも問題はないが，相互交流型の対話となると，目的が会話の雰囲気そのものを楽しむことにあるため，会話の口調も口語的となる．口語口調の音声認識は，実用にはほど遠い．米国の音声認識プロジェクトは，タスク依存のコマンド音声，新聞記事読上げ音声，放送音声の認識と，設定したタスクの多くについて，数年で認識率90％を超えるという驚異的な技術革新を達成し続けてきたが，こと口語音声の認識については，長い間性能を向上させることができないでいる．日本でも，開放融合プロジェクトの一環で，講演音声の認識が試みられるようになってきた [Furui 00]．今後の成果に期待がかかる．

相互交流型の対話においては，対話のリズムも重要な問題である．古くから，対話のターンテーキングを韻律情報でモデル化しようとした研究が見られる [Osaka 90, 六井 99]．適切なタイミングでうなずきを打たせることで対話のリズムをつくろうとする研究 [Ward 96] などは，この種の研究の一部として位置づけられる．発話計画に割込みの可能性を考慮して発話計画を行う方式 [菊池 94] や，発話途中でも暫時的な音声認識機能を実装することで，発話内容を加味したうえでターンテーキングのリズムをつくろうとした研究もある [Nakano 99]．今後，これらの研究が結びつくことによって，自然なリズムのある対話システムが構築されることが期待される．

このような相互交流型の対話システムの用途としては，痴呆の方の話し相手などが真剣に検討されている．筆者自身は，表層的な受け答えしか当面はできないであろう機械を，こういった方々の相手に充てることには，当初抵抗感が強く疑問視していたが，実際に痴呆の方の面倒を見ていらっしゃる方には切実な問題であるとのことだ．

4. グループ会話研究への展開

グループ会話とは，同じ話題について3人以上の対話参加者が情報のやり取りをするタイプの対話であって，グループウェアのインタフェースのために必要とされる技術である．遠隔会議システムにも応用可能であり，また要素技術は会議記録にも利用できる [Waibel 01]．

グループ会話において，対話の推移を正確に捉えるためには，誰が誰に情報を伝えようとしているのかを推定する問題を解く必要が生じることになる．この問題は視覚的な状況判断によって解かれる部分が多いため，グループ会話は本質的にマルチモーダルのシステムとなる [松坂 01b]．また，ハンズフリーの音声認識装置 [ATR 01] や，複数の人の声が混じった入力でそれらを分離・認識する技術 [村井 00] が求められるなど，技術的にも興味深い課題を多く含むチャレンジングなテーマといえる．

結局のところ，人間と機械の一対一の対話においては，人間が機械にいくらでも合わせることができる．このような状況では，人間は必ずしも音声を利用したいとは思わないかもしれない．しかし，音声が使える状況でありながら，人間と人間とが音声以外の方法で情報交換するというのは，かなり不自然と思われる．こうして考えたとき，人間と人間とのコミュニケーションに割って入る機械があったならば，その機械には音声認識技術が必須となる．なぜなら，このような状況では，対話に参加者の間で効率的に情報を共有する必要が生じるわけであるが，そこで生じる人間と人間との情報交換が，新たに加わった機械の都合で，わざわざ音声対話以外の方法で行われるようになるとは考え難いからである．人間どうしが自然な形で情報交換しながら作業を進めている状況において，その作業を支援できるシステムがあれば有用に違いない．そして，そのようなシステムにはグループ会話の機能が必要となる．筆者が，グループウェアにグループ会話インタフェースが必要と考えている理由がそこにある．

リビングルームにおいて家族が談笑しながら情報家電を利用するといった状況も，家族どうしの対話と情報機器への指令発話とを区別する必要が生じることから，部分的にグループ会話的な要素をもつ．情報機器の家電化，あるいは家電の情報機器化が進む状況の中で，今後グループ会話の重要性は高まるに違いない．

5. 音声対話はヒューマンインタフェースを変えるか

さて，このように研究開発が進む音声対話システムであるが，期待どおりヒューマンインタフェースの世界を変えていく可能性はあるのだろうか．残念ながら，最近運用が始まった音声ポータルサービスを利用してみても音声カーナビのデモを見ても，認識率は飛躍的に向上しているにもかかわらず，音声対話がインタフェースを根本的に変えていく予感を与えるほどの利便性を感じるこ

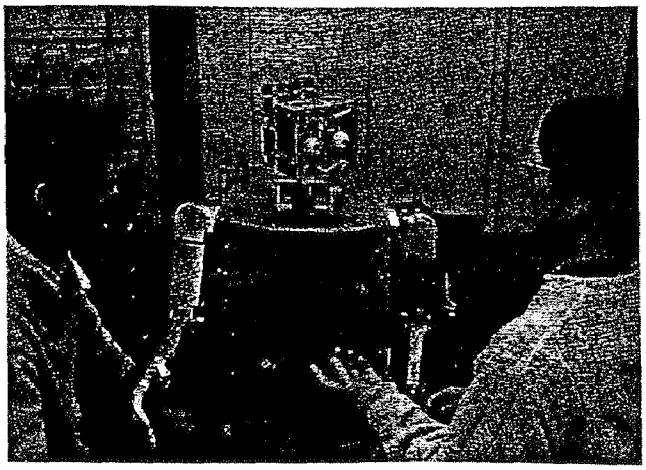


図3 身振りを交えてグループ会話をするROBITA

とは少ない．

アプリケーションの貧弱さもさることながら，音声の利点・欠点を理解したうえでシステムがつくられていないことが，利便性を感じることができない要因のように思われる．

使いやすいインタフェースが満たすべき要件に，透過性の原則 [西本 96, Norman 90] がある．インタフェースは，今システムがどういう状態にあるかが利用者に容易に理解できなければならず（状態理解容易性の原則），そこでどういう操作をすればよいのかが簡単に想像できなければならない（手順連想容易性の原則）．また，利用者の作用の効果がすぐに利用者に対しフィードバックされなければならない（フィードバックの原則）．音声入力においては，本来ならば，思ったことを言葉で伝えれば，相手に指令が届くことが期待される．この意味からは，音声は操作手順の連想は容易であるべきである．しかしながら，Voice XML に関連して述べたように，実際には個々の場面での操作手順（認識できること）にシステム内部の状態が深くかかわる．このため，利用においては状態理解の容易性がより重要になるが，このことに関しては，音声は本質的に不得手である．状態を提示するためには，長々と話さなければならない．状態提示のためのメッセージは，なければ使えないし，あればあったで円滑さを妨げる要因となるという厄介な代物である．また，音は瞬時に消えてしまう性質のものであるから，仮に状態を提示したとしても，それを長く記憶にとどめておくことは難しい．これらの意味で，音声は状態理解の容易性に関してはもともとハンデキャップを背負ったメディアということになる．音声入力に対する期待と，実際のシステムがもつ制約からくる使い勝手とのギャップが，より一層音声入力への失望感を助長させているようにも思う．

しかしながら，音声はもともと状態理解の容易性に問題があることを知ったうえでシステムを構成するならば，状況は改善する可能性をもつ．すなわち，システムの状態依存性を極力排除するとともに，発話以外の方法（例えば簡単な音）で状態を伝える方法を決め，それを定着させる努力をすることが考えられる．

この利用法を定着させるための努力というのは，音声インタフェースを広めるうえで重要な意味をもつ．例えば，単機能だけでもよいから使いこなしたいという要求は多いはずである．であれば，使い込んだとき便利と感じる利用法を用意したうえで，その利用法の周知に手間をかけるべきである．インタフェースの構成原則 [西本 96] における，熟練者優遇の原則（アプリケーションに特化した，効率的な入力手段を用意すべき）と上級利用意向支援の原則（特化手段の存在を知らせ，その利用を促す自然な枠組みを用意すべき）である．利用を促すしくみとしては，インタフェースの中で完結していることが望ましいが，ホームページなどを援用することも考慮すべきだろう．残念ながら，現状のポータルサイトの作りには，そういったものへの配慮がないようだ．

また，カーナビへの応用に関していえば，音声はハンズフリー，アイズフリーの入力であることから，利用時の Distraction を減じることが期待されているわけであるが，音声を使えばそれだけで Distraction の問題から解放されるかといえば，それはあり得ない [Lee 01]．当然のことながら，インタフェースの構成によっては，ハンズフリー，アイズフリーではあっても，マインドフリーでなくなる（思考の束縛を受ける）からである．車内での音声の利用に道が開かれるためには，どのような音声の利用形態であれば思考が拘束されることなく，運転への集中を切らすことがないかを十分に調査する必要がある [自走協 01]．少なくとも，システムの状態を意識しなければ発話できない状態を脱しなければ，音声入力は Distraction フリーであり得ない．

以上述べたように，音声の特質を十分に理解したうえでインタフェースを構成しなければ，音声対話の潜在的な有用性は引き出せない．実用化を指向するならば，今後この部分の検討が不可欠である．インタフェースの原則論は多くの示唆に富んでいる．原則論ベースのシステム設計を今後より重要視すべきであろう．

6. む　す　び

対話を分類したうえで，それぞれの対話形態における現状と動向，さらには今後検討すべき課題について述べた．研究対象として魅力的な展開が見込まれる一方で，実用期を迎えたとはいいながらも真の実用化にはまだ距離がある．今後この距離を埋めるためには，技術的問題の解決もさることながら，インタフェースにおけるヒューマンファクタに関する知見を深め，それをシステム構成に生かすとともに，利用法の教育のあり方についても検討を進めるなど，総合的な取組みが必要である．

◇ 参 考 文 献 ◇

[荒木 01] 荒木雅弘: 音声対話システムと Voice XML, 人工知能学会研究会資料, SIG-SLUD-A103-7 (2001)

[Asoh 99] H. Asoh, T. Matsui, John Fry, F. Asano and S. Hayamizu: A spoken dialog system for a mobile office robot, Proc. Eurospeech'99, pp. 1139-1142 (1999)

[ATR 01] Proc. Workshop on Hands-Free Speech Communication (2001)

[Cohen 90] P. R. Cohen, J. Morgan and M. E. Pollack: Intentions in Communication, The MIT Press (1990)

[Communicator] http://www.darpa.mil/ito/research/com/

[堂坂 02] 堂坂浩二, 安田宜仁, 相川清明: システム知識制限下での効率的音声対話制御法, 自然言語処理, Vol. 9, No. 1, pp. 43-63 (2002)

[Furui 00] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki and T. Ohdaira: Toward the realization of spontaneous speech recognition — Introduction of a Japanese priority program —, Proc. International Conference on Spoken Language

Processing, Vol. 3, pp. 518-521 (2000)
[Hashimoto 97] S. Hashimoto, S, Narita, H. Kasahara, A. Takanishi, S. Sugano, K. Shirai, T. Kobayashi, H. Takanobu, T. Kurata, K. Fujiwara, T. Matsuno, T. Kawasaki and K. Hoashi: Humanoid Robot — Development of an Information Assistant Robot Hadaly —, 6th IEEE International Workshop on Robot and Human Communication(RO-MAN'97) (1997)
[堀内 99] 堀内靖雄, 中野有紀子, 小磯花絵, 石崎雅人, 鈴木浩之, 岡田美智男, 仲真紀子, 土屋 俊, 市川 熹: 日本語地図課題対話コーパスの設計と特徴, 人工知能学会誌, Vol. 14, No. 2, pp. 261-272 (1999)
[自走協 01] 車載システムにおけるネットワーク型音声利用技術に関する調査研究報告書, (財) 自動車走行電子技術協会 12-R-4 (2001)
[菊池 94] 菊池英明, 工藤育夫, 小林哲則, 白井克彦: 音声インタフェースにおける発話権管理による割り込みへの対処, 信学論 (D-II), Vol. J77-D-II, No. 8, pp. 1502-1511 (1994)
[菊池 00] 菊池英明, 白井克彦: 対話効率の向上を目的とした音声対話制御のモデル化, ヒューマンインタフェース学会, Vol. 2, No. 2, pp. 145-152 (2000)
[Kobsa 89] A. Kobsa and W. Wahlster (Eds.): User Models in Dialog Systems, Springer-Verlag (1989)
[鯨井 00] 鯨井俊宏, 高橋 久, 天野明雄, 畑岡信夫: Voice XML インタプリタと連続単語認識エンジンの開発—音声ポータル向け音声認識技術の開発—, 情報処理学会研究報告, SLP-33-12 (2000)
[Lee 01] J. Lee, K. Schmidt and T. Bral: Distraction potential of speech-based driver interfaces, International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design (2001)
[Levin 00a] E. Levin, R. Pieraccini and W. Eckert: A stochastic model of human-machine interaction for learning dialogue strategies, IEEE Trans. on Speech and Audio Processing, Vol. 8, No. 1, pp. 11-23 (2000)
[Levin 00b] E. Levin, S. Narayanan, R. Pieraccini, K. Biatov, E. Bocchieri, G. DiFabbrizio, W. Eckert, S. Lee, A. Pokrovsky, M. Rahim, P. Ruscitti and M. Walker: The AT & T-DARPA Communicator mixed initiative spoken dialogue system, Proc. International Conference on Spoken Language Processing (2000)
[松坂 01a] 松坂要佐, 小林哲則: 異分野エキスパート集団によるインタラクティブロボットの共同開発のための基本ソフトウェア, 情報処理学会第126回知能と複雑系研究会 (2001)
[松坂 01b] 松坂要佐, 東條剛史, 小林哲則: グループ会話に参与する対話ロボットの構築, 信学論 (D-II), Vol. J84-D-II, No. 6, pp. 898-908 (2001)
[村井 00] 村井則之, 小林哲則: 話者性と発話交替を考慮した複数話者対話音声の認識, 信学論 (D-II), Vol. J83-D-II, No. 11, pp. 2465-2472 (2000)
[Nakano 99] M. Nakano, N. Miyazaki, J. Hirasawa, K. Dohsaka and T. Kawabata: Understanding un-segmented user utterances in real time spoken dialogue system, Proc. 37th ACL, pp. 200-207 (1999)
[中野 02] 中野幹生, 堂坂浩二: 音声対話システムの言語・対話処理, 人工知能学会誌, Vol. 17, No. 3, pp. 271-278 (2002)
[西本 96] 西本卓也, 志田修利, 小林哲則, 白井克彦: マルチモーダル入力環境下における音声の協調的利用—音声作図システム S-tgif の設計と評価—, 信学論(D-II), Vol. J79-D-II, No. 12, pp. 2176-2183 (1996)
[Norman 90] D. A. Norman, 野島 久雄 訳: 誰のためのデザイン? —認知科学者のデザイン原論, 新曜社 (1990)
[Os 99] E. D. Os, L. Boves, L. Lamel and P. Baggia: Overview of the arise project, Proc. European Conference on Speech Technology, Eurospeech99, pp. 1527-1530 (Sep. 1999)
[Osaka 90] N. Osaka: Conversational Turn-taking Model Using Petri Net, International Conference on Spoken Language Processing 90, pp. 1297-1300 (1990)
[Polifroni 00] J. Polifroni, S. Seneff: Galaxy-II as an Architecture for Spoken Dialogue Evaluation, International Conference on Language Resources and Evaluation (LREC) (2000)
[六井 99] 六井 淳, 坂本彰司, 岡田美智男: 多人数会話における会話調整とその方略の獲得, 情報処理学会研究報告, SLP-28-8 (1999)
[Rudnicky 00] A. Rudnicky, C. Bennett, A. Black, A. Chotomongcol, K. Lenzo, A. Oh and Singh, R.: Task and domain specific modeling in the Carnegie Mellon communicator system, Proc. International Conference on Spoken Language Processing (2000)
[白井 97] 白井克彦: 音声対話コーパスと対話過程のモデル, 人工知能学会誌, Vol. 12, No. 1, pp. 30-35 (1997)
[田中 96] 田中修一, 中里 収, 帆足啓一郎, 白井克彦: 複合作業時における音声インタフェースの設計と評価, 信学論 (D-II), Vol. J79-D-II, No. 12, pp. 2163-2169 (1996)
[VOIZI] http://www.voizi.net/
[V ポータル] http://www.ntt.com/v-portal/
[Waibel 01] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu and K. Zechner: Advances in automatic meeting record creation and access, Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2001 (May 2001)
[Ward 96] N. Ward: Using prosodic clues to decide when to produce back-channel utterances, Proc. International Conference on Spoken Language Processing (1996)

2002年3月26日 受理

著者紹介

小林 哲則 (正会員)

1980年早稲田大学理工学部電気工学科卒業, 1985年同大学院博士課程修了. 工学博士. 法政大学講師・助教授, 早稲田大学助教授を経て, 1997年より早稲田大学理工学部電気工学科教授. 音声情報処理, 画像情報処理, およびそれらを統合したヒューマンインタフェースに興味をもつ. 電子情報通信学会, 情報処理学会, 日本音響学会, 日本ロボット学会などの各会員.

特徴量レベルでの統合法に基づくマルチバンド型モデルによる雑音環境下音声認識

大川茂樹† 白井克彦††

本論文では，マルチバンド型と呼ばれる音声認識モデルにおいて，部分周波数帯域の情報を統合する新しい方法を提案する．マルチバンド型音声認識は，入力音声信号を複数の周波数帯域に分割し，各帯域を独立に処理した後に再統合するという認識手法であり，近年，特に雑音環境下で良い性能を与えることが示されている．このようなモデル化では，(i) 再統合の方法，(ii) 帯域分割の方法について検討する必要がある．本研究では，まず (i) の問題に対して，すでに Bourlard らにより提案されている HMM の尤度レベルの再統合に基づく方法（LC 法）の追試を行うとともに，新たに音響特徴量レベルの再統合法（FC 法）を提案し，比較評価を行う．また，(ii) の問題に対しては，FC 法において部分帯域の特徴量と音素モデルとの相互情報量を評価基準とした分割の最適化を試みる．偏帯域性雑音が付加された音声を用いた実験の結果，FC 法に基づくシステムは，従来型システムおよび LC 法に基づくシステムの双方に対してより高い認識性能を与えた．また，情報量を基準とした分割点最適化の効果が確認された．

Noisy Speech Recognition by Multi-band Modeling Based on Feature-level Combination

SHIGEKI OKAWA† and KATSUHIKO SHIRAI††

This paper presents a new approach for sub-band recombination in the framework of multi-band ASR. Recent works suggest that multi-band ASR, which is based on independent processing and recombination of partial frequency bands of input speech, gives more accurate recognition, especially in noisy acoustic environments. In the case, we need to discuss (i) how to recombine the sub-band output, and (ii) how to split the input speech frequencies. We propose and evaluate "feature combination" (FC) approach, as a solution of the above point (i), instead of "likelihood combination" (LC) approach proposed by Bourlard et al. Also for the point (ii), we introduce the mutual information between sub-band features and target phoneme categories to find the optimal splitting frequencies. The experimental results show that the FC-based system can yield better performance both the conventional ASR and the LC-based system for band-limited noisy speech. Also, we could obtain a favorable band-splitting strategy by using the optimization method.

1. はじめに

音声認識技術の実用化において，頑健性すなわち雑音や環境の変動に対する耐性は重要な問題である[1]．従来より，雑音環境下などにおける音声認識の頑健性を論じた研究は数多く行われており，様々な知見や指針を得ることができる．たとえば，入力音声中の雑音成分を推定して周波数軸上で差し引く方法[2]や，雑音自体をモデル化する方法[3]などが代表的なものである．

ところで，伝統的な音声認識では，入力音声の全周波数帯域から計算された音響特徴量（ケプストラムなど）をベクトル特徴として用いることが多い．この場合，もしも周波数帯域の一部が実環境下で多く見られるような偏帯域性雑音☆ などにより汚損されていると，ベクトル特徴のすべての要素が少なからず影響を受けてしまう．この問題を回避するため，近年，複数の部分周波数帯域の特徴量をそれぞれ独立にモデル化するマルチバンド（複合周波数帯域；Multi-band）型の音声認識手法（以下 MB-ASR と記すことがある）が提

† 千葉工業大学情報科学部情報ネットワーク学科
Department of Information and Network Science, Chiba Institute of Technology
†† 早稲田大学理工学部情報学科
Department of Information and Computer Science, Waseda University

☆ 特定の周波数帯域にエネルギー成分が集中した雑音．自動車のエンジン音や計算機のファンの音などもこの性質を持つ．

案され，雑音環境下あるいは不整合な環境下において良好な頑健性を示している[4)~8)]．

本論文では，MB-ASR の枠組みにおいて，部分周波数帯域から得られた情報を特徴量のレベルで統合する方法を提案し，HMM の尤度レベルでの統合法との比較および種々の雑音環境下における有効性を検証する．また，情報量を評価基準とした帯域分割点の最適化について言及する．

以下，2 章では MB-ASR の基本的な考え方とその利点を説明し，3 章では 2 種類の再統合レベルについて検討する．また，4 章では帯域分割の最適化の手法を示す．5 章では実験におけるデータや分析の条件を述べ，6 章で音声認識実験の結果を示し考察する．最後に 7 章で結論を述べる．

2. マルチバンド型音声認識

2.1 基本的な考え方

マルチバンド型音声認識（MB-ASR）は，入力音声信号の周波数帯域全体を複数の部分に分割し，それぞれ独立に分析や音響尤度計算の処理を施し，後段のどこかの時点で個々の出力（尤度など）を再統合するという考え方に基づいている．前章で触れた耐雑音性のみならず，以下で述べるように人間の聴覚機構にも通ずるアプローチであることから，音声認識の新しいパラダイムとして注目され始めている．

図 1 に，MB-ASR の基本的な考え方を示す．図中で，入力音声の低い周波数帯域が局所的に雑音で汚損されている．このとき，(a) 従来型の認識手法では，生成される特徴量ベクトルの全体に局所雑音の影響が及んでしまうのに対して，(b) マルチバンド型モデルでは，低い帯域に対応するベクトル要素のみに雑音の影響を「封じ込め」，他の帯域の情報損失を軽減できる．

2.2 研究の意義

MB-ASR を積極的に追求する意義として，以下にいくつかの根拠をあげる．

- マルチストリーム音声処理：
  複数の情報源から得られた観測あるいは尤度を統合することで，より安定した情報抽出が可能になるという「マルチストリーム音声処理」（Multi-stream ASR；多元情報に基づく音声処理）の考え方に包含される[9)]．
- 偏帯域性雑音：
  我々の身の周りに存在する雑音は，一部の周波数帯域に成分が集中している（偏帯域性である）ことが多い．上で述べたように，部分帯域からの出力を再統合する際に，雑音を含む帯域の影響を低減させる仕組みを導入することが可能である．
- 心理音響的な裏付け：
  人間の聴覚機構は，周波数帯域をいくつかの狭い区間に分けてそれぞれ独立に処理しているといわれている[10)]．
- 「次元の呪い」の回避：
  分析する帯域が広いほど特徴量空間の次数が増大し，いわゆる「次元の呪い（curse of dimensionality）」問題をもたらす．各部分帯域の特徴量空間を高い次元で表現すれば，同じ計算量で全帯域型よりも精密なモデル化が行える可能性がある．

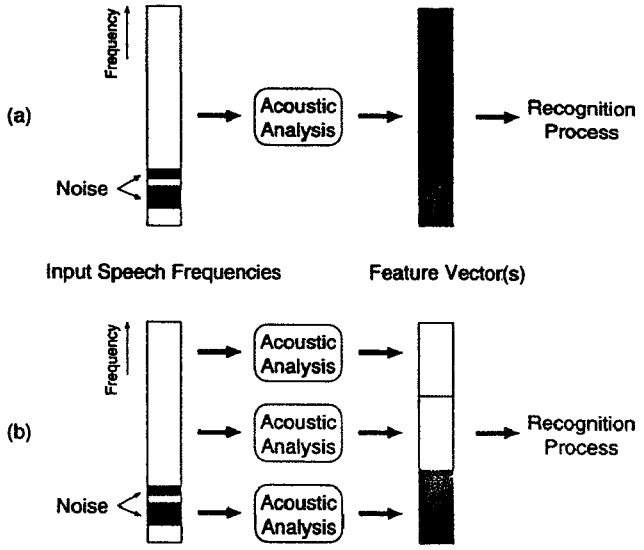


図 1 (a) 全帯域型（従来型）音声認識と (b) マルチバンド型音声認識の考え方
Fig. 1 Schematic diagrams of (a) full-band (conventional) ASR and (b) multi-band ASR.

他にも，異なる情報源を取り扱うマルチストリーム処理に比べ信号処理がしやすい，計算機への実装時に並列処理がしやすいなどの利点がある．

2.3 検討すべき問題

MB-ASR の実現に際しては，従来型の音声認識における様々な条件に加えて，次の 2 つの主要な検討事項がある．

(1) 各部分帯域の処理結果より最終的な決定（認識）をする際，「どの時点」で「どのように」再統合すればよいかという問題．

(2) 部分帯域を定義する際，「いくつの帯域」を「どこで」分割して用いるかという問題．

本論文では，まず (1) の問題に対して，特徴量を生成する時点で再統合する方法を提案・検討し，次に (2) の問題に対して，情報量を評価基準とした分割数および分割点の最適化を試みる．

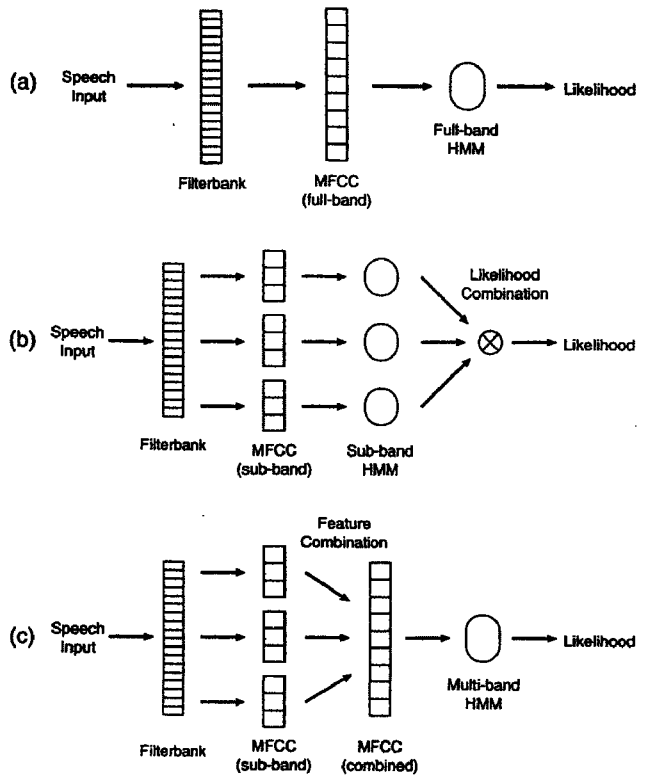


図2 (a) 全帯域システム（従来型；FB-ASR），(b) LC 法に基づく MB-ASR，(c) FC 法に基づく MB-ASR のデータの流れ．
Fig. 2 Diagrams of (a) full-band (conventional) ASR, (b) MB-ASR (likelihood combination), and (c) MB-ASR (feature combination).

3. 再統合レベルの検討

MB-ASR の原型は，1996 年に Bourlard らと Hermansky らによってほぼ同時に提案された[4),5)]．彼らの研究では，各部分帯域に対して異なる識別器が適用された後，いくつかの統合時点（HMM の状態・音素・単語など）において尤度の統合が試みられている．

本研究では，Bourlard らや Hermansky らの方法（尤度レベルの統合法；以下「LC 法」と呼ぶ）の追試に加え，異なる統合法を導入し，種々の雑音条件下における比較を試みる．新たな方法は，部分帯域から得られた特徴量を識別処理の前に連結し，単一の特徴ベクトルとして用いるものである（特徴量レベルの統合法；以下「FC 法」と呼ぶ）．図 2 に，(a) 全帯域型（従来型）の音声認識手法（Full-band ASR；以下 FB-ASR），(b)LC 法に基づく MB-ASR，(c) FC 法に基づく MB-ASR におけるデータの流れを示す．

本章では，LC 法および FC 法の概念および特長をそれぞれ説明する．

3.1 尤度レベルの統合法（LC 法）

まず，LC 法では，各部分帯域から得られた特徴量ベクトルを用いて，あらかじめ独立に HMM を学習する．認識時にも，各帯域に対してそれぞれ異なる識別器を独立に適用し，各識別器から認識仮説（音素や単語の候補）とその尤度を得る．その後，すべての識別器からの出力を統合して全体の尤度を計算することで認識結果が得られる（図 2 (b) 参照）．

各帯域の尤度は，それぞれの帯域に対応する HMM により計算・出力される．帯域間の独立性を仮定すれば，すべての出力確率を単純に掛け合わせることで全体の尤度が得られる．

この際，不要と判断された帯域を使用しない方法[11)]や，帯域ごとに何らかの重み係数を設定し，$p(o_i^b|s_j)$ を HMM の出力確率（状態 s_j，時刻 i，帯域 b），w_b を帯域 b に対する重み係数として，

$$p(o_i|s_j) = \prod_b p(o_i^b|s_j)^{w_b}, \tag{1}$$

のように重み付け確率計算を行う方法[12)]なども考えられるが，本論文では，MB-ASR において部分帯域への分割と独立処理そのものが持つ効果に注目するため，重み付けは行わずに議論を進める．すなわち，LC 法における確率計算は，

$$p(o_i|s_j) = \prod_b p(o_i^b|s_j), \tag{2}$$

のように行う．

なお，Bourlard らの報告[4)]によると，尤度レベルで統合を行う単位（尤度の統合計算を行う時点）として，HMM の状態（=分析フレーム）を用いても，言語的にさらに長い単位（たとえば音素や単語）を用いても，ほぼ同等の性能が得られるとされている．HMM の状態を単位とした統合は，明らかに実装が簡単であるので，ここでは LC 法での統合の単位には HMM の状態を用いるものとする．

3.2 特徴量レベルの統合法（FC 法）

次に，本研究で新たに提案する FC 法について述べる．

LC 法では，各部分帯域から得られた特徴量に対してそれぞれ尤度を計算した後で統合したが，FC 法では，各帯域の特徴量を得た直後に（尤度計算以前に）それらを統合する（図 2 (c) 参照）．

具体的には，適当な数に分割した部分帯域の各々について，音響特徴量を独立に計算した後，連結して 1 つのベクトルを構成し，識別器への入力として用いる．LC 法での式 (2) に対して，FC 法での出力確率は，

$$p(o_i|s_j) = p(o_i^1, \ldots, o_i^B|s_j), \tag{3}$$

(B は帯域数)

と表すことができる．この場合，識別器以降の処理は，基本的に従来型の音声認識と同等になるため，認識システムへの実装は比較的容易である．

このFC法は，MB-ASRの一般的な利点とFB-ASR（従来型の音声認識）の利点との両者をあわせ持つと考えられる．すなわち，

- 一部の帯域が汚損されている場合，その影響は，分割の結果として特徴量ベクトルの一部の要素のみに影響し，他の帯域の情報は極力損失させないように作用するであろう（MB-ASRの利点）．
- 特徴ベクトルのすべての要素が統計モデル内で一括表現されるので，複雑な再統合機構を必要としない．また，HMMの出力確率計算時には，一般的な特徴ベクトルと同様に全相関あるいは混合確率密度分布を用いることにより，各帯域間の相関を考慮することが容易である（FB-ASRの利点）．

ただし，FC法では，音素や単語などを単位とした統合は不可能である．本研究では，比較するLC法においても，前述のようにBourlardらの実験結果をふまえてHMMの状態（=分析フレーム）を統合の単位とするが，仮に音素や単語などの単位での統合がより良い結果を与えるような条件が存在する場合，それはLC法においてのみ可能であることを付記しておく．

なお，FC法においても，各帯域が音声認識に対して持つ情報の量に応じてベクトル構成時に重み付けを行う方法が考えられるが，本研究では，LC法の場合と同様の理由により，特に重み付けなどの処理は施さず，各帯域から得られた特徴ベクトルを単純に連結してHMMへの入力として用いている．

4. 情報量に基づく帯域分割の最適化

マルチバンド型モデルを構成する際，どのように周波数帯域を分割するかは重要であるが複雑な問題でもある．特に，音声信号に偏帯域性雑音が含まれるような場合，雑音の性質により最適な分割点は変動するので，入力音声に応じた最適化が必要となる．雑音なし音声の場合も，音声のスペクトル特徴自体が持つ周波数構造（たとえば母音のホルマント位置との関連など）により，帯域により分割の影響が異なると考えられる．

スペクトルの重要な成分のみを選択的に用いて音声認識を行う方法としては文献13)などがあるが，MB-ASRの枠組みの中で，入力に応じた分割点の最適化に言及した研究はほとんどない．このことから，本論文では，音声および雑音の双方の性質をふまえた帯域分割最適化の提案として，FC法において，あらかじめ設定した複数の帯域分割条件に対し，特徴ベクトルと音素HMMの全状態との相互情報量が最大となるものを逐次選択する方法を導入する．

この方法では，全帯域の特徴量ベクトルをそのまま用いる場合も選択候補に含めることができるため，従来型の音声認識が最も良い性能を与える条件でも最適なモデルが選択できるという利点がある．

部分帯域が音声認識に対して与える情報の量は，観測系列の条件付きエントロピーを用いると都合良く定義できる．ただし，条件付きエントロピーは，各帯域間での相対値として扱う必要があるので，事後確率により計算しなければならない．ここで，部分帯域 b に対する時刻（フレーム）i のHMM出力確率を $p(o_i^b|s_j)$ とすると，Bayesの定理により事後確率は，

$$p(s_j|o_i^b) = \frac{p(s_j)p(o_i^b|s_j)}{p(o_i^b)}, \tag{4}$$

となる．ここで，$p(o_i^b)$ は $p(s_j|o_i^b)$ の大小には無関係であり，また言語確率 $p(s_j)$ は，単独の確率変量としては意味を持つものの，部分帯域間の比較においては同じ条件（本論文ではHMMの状態遷移を帯域ごとに変えるような機構は用いていない）であるため，次の近似式により事後確率を計算する．

$$p(s_j|o_i^b) \cong \frac{p(o_i^b|s_j)}{\sum_j p(o_i^b|s_j)}. \tag{5}$$

すべての音素カテゴリおよびHMM状態に対応する条件付きエントロピー $H(S|o_i^b)$ は，観測系列の特徴量ベクトル o_i^b が与えられたとき，次のようになる．

$$H(S|o_i^b) = \sum_j -p(s_j|o_i^b)\log p(s_j|o_i^b). \tag{6}$$

さらに，適当な時間長の複数フレームに対する平均エントロピー $H(S|O^b)$ は次のようになる．

$$H(S|O^b) = \sum_i \sum_j -p(s_j, o_i^b)\log p(s_j|o_i^b). \tag{7}$$

この値は，系列 $O^b = \{o_i^b\}$ が観測されたときに，最も起こりうるHMMの状態を決定する際の曖昧さ（情報量の減少の度合い）を意味する．

同様に，HMM状態の集合 S の自己エントロピーは，事前知識（たとえば言語確率など）に基づき次のように O^b に無関係な定数として定義できる．

$$H(S) = \sum_j -p(s_j)\log p(s_j). \tag{8}$$

こうして得られた $H(S|O^b)$ と $H(S)$ により，状態の集合 S と適当な時間長の観測ベクトルの集合 O^b との間の相互情報量 $I(S;O^b)$ が次式で定義できる☆．

☆ ここで，$H(S)$ は O^b に無関係であるが，$I(S;O^b)$ の計算に必要である．本研究では，学習データ中の音素数から実際に $H(S)$ を計算した4.977という値を用いた．

$$I(S;O^b) = H(S) - H(S|O^b). \quad (9)$$

この値は，音声認識に対して帯域 b の観測 O^b が持つ情報量を表す．本研究では，様々な分割点または分割点の組合せに対して $\sum_b I(S;O^b)$ を計算したとき，最も大きな値を与える帯域数および分割点を選択し，帯域分割の最適化を行う．

5. 実験条件

本章では，実験に用いたデータや音響分析の条件，認識モデルや帯域分割の条件などについて説明する．

5.1 音声・雑音データ

音声データには，ATIS（Air Travel Information Service）連続音声認識タスク（英語）を用いた[14]．音声は，実験室環境で接話型マイクロフォンを用いて録音されたものである．学習には，19,507 文章（話者 528 名）を用いて，評価には，別の 981 文章（話者 24 名）を用いた．

付加雑音としては，まず各帯域数の条件で最初の（最も低い）周波数帯域のみに白色雑音を加えた 'LPW' 雑音を定義し生成した．これは，いわば MB-ASR における「理想的な」雑音である．帯域を制限した白色雑音を生成する処理は，計算機上で FIR フィルタを設計して行った．次に，実雑音として，NOISEX-92 データベース[15]に含まれる 15 種類の雑音を計算機上でそれぞれ音声データに加えた．いずれの場合も，評価用データには雑音を加えるが，学習用データには加えていない．すなわち，HMM はつねに雑音を含まない条件下で学習される．

5.2 音響分析と認識モデル

認識のフロントエンドとなる音響分析は，フィルタバンク分析および離散コサイン変換（DCT）に基づく MFCC（Mel-Filterbank Cepstrum Coefficients）分析を用いた．

まず 16 kHz で標本化された入力音声をフレーム長 20 ms，分析周期 10 ms のハミング窓で切り出し，FFT 計算によりメルフィルタバンク係数を得る．これを帯域ごとの DCT によりメルケプストラム係数（MFCC）に変換する．この際，各入力に対して，全フレームの MFCC ベクトルから 1 入力あたりの平均ベクトルを差し引く CMS（Cepstrum Mean Subtraction）法を適用している．

認識器は，4 状態 3 ループ，16 混合の文脈独立音素 HMM（音素数 49）とし，言語モデルには単語バイグラムのみを用いる．認識タスクは，連続単語認識であり，以下の結果は単語誤り率で示す．

5.3 帯域分割と特徴ベクトルの構成

フィルタバンク分析は，メル周波数軸上で 31 個の三角窓を等間隔に配置して実現する．比較実験のために設計する全帯域システム（従来型の音声認識手法に相当）では，この 31 次フィルタバンクのすべての次元に対して DCT を適用して MFCC 特徴量を得る．マルチバンド型システムでは，あらかじめ定義した条件（後述）により 31 次フィルタバンクをいくつかの帯域に分割し，それぞれ独立に DCT を適用して複数の MFCC 特徴量を得る（図 2 (b)，(c) 参照）．

帯域数として，当初 1（全帯域），2，3，4 および 6 の場合について実験を行ったが，このうち 4 と 6 の場合は，予備実験の結果，本研究の条件下では十分な性能が得られないことが確認されたため，以下の実験は 1，2 および 3 について行っている．

ちなみに，各帯域数においてメル周波数軸上で等間隔に帯域分割した場合，分割点に相当する周波数（各フィルタに対応する三角窓の中心周波数）は次のようになる（単位は Hz）．

- 2 帯域：0-1772-8000
- 3 帯域：0-1100-2853-8000
- 4 帯域：0-900-1772-3452-8000
- 6 帯域：0-600-1100-1772-2853-4595-8000

MFCC の次数は，全帯域に対して 12 とし，部分帯域のフィルタ数におおむね比例させて決定した．たとえば，等間隔に分割する場合，2 帯域では 6，3 帯域では 4 のようになる．

各フレームの特徴量ベクトルは，この MFCC に dB を単位とする帯域内のフレームパワー，両者の Δ 特徴および ΔΔ 特徴を加えて構成する．FC 法（特徴量レベルの統合法）では，パワー特徴が加わることにより，ベクトルの次元は，たとえば 3 帯域の場合 $3\times3\times(4+1) = 45$，6 帯域の場合 $6\times3\times(2+1) = 54$ のようになる（いずれも分割点の最適化を行わない等間隔分割の場合）．

5.4 分割点最適化

情報量を評価基準とした帯域分割の最適化は，フィルタバンクの分割点を変えることにより実現する．たとえば，3 帯域の場合，(0-a) (a-b) (b-8000) Hz のような分割における a と b のあらゆる組合せ（フィルタバンクを構成する各フィルタが最小単位）に対して，4 章で述べた相互情報量を計算し，最大値を与えた分割点を利用する．

本研究では，評価用データのうち各話者につき 1 文ずつ計 24 文を分割点最適化のための適応化学習に用いることとし，式 (9) に基づき帯域数および分割点を

決定した後，残りの957文について認識実験を行った．

6. 実験結果と考察

6.1 LPW雑音によるLC法とFC法の比較

まず，最初の帯域のみに白色雑音を付加した'LPW'雑音を用いて音声認識実験を行う．先にも述べたように，これは，MB-ASRにおける理想的な雑音であるので，最も効果を確認しやすい条件といえる．

図3に，様々なSNRでLPW雑音を加えた場合と雑音なし（clean）の場合の認識性能の変化を，全帯域（従来型）システム，マルチバンド型システムのLC法およびFC法について示す．帯域数はすべて3とした（等間隔分割）☆．

図3から，clean（雑音なし）の場合を除きマルチバンド型システムが従来型システムよりも高い性能を与え，また，非常に低いSNR（0〜5 dB）以外では，FC法がLC法よりも高い性能を与えていることが分かる．SNR = 10 dBの場合，誤り改善率はLC法で25.3%，FC法で34.0%であった．

6.2 各種雑音に対する頑健性

次に，種々の雑音に対する認識性能について比較評価する．ここでは，LPW雑音のほかに，NOISEX-92データベースの15種類の実雑音を評価用データにSNR = 10 dBで付加した．

図4は，雑音を加えた評価用データに対して，全帯域（従来型）システムと，LC法およびFC法に基づくマルチバンド型システム（帯域数は2および3；等間隔分割）で実験を行った結果をまとめたものである．最も右側のaverageは，LPW雑音を除く15種類の雑音に対する結果の平均値を示す．

図4では，すべての雑音条件に対してFC法がLC法よりも良い性能を与えている．また，帯域数2および3の場合ともに，多くの雑音条件下において全帯域（従来型）システムよりも良い性能を与えている．総合的には，帯域数2のFC法が最も高く，かつ全帯域システムを上回る性能を与えた（全雑音に対する平均値であまり効果が見られないのは，後述するように，一部の雑音に対してMB-ASRがまったく有効に働かなかったためである）．全帯域システムに対する認識性能の向上は，雑音の種類に依存し，destroyer-engine型雑音では17.6%もの誤り改善（全帯域で30.6% → 帯域数2のFC法で25.2%）が得られた．

個々の雑音について見ると，babble，destroyer-engine，factory2，hfchannel，machinegunの場合，マルチバンド型システムは全帯域システムより高い性能を与える．これらの雑音は，信号のエネルギーが一部の周波数帯域に集中するという共通した性質を持つ．これに対して，buccaneer2，destroyer-ops，pink，whiteの場合，マルチバンド型システムは全帯域システムよりも性能が低下してしまう（whiteの場合19.2%の誤り率増加）．これは，これらの雑音の信号エネルギーが周波数軸上のほとんどすべての領域に対して広がっているためであると考えられる．このことから，周波数帯域全体に影響を及ぼすような雑音に対してはMB-ASRは有効でないといえる．ちなみにこの結果は，Hermanskyらの実験結果に符合する[16]．

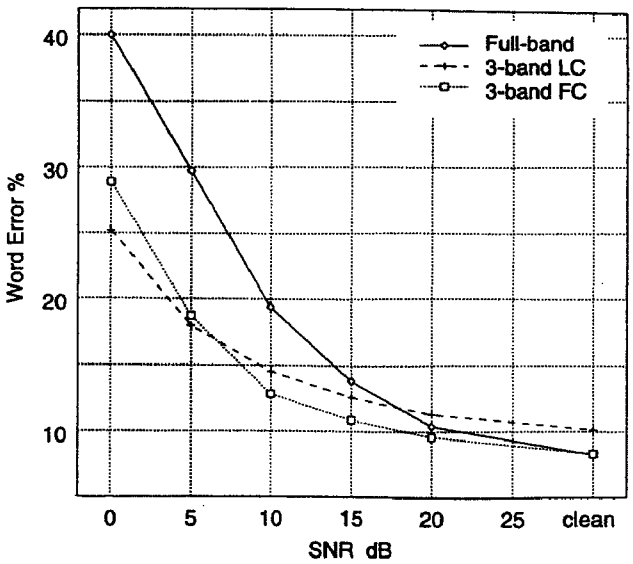


図3 LPW雑音を付加した場合の単語誤り率の比較
Fig. 3 Word error rate with 'LPW' additive noise.

6.3 分割点最適化に関する予備実験

次に，FC法に基づき，情報量を評価基準とした帯域分割の最適化を試みる．認識実験に先立って，部分周波数帯域における相互情報量の性質を調べるための予備実験を行う．本論文では，雑音付加音声に対する認識性能向上を目指しているが，ここでは特に音声のスペクトル特性や人間の聴覚特性との関係を調べるために，雑音なし音声についての結果を示す．

図5には，低帯域のみ，あるいは高帯域のみを取り出した偏帯域音声において遮断周波数を変化させた場合と，2帯域システムの帯域分割点を変化させた場合の相互情報量を同時に示す．

偏帯域音声に対するグラフがちょうど交叉する点（約1,450 Hz）付近で，2帯域システムにおける相互情報量の最大値が得られている事実は興味深い．なぜなら，偏帯域音声に対する人間の聴覚特性を調査した実験結果[10]によると，本実験と同様に遮断周波数を

☆ LPW雑音の場合，帯域数により雑音成分を含む帯域幅が異なるため，帯域数を変えた場合の比較はあまり意味を持たない．ここでは，典型的な場合として帯域数3の結果を示している．

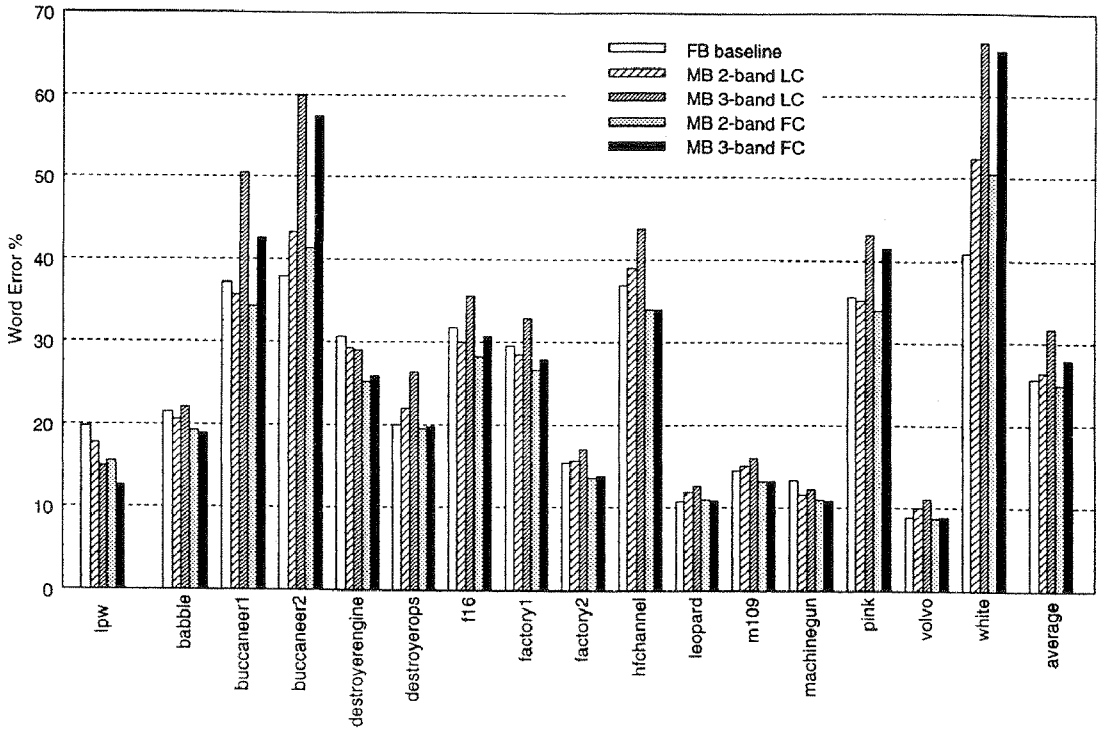


図4 様々な雑音を付加した場合の単語誤り率の比較
Fig. 4 Word error rate with various additive noise.

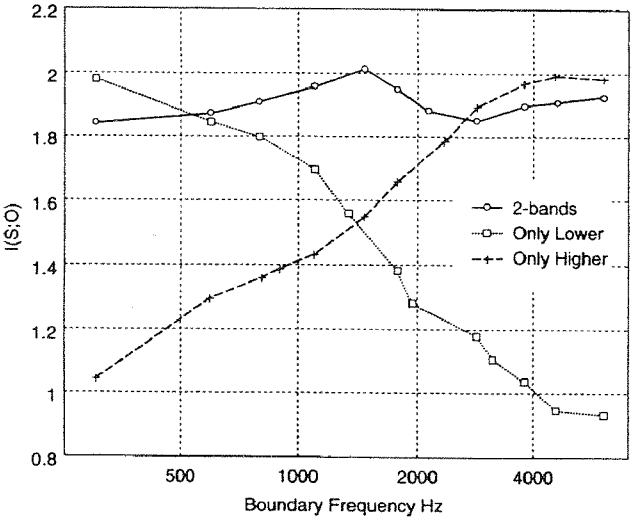


図5 偏帯域システム（low/high）および2帯域システムにおける遮断/分割周波数と相互情報量の関係
Fig. 5 Mutual information for several cut-off/boundary frequencies by partial-band system (low/high) and 2-band system.

変化させた場合，約1,550 Hz付近で聴取率が交叉しているからである（文献10）中でArticulation Indexとして説明）．これは，MB-ASRの根拠を議論するうえでの手がかりともなる重要な結果である．

この結果を得た条件の場合，最大相互情報量を与える周波数が2帯域における最適分割点となる．

6.4 帯域分割の最適化の効果

最後に，様々な雑音付加条件に対して分割点の最適化実験を行う．付加雑音の種類ごとにそれぞれ5.4節で述べた適応化学習を行い，最適な帯域数と分割点を決定する．帯域数としては，先述のとおり1（全帯域），2，および3を用い，あらゆる分割点の組合せの中で最も高い情報量を与える場合を選択した．

図6に，全帯域型（FB-ASR；従来型の音声認識）を基準（0%）とした場合の単語誤り数の改善率を示す．図中で“MB equal bandwidth”は帯域分割点をメル尺度上で等間隔に設定した場合（分割の最適化を行わない図4における最良値に相当）を，“MB optimized bandwidth”は情報量に基づく最適化学習により帯域分割点を決定した場合を示している．

LPWおよびdestroyer-ops, f16, factory1, hfchannel, leopard, pinkの各雑音において，分割点の最適化を施すとさらに性能が向上することが確認できる．また，FB-ASR（従来法）よりかえって性能が劣化してしまうbuccaneer2およびwhiteの両雑音では，最適化において帯域数1（全帯域）での情報量が最大となり，全帯域モデルが選択された．これは，MB-ASRが効果的でない条件においても，少なくとも従来法と同等の性能が得られることを意味している．

図7には，15種類の付加雑音の長時間平均スペクトル（横軸が周波数[kHz]，縦軸は正規化振幅値）と，本手法による最適分割点の位置（太線）をそれぞれ示す．偏帯域性の強い雑音（babble, destroyer-engineなど）において，雑音成分が集中する周波数帯域（の周辺）が良好に分割できていることが確認できる．

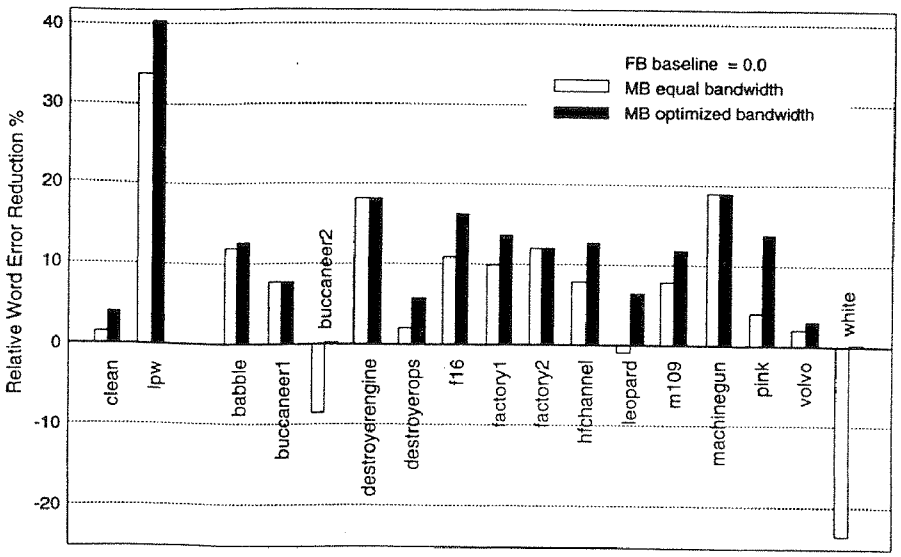


図6 様々な雑音付加音声に対する分割点最適化の効果
Fig. 6 Effect of optimal band-splitting with various additive noise.

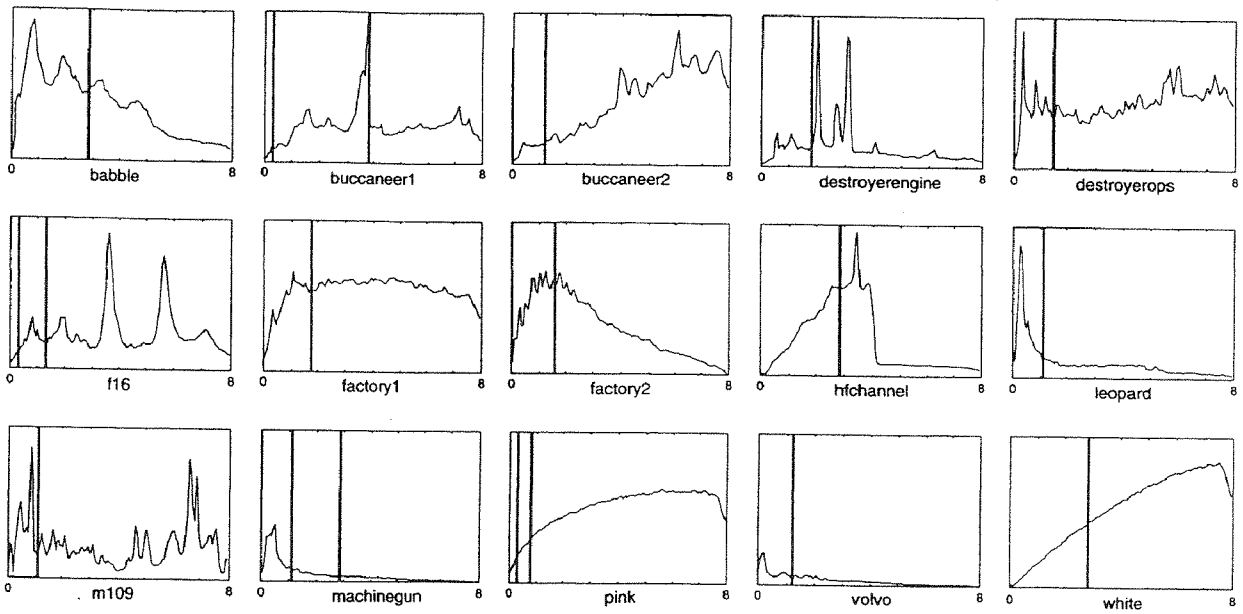


図7 各種付加雑音の長時間スペクトルと最適分割点
(横軸が周波数 [kHz]，縦軸は高域強調後の正規化振幅値，太線が最適分割点)
Fig. 7 Long-term power spectrum and the optimal splitting frequencies for various additive noise.

7. おわりに

本論文では，マルチバンド型音声認識の方法といくつかの実験結果について述べた．認識手法としては，特に次の2つのアプローチを比較検討した．

(1) 尤度レベルの統合法（LC法）：部分周波数帯域からそれぞれ得られた尤度を，HMM状態を単位として統合する．

(2) 特徴量レベルの統合法（FC法）：各帯域に対して音響分析を独立に行い，部分帯域の特徴量ベクトルを連結することにより統合する．

認識実験は，入力音声に対していくつかの異なる種類の雑音を加えて行った．その結果，一般に加えた雑音のエネルギー成分が一部の周波数に集中している場合，マルチバンド型音声認識は従来型の音声認識よりも良い性能を与えることが確認された．また，本論文で提案したFC法は，一般にLC法よりも良い性能を与えることが確認された．

さらに，FC法に基づき，特徴量ベクトルとHMM状態（音素カテゴリ）との相互情報量を評価基準として，最適な分割点を決定する手法を用いた結果，様々な雑音付加条件において，雑音の周波数特性に応じて最も適した分割点を効果的に決定することができ，さらなる認識性能の改善を得た．

今後は，部分帯域特徴量そのものの改善や，統合時における重み付け基準の導入，さらに人間の聴覚機構との関係などの問題について検討を進める予定である．

謝辞 本研究を進めるにあたり，実験環境の提供と有益な助言をいただいた米国 AT&T 研究所の Enrico Bocchieri, Alex Potamianos（現在 Lucent Technologies），David Roe（現在 Computer Motion Inc.）の各氏ならびに早稲田大学理工学総合研究センター音声言語研究室の諸氏に感謝します．

参 考 文 献

1) Junqua, J.-C. and Haton, J.-P.: *Robustness in Automatic Speech Recognition — Fundamentals and Applications*, Kluwer Academic Publishers, Boston (1996).
2) Van Compernolle, D.: Increased Noise Immunity in Large Vocabulary Speech Recognition with the Aid of Spectral Subtraction, *Proc. ICASSP*, pp.1143–1146 (1987).
3) Varga, A. and Moore, R.: Hidden Markov Model Decomposition of Speech and Noise, *Proc. ICASSP*, pp.845–848 (1990).
4) Bourlard, H. and Dupont, S.: A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands, *Proc. ICSLP*, pp.426–429 (1996).
5) Hermansky, H., Tibrewala, S. and Pavel, M.: Towards ASR on Partially Corrupted Speech, *Proc. ICSLP*, pp.1579–1582 (1996).
6) Mirghafori, N. and Morgan, N.: Combining Connectionist Multi-band and Full-band Probability Streams for Speech Recognition of Natural Numbers, *Proc. ICSLP*, pp.743–746 (1998).
7) Okawa, S., Bocchieri, E. and Potamianos, A.: Multi-band Speech Recognition in Noisy Environments, *Proc. ICASSP*, pp.641–644 (1998).
8) Cerisara, C. and Fohr, D.: Multi-band Automatic Speech Recognition, *Computer, Speech & Language*, Vol.15, pp.151–174 (2001).
9) 武田一哉：頑健な音声処理手法，電子情報通信学会技術研究報告，SP2000-75, pp.1–6 (2000).
10) Allen, J.B.: How Do Humans Process and Recognize Speech?, *IEEE Trans. on Speech and Audio Processing*, Vol.2, No.4, pp.567–577 (1994).
11) 河村尊良，武田一哉，板倉文忠：特定帯域不使用モデルを用いた雑音環境下の音声認識，日本音響学会講演論文集，2-Q-16, pp.121–122 (2000).
12) Okawa, S., Nakajima, T. and Shirai, K.: A Recombination Strategy for Multi-band Speech Recognition Based on Mutual Information Criterion, *Proc. European Conf. on Speech Communication and Technology*, pp.603–606 (1999).
13) 金寺 登，荒井隆行，船田哲男：変調スペクトルの重要な成分のみを選択的に用いた雑音に強い音声認識，電子情報通信学会論文誌，Vol.J84-D-II, No.7, pp.1261–1269 (2001).
14) Dahl, D.A., et al.: Expanding the Scope of the ATIS Task: The ATIS-3 Corpus, *ARPA Spoken Language Technology Workshop*, pp.3–8 (1994).
15) Varga, A. and Steeneken, H.M.: Assessment for Automatic Speech Recognition, NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Automatic Speech Recognition Systems, *Speech Communication*, Vol.12, No.3, pp.247–251 (1993).
16) Tibrewala, S. and Hermansky, H.: Sub-band Based Recognition of Noisy Speech, *Proc. ICASSP*, pp.1255–1258 (1997).

(平成 13 年 11 月 14 日受付)
(平成 14 年 4 月 16 日採録)

大川 茂樹（正会員）

1992 年早稲田大学理工学部電気工学科卒業．1996 年同大学院理工学研究科電気工学専攻博士後期課程修了．1996～1997 年米国 AT&T 研究所客員研究員．1998 年千葉工業大学情報ネットワーク学科助教授．博士（工学）．音声言語処理，ヒューマンインタフェース等に関する研究に従事．IEEE, ISCA，電子情報通信学会，日本音響学会，人工知能学会等会員．

白井 克彦（正会員）

1963 年早稲田大学理工学部電気工学科卒業．1968 年同大学院理工学研究科電気工学専攻博士後期課程修了．同年同大学電気工学科専任講師．1975 年同教授．1991 年より同大学情報学科教授．1998 年同大学常任理事・副総長．工学博士．音声言語処理，画像処理，ヒューマンインタフェース，教育用マルチメディアシステム，信号処理用アーキテクチャ設計等に関する研究に従事．IEEE，電子情報通信学会，電気学会，日本音響学会，人工知能学会等会員．

音声対話における心的状態変化の予測をともなうメタ発話生成機構

菊池英明[†,††] 白井克彦[†††]

音声対話制御において，コミュニケーションの調整をはかるメタ的な発話の生成は重要な課題である．音声対話では，音声メディアの特性によって情報の授受とコミュニケーションの調整を同じチャネルで行うために時間のリソースを適切に分配する必要があるが，行為遂行中にも対話の状況や相互の心的状態が変化するためリソースの適切な分配は容易ではない．本研究では，ユーザの目標を協調的に達成するための計画立案の枠組みの中で，対話における相互の心的状態を認識・予測しながら，状況に適したメタ発話を生成する機構を提案する．具体的には，まず対話における状況を表現するための多次元心的状態モデルを定義する．多次元にすることにより各次元の独立性を高め，複雑な制御を可能にするとともにメンテナンス効率を向上させる．この多次元心的状態モデルを用いて認識・予測した心的状態を表す．次に，プラニングにおいて心的状態を参照しながらあらかじめ記述されたプラニングルールに従って目標を達成するために最適な行為を決定する機構を実現する．提案した機構を音声対話検索システムに導入し，評価実験を行った結果，対話の円滑さ向上に効果があることが確認できた．本稿では心的状態として発話権の状態およびシステム内部の処理状態を扱い，状態提示発話を適切に出力することを具体的に示したが，本機構では一般的な心的状態表現およびメタ発話行為を扱っている．

Mechanism of Generating Meta-utterance with Predicting Changes of Mental States in Spoken Dialogue

HIDEAKI KIKUCHI[†,††] and KATSUHIKO SHIRAI[†††]

In controlling spoken dialogues, it is important to generate meta-utterances which coordinate communication. To realize smooth conversation between human and machine, time resources should be appropriately divided into speakers because exchanges for problem-solving and coordinating communication are done in same channel. This is a feature of the speech dialogue medium. This research will propose the mechanism of generating meta-utterances with predicting changes of mental states in spoken dialogues. At first, we defined the model of multi-dimensional mental states for expressing situation in dialogues. By multi-dimensionality, independency of each dimension of states and efficiency of maintenance of them can be improved. Mental states, which are recognized and predicted, will be expressed by using this model. At next, we described the mechanism, which determine an appropriate action to achieve a goal by planning with the planning rules. These rules, in advance, are described with referring each dimension of mental states. By the preliminary experiments, we confirmed that the proposed model is efficient for improvement of smoothness of conversation and efficiency of problem-solving.

1. はじめに

音声による対話は人間同士の通常のコミュニケーション形態の1つであることから，音声対話処理技術は将来のマンマシンインタフェース技術の1つとして期待される．マンマシンインタフェースのメディアとしての音声対話では，ユーザが自らの思考を妨げられることなく発話できることが望まれ，そのためには話題展開や発話タイミングなどの自由を保証する必要がある[1]．しかし，ユーザの行為に制約を与えなければ，その結果として問題解決が阻害されたり発話が重複してしまったりなど，対話の進行が滞るケースが増える．このような場合に，たとえば，問題解決の状況を説明したり，再発話を要求したり，発話の番を譲ったりするなどの，コミュニケーションの調整をはかるシステムのメタ的な行為が対話の円滑な進行に役立つ[2),3)]．

† 国立国語研究所
The National Institute for Japanese Language
†† 早稲田大学人間科学部
School of Human Sciences, Waseda University
††† 早稲田大学理工学部
School of Science and Engineering, Waseda University

Batesonはコミュニケーションにおける参加者間のやりとりのうち，こうしたコミュニケーションの調整をはかるレベルのメッセージをメタメッセージと呼び，メッセージとメタメッセージの二重性はコミュニケーションの場につねに存在するとしている[4]．

ここで，音声対話の形態は対話様式や目的などによってさまざまであるが，本研究では，音声対話処理技術の実現の容易さと市場の広さから電話会話のような音声のみによる目的指向対話に焦点をあてる．目的指向対話においてユーザの行為に制約を与えず，なおかつ円滑な進行をはかるには，目的を達成するための情報授受のやりとりを中心として，先に述べたコミュニケーションの調整をはかるメタ的な行為をシステムが適宜行う必要がある．たとえば画面を用いたユーザインタフェースでは，メタ的行為の1つである状態提示行為をメッセージの画面表示などで遂行できる．しかしながら，音声のみによる対話においては，単一のチャネルしか持たないメディアとしての特性により，メタ的行為を発話によって行わなければならない．その結果，情報の授受とコミュニケーションの調整を同じチャネルで行うために時間のリソースを適切に分配しなければならないという問題が生じる（単一チャネルにおける時間リソース分配の問題）．これには，発話行為の遂行に若干の時間を要するという音声対話メディアの別の特性も関わっている．これまでに我々は問題解決行為の合間を埋めるような状態提示発話出力が，全体の発話数や発話時間を増加させても効率評価の向上に有効であることを確認した[5]．本研究ではこの成果をさらに一般化し，目的指向対話におけるさまざまな状況下でメタ発話を適切に出力する機構を構築することを目的とする．

従来，対話参加者の真の目標を協調的に達成するためのプラニングの枠組みがさまざまに提案されており[6]，それらの多くは談話を階層的に構造化する中で発話行為の1つとしてメタ発話を扱う．コミュニケーションの調整をはかるメタ的発話行為を扱うために，行為レベルでの状況の変化に対応して再プラニングを行う従来のモデルを拡張し，発話権移行レベルの頻繁な状況変化に対応してプラニングを行うモデルも提案されている[7)~9]．ここで，音声対話でのメタ発話生成においては，先に述べたとおり単一チャネルにおける時間リソース分配の問題がある．発話権の所在が固定されている（あらかじめ時間リソースが対話参加者に分配されている）対話様式であれば，プラニングによって決定されたメタ発話を通常のやりとりの中に埋め込むことが容易である．しかしながら，発話権の移動が自由な様式の場合には，発話権の所在や内外の状況の変化を予測しなければ通常のやりとりを阻害しないようにメタ発話を出力することが難しい．

本稿では，音声対話における状況変化の予測に応じてプラニングにより状況に適したメタ発話生成を行う機構を提案する．具体的には，対話における状況を複数の独立した次元の心的状態として表現し，その変化の予測をプラニングにあわせて行うことによってより状況に適した行為の決定を可能にする．以下，2章にメタ発話の例とその中での状況変化について示し，提案するメタ発話生成機構で扱う多次元心的状態モデルを定義する．3章ではメタ発話生成機構について説明し，実際の音声対話システムに導入して評価を行う．4章では提案機構の一般性を議論する．

2. メタ発話と多次元心的状態モデル

ここでは対話における状況を表現する心的状態モデルについて定義する．人間同士の会話における発話の機能を社会学的にとらえようとする際，複数の側面について考察する手法が用いられる[10]．また，人間の思考そのものが複数の心の競合もしくは協調によるものとする考え方がある[11),12]．本稿で扱う「メタ発話」は，先に述べたBatesonのメタメッセージを発話に特化したものであり，すなわちコミュニケーションの過程におけるさまざまな思考に関わるものである．したがって，メタ発話を含む発話行為のプラニングにおいて対話の状況をとらえるモデルとして多次元性を持った心的状態のモデルを用いることにする．

以下では，音声対話におけるメタ発話の具体例を示したうえで，本研究で扱う多次元心的状態モデルを定義する．なお，本稿での議論は，以降，学術論文検索をタスクとした目的指向対話を例として行う．

2.1 メタ発話と心的状態変化

図1に論文検索対話の一例を示す．この例において，メタ発話は太字で示したS4，S5，S6であり，いずれもシステムの処理状態を提示する役割を持つ．このうち，S5とS6を含むやりとりにおいてシステムがとらえるべき心的状態の変化を図2に示す．

図2において，心的状態としては仮に，ユーザ・システムそれぞれから見た発話権の状態（発話権がどちらにあるか）とシステム内部の処理状態を示した．

この図は，図1で示したようにユーザの検索要求にシステムが応じて検索処理を行っている間のやりとりであり，このやりとりが始まるまでの発話権はシステム側にあることを，ユーザ・システムともに共有信念として持っていると仮定する．また，ユーザ発話U4

S1: こんにちは
U1: うん
U2: 細菌学の論文なんだけど
S2: はい
U3: 炭疽菌の治療法についてあるかな
S3: 炭疽菌と治療法で検索します
S4: しばらく時間がかかるのでお待ち下さい
<しばらく待ち状態>
U4: あのー
S5: もうすぐ検索が終わります
<若干の待ち状態>
S6: 検索が終わりました
S7: 16件の論文が見つかりました
S8: これでいいですか
U5: はいありがとう

図1 論文検索対話におけるメタ発話の例
(Sはシステム，Uはユーザの発話，太字はメタ発話)
Fig. 1 Example of meta-utterance in dialogue for academic paper retrieval.
(S: system, U: user, Gothic style: meta-utterance)

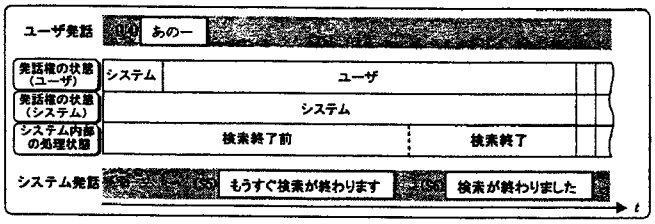


図2 検索終了前後の状態提示発話
Fig. 2 Utterance for presenting states.

は検索処理の終了を待てずに検索条件を変更しようとするものと仮定する．

まず，ユーザ発話U4を発した時点で，ユーザにとっては発話権を獲得しようとしているため，ユーザ側からみた発話権の状態は「ユーザ」に移ったといえる．一方，システム内部の処理状態は図の点線で示す位置が「検索終了」のタイミングとして予測されるとしたとき，U4開始の時点では，システム内部の処理状態が「検索終了前」と予測されている．したがって，ユーザの大局的な目標達成のためには発話権を譲ることが得策でないと判断し，代わりとして発話権獲得を撤回させるための状態提示発話S5を生成する．なお，この間，システム側からみた発話権の状態は一貫して「システム」のままである．「検索終了前」であることを示す状態提示発話S5は，煩わしさを抑えるために通常は行わないものであるとしたとき，この状況下に限ってこの発話を行う制御を実現するには，ユーザ・システムそれぞれからみた発話権の状態およびシステム内部の処理状態を認識および予測しておく必要があるといえる．なお次に，実際に検索処理が終了した時点で(図2では予測と同じタイミングで検索処理が終了したとしている)，その旨を伝える発話S6を生成する．

ここでは検索処理終了の予測を例として説明したが，こうした内部状態のほかにも発話権状態，さらにはユーザの注意，積極性などの状態も心的状態の次元の1つとして同様に扱い，認識および予測を行うことは状況にあったメタ発話の生成にとって有効である．

2.2 多次元心的状態モデルの定義

前節に示したように，心的状態は複数の次元に存在しうる．対話制御機構において心的状態を多次元的に扱うことによって，各状態が相互に作用しあうような複雑な制御が可能になるとともに，モデルとしてのメンテナンスが容易になるという利点がある[13]．本研究で提案するメタ発話生成機構では，複数の次元の心的状態を独立にモデル化し，プラニング時に利用する．以下には，まず多次元心的状態モデルの一般的な定義を行う．

まずはじめに目的指向対話を協調的問題解決の手段と位置づけ，プラニングについて以下の前提をもうける．なおここで示すプラニングは具体的な行為を決定するのではなく，プランを複数の行為の集合と見なし，状況に応じた適切なプランを決定することを目的とする．

前提1：問題解決のゴールは複数のサブゴールに分割され，その構造はあらかじめ規定されない．

$G = \{G1, G2, \cdots, Gn\}$

G：問題解決のゴール
(ex. “会議室予約”,“論文検索”,“チケット予約”,$\cdots$)
Gi：ゴール Gのサブゴール
(ex. G=“論文検索” の場合，$G1$=“検索条件作成”，$G2$=“検索結果提示”,$\cdots$)

前提2：サブゴールは1つ以上のプランの実行により達成される．

$P_i = \{Pi1, Pi2, \cdots, Pim\}$

Pi：問題解決のサブゴール Gi 達成のためのプラン群
Pij：サブゴール Gi を達成するためのプランの1つ

これらの前提1，2から，1つのゴールに対してサブゴールの系列は複数あり，さらに1つのサブゴールに対してプラン系列も複数ありうるといえる．たとえば，G=“論文検索” の場合，サブゴール系列は [“検索

条件作成”,“検索結果提示”,“検索条件修正”,“検索結果提示”…] のように，検索条件をあまり吟味せずに結果を取得してから判断する方法と，[“検索条件作成”,“検索条件修正”,“検索結果提示”,…] のように，検索条件をよく吟味してから検索を行う方法がありうる．前提1は，このようなサブゴール系列を問題解決の前に決定するのではなく，変化する状況に応じて修正可能であることを示す．同様に前提2より，あるサブゴールを達成するためのプラン系列も状況に応じて修正可能とすることにより，1つのゴールに対するサブゴールおよびプラン系列は状況に応じて柔軟に修正することが可能となる．具体的なサブゴール系列およびプラン系列の構成方法については，3.2節の動作例の説明で述べる．

以下に，複数次元の心的状態モデルを定義する．

$$S = \begin{pmatrix} S_1 \\ S_2 \\ \cdots \\ S_n \end{pmatrix}$$

$$Sp = Sp1, Sp2, \cdots, Sps$$

S：多次元心的状態の集合

Sp：p次元目の心的状態の集合

(ex. “発話権の状態”，“処理状態”，“情報授受の状態”,…)

Spq：p次元目の心的状態の1つ

(ex. Sp=“発話権の状態” の場合，$Sp1$=“ON”, $Sp2$=“OFF”)

なお，各状態の認識方法はその種類によってさまざまであるが，他の次元に対して依存関係を持つべきではない．つまり，何らかの障害により1つの次元の状態が認識不能となっても対話を続けることができるようにすべきである．

先に述べたように，状況に適したメタ発話の生成には，心的状態変化の予測が必要になる．ここまでの説明で示したように，心的状態を離散的な状態系列として表現するために，予測可能な状態変化の過程を離散的な状態の1つとして加える方法（ex. Sp=“発話権の状態” の場合，発話権譲渡を予測して Spq=“OFF直前” など）を用いる．

以上から，複数の行為の集合であるプランはゴールGと，それに対応したプラン群Pおよび，多次元心的状態Sを用いたプラニングによって決定されるといえる．次章では，ここで示した心的状態モデルを導入したメタ発話生成機構を説明する．

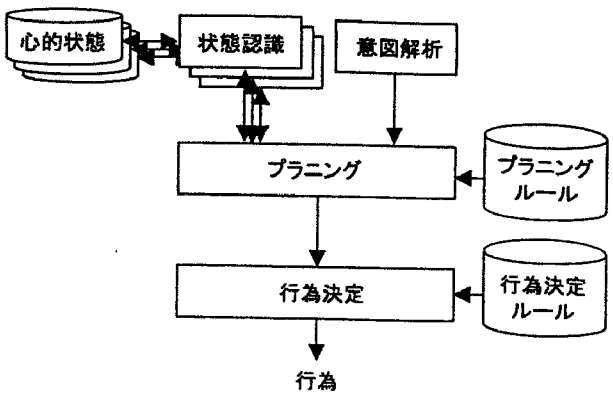


図3 メタ発話生成機構
Fig. 3 Mechanism of generating meta-utterance.

3. メタ発話生成機構

本章では，前章で示した多次元心的状態モデルを導入し，対話参加者の多次元の心的状態予測をともなうプラニングによって，状況に適したメタ発話を生成する機構を説明する．

3.1 構成

提案するメタ発話生成機構の構成を図3に示す．図3の各部の処理について以下に述べる．

(1) 意図解析 ユーザ発話についての表層的な意図を解析する．ユーザ発話の音声終端検出を契機に処理が開始される．

(2) 状態認識 次元ごとに状態の変化を認識・予測する．具体的には，ユーザ発話の音声始終端検出やデータベース検索処理終了など，心的状態変化に関わるイベントをつねに監視して，心的状態の値をリアルタイムに更新する．さらに韻律パターン照合によってユーザ発話の音声終了を予測したり，後述するように内部処理終了を予測したりするなどによって，状態変化の予測を行う．ここで状態変化の予測を行う点が本機構の特徴である．状態変化が検出された場合にはプラニング部に信号を送る．

(3) プラニング あらかじめ記述されたプラニングルールを用いて，意図解析処理の結果あるいは状態認識による状態変化検出信号が得られたときにプラニング処理を開始する．プラニングの結果として，行為群に相当するプランが決定される．

(4) 行為決定 プラニング処理によるプラン決定を受けて，対応する行為決定ルールを用いて，具体的な行為を決定する．

3.2 動作例

以下には，2章の図2で示したやりとりを実現するためのプラニングルールとそれを用いた処理の説明を行う．心的状態としては，ユーザ・システムそれぞれ

(Rule1)SubGoal("検索条件作成")
&& *Act*(Process("データベース検索"))
&& *Act*(Plan("検索結果提示"))
⇒ SubGoal("検索結果提示")
(Rule2)Ready_SubGoal("検索結果提示")
&& [$S_{process}$== "検索終了"]
&& *Act*(Plan("検索結果提示"))
⇒ SubGoal("検索結果提示")
(Rule3)[$S_{process}$== "検索終了前"]
&& Change_State(S_{usr_turn},"ON", "OFF")
⇒ Ready_SubGoal("検索結果提示")
(Rule4) [S_{usr_turn} == "ON"]
&& [S_{sys_turn}== "ON"]
&& [$S_{process}$== "検索終了前"]
&& *Act*(Plan("状態提示"))
⇒ Change_State(S_{usr_turn},"ON", "OFF")

図4 心的状態を参照するプラニングルールの例
Fig. 4 Example of planning rules referring mental states.

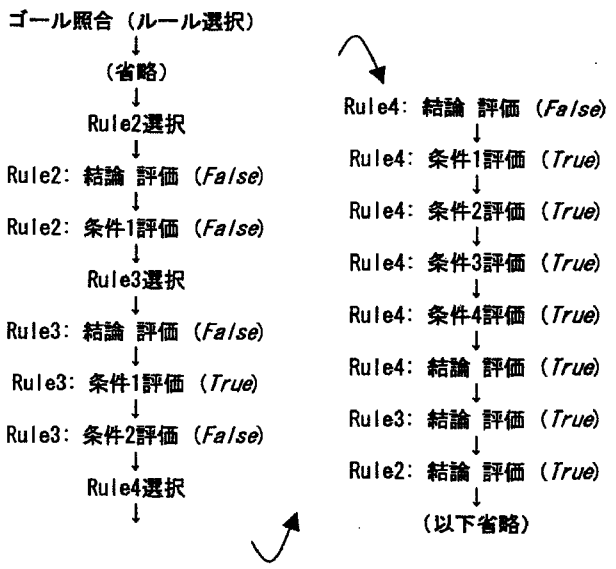


図5 プラニング過程の例
Fig. 5 Example of planning process.

の発話権の状態を S_{usr_turn}, S_{sys_turn} とし，システム内部の処理状態を $S_{process}$ として導入する．発話権の状態は値として "ON", "OFF" をとる．処理状態は "検索終了", "検索終了前" など，あらかじめ定義された内部処理の離散的な状態を値としてとる．

図4に，状態提示発話 S5 を生成するのに必要なプラニングルールを示す．"⇒" の前部が条件部，後部が結論部であり，"&&" は AND を示す．各節は命題を示すが，斜体で示した Act() は手続きをともなう命題である．プラニング処理ではあらかじめこれらの標準的な命題に対する振舞いが定義されている．なお，ルールは後ろ向き推論を仮定しており，まずすべてのルールの結論部が評価され，意図解析により得られたユーザのゴールと一致するルールが選択される．次に，選択されたルールの条件部のうち最初の条件から順に評価される．評価された命題が真でない場合にはその命題を結論部に持つルールが選択される．条件部の命題がすべて真となった場合に結論部の命題が真となる．なお，評価されていない命題については便宜上，偽と同様に扱う．

図4において，Rule1 は通常の問題解決に関するやりとりを遂行するためのルールであり，サブゴール "検索結果提示" を達成するためには，サブゴール "検索条件作成" の達成と，処理 "データベース検索" の実行が必要であることを示している．それに対して Rule2 から Rule4 は，処理 "データベース検索" の終了待ち状態のユーザ割込みに対して状況に応じてメタ発話を出力するためのルールである．Rule2 は，サブゴール "検索結果提示" を達成するためには，Ready_SubGoal ("検索結果提示") を達成して，なおかつシステム内部の処理状態 $S_{process}$ が "検索終了" となる必要があることを示す．なお，Ready_SubGoal() はサブゴール達成前の状態に相当する．Rule3 は，Ready_SubGoal ("検索結果提示") を達成するためには，システム内部の処理状態 $S_{process}$ が "検索終了前" であり，なおかつユーザの発話権の状態 S_{usr_turn} を "ON" から "OFF" にする必要があることを示す．なお，Change_State() は指定された状態の値を変更するという命題である．Rule4 は，ユーザの発話権の状態 S_{usr_turn} を "ON" から "OFF" にするには，ユーザとシステムの発話権の状態がともに "ON" であり，システム内部の処理状態 $S_{process}$ が "検索終了前" であり，なおかつ Plan ("状態提示") すなわち状態提示プランが遂行される必要があることを示す．

なお，システム内部の処理状態 $S_{process}$ が "検索終了前" となるのは，状態認識処理における状態変化予測で検索終了時期を推定した結果である．処理の終了時期の推定は処理時間の統計量に基づいて行う[5]ことを想定している．また，"検索終了前" 状態の始端位置は，その区間に状態提示発話を生成する可能性があるのだから少なくとも発話時間長の分を確保するように決定する必要がある．

以下には図5を用いて，具体的なプラニング過程を示す．ユーザ発話の意図解析結果からゴールを "特定論文検索" と判断してプラニングによりそのサブゴールとして "検索結果提示" が設定されたとする．次に，Rule1 と Rule2 により，サブゴール "検索結果提示" を達成するための，2つのプラン系列が候補として生成される．すでに Rule1 により生成されたプラン系列

表1 メタ発話生成機構の評価結果
((S) はシステム発話，(U) は被験者発話を計測の対象としていることを示す)
Table 1 Evaluation result of proposed model.
((S) indicates that the item was counted in system's utterances, and (U) indicates subjects' utterances)

メタ発話制御		総発話 (U)	検索要求 (U)	メタ発話 (S)	重複発話 (割合)	円滑さ
(1) なし		47	31	0	2 (4.2%)	2.2
あり	(2) 常時出力	92	49	38	11 (11.9%)	3.1
	(3) 本手法	62	42	35	3 (4.8%)	3.8
計		201	122	73	16 (7.9%)	3.0

が選択および遂行され，Rule1 に条件として記述された Act が遂行されていても，それが完了しない限りは Rule2 から生成されたプラン系列が選択されうる．プラン系列の選択過程は図5に示したように，Rule2 選択の後，結論部が評価され False であるため Rule2 の条件1の命題 "Ready_SubGoal" が評価される．これが False となるためこれを結論部に持つルール Rule3 が選択される．同様にしてこのルールの条件を満たすようにルールを選択した結果，[S_{usr_turn}=="ON"]，[S_{sys_turn}=="ON"]，[$S_{process}$=="検索終了前] の状況でプラン "状態提示" を遂行することによって，結果としてサブゴール "検索結果提示" が満たされることになる．このようにして，論文検索処理実行中にユーザからの発話があった場合，Rule2 から Rule4 で決定されるプラン系列の1つとして状態提示プランを遂行できることになる．なお，ここで示したルールは，論文検索処理終了をユーザが待ちきれずに発話した場合のメタ発話出力を実現するものである．ただしあいづちや無意味語などに対しては持続時間の短さから発話権状態更新を行わないためメタ発話は出力されない．また，ゴールを変更するような発話に対しては発話意図が解析されてプラニングにより新たなサブゴール系列・プラン系列が決定されることになる．

行為決定処理で用いる行為決定ルールについてはイベントをアーク，行為をノードとする有限状態オートマトンによる記述[14]を想定しているが，本稿の範囲では，行為としてメタ発話を対象としており，プランと行為の関係がほぼ一定であるため，具体的な行為そのものの形式は特に言及しない．

3.3 実用性評価

3.1 節で示したメタ発話生成機構と図4のルールを含む17のプラニングルールを音声対話システム汎用プラットフォーム[14]に導入してその実用性評価を行った．具体的には，12名の被験者にあらかじめ3種類の文献を与え，それぞれの理解に必要な論文をシステムとの音声対話によって検索するよう要求した．なお，被験者のうち4名のみが音声インタフェースの使用経験があり，念のために全被験者に数回の練習を実験前に行わせシステムの使い方に慣れさせた．心的状態としては発話権状態とシステム内部の処理状態を導入し，プラニングルールにおいてはシステム内部の処理状態を提示するメタ発話を生成するためのルールを加えた．システム内部の処理としては，音声認識，意図解析，データベース検索を対象とした．

各被験者が行う3対話において，各対話によってメタ発話制御方法を，(1) なし，(2) 常時出力，(3) 提案手法と変え（順不同），それぞれの対話における被験者の総発話数，検索要求数，重複発話数とシステムのメタ発話数を調べた．なお，(2) の常時出力の場合は提案手法によって出力する可能性のあるすべてのタイミングでメタ発話を出力した．また，実験後のアンケートにより対話の円滑さを被験者に5段階で評価させた．以上の結果を表1にまとめる．

表において，円滑さ評価の平均がメタ発話出力なしの場合の 2.2 から常時出力あるいは本手法による出力によって向上している（t 検定，$p < 0.01$）．この結果から，メタ発話の出力が円滑さの評価向上に有効であるといえる．なお，表においてメタ発話出力によってユーザ発話数および検索要求数が大幅に増加している．状態提示メタ発話により無駄なやりとりを回避するという当初予想した効果は認められないが，メタ発話によって対話が円滑になった結果，対話によって目的を達成しようとするやりとりが増えたものと考えられる．

また，メタ発話を常時出力する場合に比べ，本手法を適用して問題解決の進行を考慮することによって検定による有意差はないものの（$\chi_2(1) = 0.13$）重複発話の数が大幅に減った．さらに，発話数や検索要求数が減っているが，収録対話の観察から，重複から復帰するための発話や重複により遂行されない検索要求の数が減ったことによると考えられる．

以上から，音声対話におけるメタ発話生成において，メタ発話の出力によって対話を円滑にし，さらにメタ発話出力方法として本手法を適用することで，つねにメタ発話を出力するよりも効率良く問題解決を進めることが可能になることを確認した．

4. 一般性についての考察

ここでは，これまでに述べたメタ発話生成機構についていくつかの側面から一般性を考察する．

4.1 心的状態

本稿ではユーザ・システムそれぞれの発話権の状態と，システム内部の処理状態を心的状態としてとりあげた．しかし，2章で示した多次元心的状態モデルは特にこれらを規定するものではない．

なお，心的状態の値としては離散的な値だけでなく連続量をとりうる．たとえば，認識した音声の尤度を，受理した発話の信頼度についての心的状態として用いるなどが考えられる．ただし，心的状態の値が連続量である場合，プラニングにおいては状態の値に対して比較演算やマッチング処理などを可能にする必要がある．

4.2 心的状態変化の予測

本稿では，システム内部の処理状態のみを予測の対象としてとりあげたが，発話権の状態についても文脈やユーザモデルによってある程度予測することができると考えられる．

なお，状態変化の予測の難しさは，対象とする心的状態の種類によって異なる．現時点では，状態遷移のパターンや状態滞在時間の分布がある程度既知である場合を想定している．

4.3 メタ発話

プラニングにおいて，心的状態を参照して決定するプランとしては状態提示のみを示したが，他のメタ行為についても同様にプラニングの対象とすることができる．たとえば，前述した発話の信頼度についての心的状態を導入することで，再発話要求を促すプランを用意することも可能になる．また，発話権の状態変化を予測することで，ユーザ発話の合間にあいづちのような調整的行為を遂行することも可能になる．

4.4 プラニング

状態認識処理を常時行い，状態変化に応じて再プラニングを行う本機構により，プラン遂行中に，前提としていた状態の値が変化してしまうような場合，その状態に適した再プラニングが実現できる．これによって，対話における頻繁な状況の変化に対応可能になったといえる．

また本稿では，発話生成機構をプラニングと行為決定に分けて，それぞれ異なるルールを参照するようにした．これは，状態変化に応じてプラニングを行うことで，処理の頻度が著しく増加することを考慮して，プラニングの処理を減らすために反射的な制御を行為決定処理にまかなわせたためである．状態変化に応じてプラニングを行った結果，プランの変更がなく行為決定を必要としないケースは非常に多いため，このように処理を分けることは効率を大幅に向上させる．しかしながら，2種類の独立したルールが同じ心的状態モデルを共有するのは，メンテナンス上の効率を下げる．メンテナンス効率を向上するルール記述形式あるいはアーキテクチャを検討する余地があるといえる．

また，本稿では電話会話のような音声のみによるコミュニケーションの形態を想定して議論してきたが，音声以外のメディアを利用できる環境においては発話の代替となる手段を用いることができる．プラニングにおけるこうした環境変化の考慮は今後の課題といえる．

4.5 システムの汎用性

本研究は目的指向対話におけるさまざまな状況下でメタ発話を適切に出力する機構の構築を目指している．目的指向対話のようにある程度対話の流れが定まっている対話においては，本機構のルールベースの手法は有効である．すなわち，そのような対話においては通常のやりとりを構造化したうえでコミュニケーションの調整行為を位置づけられる．しかしながら，対話の流れの予測がつかない場合，通常のやりとりを構造化できないためにメタレベルのやりとりを位置づけることは困難である．その場合にはむしろプラニングとは独立した枠組みとすることが有効と考える．

5. おわりに

本研究では，ユーザの目標を協調的に達成するためのプラニングの枠組みの中で，対話における相互の心的状態を認識・予測しながら，状況に適したメタ発話を生成する機構を提案した．具体的には，多次元の心的状態モデルを定義し，その状態変化をつねに予測する機能を設け，あらかじめ状態の予測結果を参照するよう記述したルールをプラニングにおいて扱うことを可能にした．提案した機構を音声対話検索システムに導入し，評価実験を行った結果，対話の円滑さと効率の向上に効果があることが確認できた．ここで示した機構は，先の研究[5]における状態提示発話の制御モデルを，一般的なメタ発話制御に拡張したものである．考察の結果，心的状態の定義およびメタ発話の一般性はあるが，状態変化の予測およびプラニング処理についてはいくつか課題が残るといえる．

参考文献

1) 菊池英明，工藤育男，小林哲則，白井克彦：音声

対話インタフェースにおける発話権管理による割込みへの対処，電子情報通信学会論文誌，Vol.J77-D-II, No.8, pp.1502–1511 (1994).
2) DISCプロジェクト報告，http:www.disc2.dkslds など
3) Bell, L. and Gustafson, J.: Interactive Dialogue in Multi-Modal Systems, *Proc. ESCA workshop on Interactive Dialogue in Multi-Modal Systems* (*IDS'99*), Kloster Irsee, Germany (1999).
4) Bateson, G.: *Steps to an ecology of mind*, Chicago University Press (1972).
5) 菊池英明，白井克彦：対話効率の向上を目的とした音声対話制御のモデル化，ヒューマンインタフェース学会誌，Vol.2, No.2, pp.145–152 (2000).
6) Allen, J.F., Hendler, J. and Tate, A. (Eds): *Readings in planning*, Morgan Kaufmann (1990).
7) Cohen, P.R.: The pragmatics of referring and the modality of communication, *Computational Linguistics*, Vol.10, No.2, pp.97–146 (1984).
8) Clark, H.H. and Wilkes-Gibbs, D.: Referring as a collaborative process, *Cognition*, No.22, pp.1–39 (1986).
9) Novick, D.: Controlling interaction with meta-acts, *Conference on Human Factors in Computing Systems* (*CHI'91*), New Orleans, USA (1991).
10) Hymes, D.: Models of the interaction of language and social life, *Directions in Sociolinguistics*, Gumperz, J. & Hymes, D. (Eds.), New York, Holt, Reinhart & Winston (1972).
11) M. ミンスキー：心の社会，産業図書 (1985).
12) 山鳥　重：脳からみた心，NHK ブックス 482，日本放送協会 (1985).
13) 鈴木堅悟，青山一美，菊池英明，白井克彦：多次元心的状態を扱う音声対話システムの構築，情報処理学会研究報告，2001-SLP-37, pp.13–18 (2001).
14) 青山一美，平野　泉，菊池英明，白井克彦：音声対話システム汎用プラットフォームの検討，情報処理学会研究報告，2000-SLP-30, pp.7–12 (2000).

(平成 13 年 11 月 19 日受付)
(平成 14 年 4 月 16 日採録)

菊池　英明（正会員）

1991 年早稲田大学理工学部電気工学科卒業．1993 年同大学大学院修士課程修了．同年（株）日立製作所入社．1997 年同退社．同年早稲田大学大学院後期博士課程入学．2000 年同単位取得退学．同年より国立国語研究所特別奨励研究員．2002 年早稲田大学人間科学部非常勤講師．ヒューマンインタフェース，音声対話の研究に従事．情報処理学会平成 8 年後期大会優秀賞受賞．人工知能学会，電子情報通信学会，日本音響学会各会員．

白井　克彦（正会員）

1963 年早稲田大学理工学部電気工学科卒業．1968 年同大学大学院博士課程修了．工学博士．同年同大学理工学部講師．1970 年同助教授．1975 年同教授（電気工学科）．1991 年同教授（情報学科）．音声認識・合成技術，自然言語処理，CAI 等を中心にヒューマンインタフェースの研究に従事．1998 年人工知能学会会長．日本音響学会，IEEE 等各会員．電子情報通信学会フェロー．

International Journal of Knowledge-Based Intelligent Engineering Systems, Vol. 6, No. 4, October 2002

Human Face Extraction using Genetic Algorithm with Similarity of Subspace Method as a Fitness Value

Makoto Murakami, Katsuhiko Shirai
School of Science and Engineering,
Waseda University
3-4-1 Okubo, Shinjuku
Tokyo 169-8555, Japan
murakami@shirai.info.waseda.ac.jp

Masahide Yoneyama
Dept. of Information and
Computer Sciences, Toyo University
2100 Kujirai, Kawagoe
Saitama 350, Japan
yoneyama@cs.toyo.ac.jp

Abstract

The purpose of this paper is to extract the human face from the images taken on the various conditions and detect the facial pose using subspace method and Genetic Algorithm. Subspace method can express facial images with various pose by linear translation into lower dimensional subspace, but it is difficult to put into practical use because it comes under the influence of background image or illumination condition. At first, we consider the feature vector to be used for construction of subspace and the construction method. Then, based on the results and considerations we extract human face from the images with many kind of background actually. In the extraction process, we use Genetic Algorithm that can be effective for the discontinuous search space. And we report the effective results were obtained actually.

Key Word: Face Recognition, Genetic Algorithm, Subspace Method

1. Introduction

Lately multimodal interface has been noticed. It is expected as the latest interface that will take the place of the interface using keyboard or mouse. Most of us would accept that human face has important information. So a considerable number of studies ever have been made on expression recognition, individual distinction and eyes recognition etc.[1],[5],[11] In these studies the accurate extraction method of human face is required.

In this paper, at the first step of the recognition of human face's motion we are concerned with human face extraction from images with various backgrounds and the pose detection. A lot of studies [4],[10] on this field ever have been done, but most of them come under the influence of background image or illumination condition, so it is difficult to put into practical use. The purpose of this paper is to recognize human face accurately under the various conditions.

In the conventional human face extraction methods by Genetic Algorithm require a high computational effort because they deal with raw image data. So we decide the facial area and detect the pose using subspace method [3],[7],[9] that can express facial images efficiently by linear translation into lower dimensional subspace. To improve the recognition accuracy in the real world, it is necessary to construct the subspace in answer to the purpose or the recognition object. In this paper we consider the feature vector to be used for construction of subspace and the construction method by experiment. Out purpose for facial pose detection is to recognize independent of individual and for face extraction is to improve the recognition accuracy reducing the influence of background or illumination condition. Then in the search of the face area, we use Genetic Algorithm [2],[6],[8] that can be effective for the discontinuous search space, and show that the effective results were obtained actually.

2 Recognition Method

2.1 Subspace Method

In this section, we describe the subspace method (Parametric Eigenspace Method [7] and CLAFIC Method [4],[10]) that is one of the object recognition method.

2.1.1 Construction of Subspace

Let N features extracted from one image show as the following N-dimensional vector x:

$$x = [a_1, a_2, \cdots, a_N]^T \tag{1}$$

Here, a vector x is normalized to $|x| = 1$.
Next, let a set of vectors which is classified as category i show as

$$\{x_1{}^{(i)}, x_2{}^{(i)}, \cdots, x_R{}^{(i)}\} \tag{2}$$

Here, R is the step number of rotation.

So we calculate orthonormal bases in the subspace of the set of vectors.
From the average vector $c^{(i)}$:

$$c^{(i)} = \frac{1}{R}\sum_{r=1}^{R} x_r{}^{(i)} \tag{3}$$

matrix $X_{cov}^{(i)}$

$$X_{cov}^{(i)} \equiv [x_1{}^{(i)} - c^{(i)}, \cdots, x_R{}^{(i)} - c^{(i)}] \tag{4}$$

and covariance matrix $Q_{cov}^{(i)}$

$$Q_{cov}^{(i)} \equiv X_{cov}^{(i)} X_{cov}^{(i)\,T} \tag{5}$$

are calculated.
In the same way from matrix $X_{cor}^{(i)}$

$$X_{cor}^{(i)} \equiv [x_1{}^{(i)}, \cdots, x_R{}^{(i)}] \tag{6}$$

correlation matrix $Q_{cor}^{(i)}$

$$Q_{cor}^{(i)} \equiv X_{cor}^{(i)} X_{cor}^{(i)\,T} \tag{7}$$

is calculated.

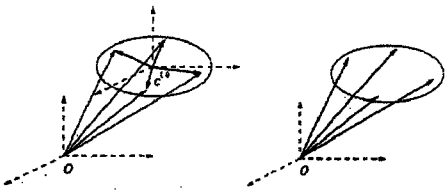
Then, we solve the following characteristic equation:

$$\lambda_j{}^{(i)} e_j{}^{(i)} = Q^{(i)} e_j{}^{(i)} \tag{8}$$

Eigenvalues $\lambda_j{}^{(i)}$, eigenvectors $e_j{}^{(i)}$ are calculated.
From k eigenvectors as follows:

$$(e_1{}^{(i)}, \cdots, e_k{}^{(i)}), (\lambda_1{}^{(i)} \geq \cdots \geq \lambda_k{}^{(i)} \cdots \geq \lambda_N{}^{(i)}) \tag{9}$$

which are orthonormal bases, k-dimensional subspace of category i is constructed.



Covariance Matrix Correlation Matrix

Figure 1. Subspace

2.1.2 Parametric Eigenspace Method

1. Projection into Subspace
   We project the feature vectors extracted from the images classified as category i into the subspace of category i as follows

$$p_r{}^{(i)} = [e_1{}^{(i)}, e_2{}^{(i)}, \cdots, e_k{}^{(i)}]^T (x_r{}^{(i)} - c^{(i)}) \tag{10}$$

   Here, $p_r{}^{(i)}$ are points in the subspace.

2. Pose Detection
   A vector y extracted from an input image that is normalized $|y = 1|$ is projected into every subspace as follows

$$z^{(i)} = [e_1{}^{(i)}, e_2{}^{(i)}, \cdots, e_k{}^{(i)}]^T (y - c^{(i)}) \tag{11}$$

   From the distance $d_r{}^{(i)}$ between the projected point $z^{(i)}$ and the points $p_r{}^{(i)}$ in the subspace of category i

$$d_r{}^{(i)} = |z^{(i)} - p_r{}^{(i)}| \tag{12}$$

   we detect the pose.
   In general, to detect the pose we calculate r that minimizes $d_r{}^{(i)}$. But to detect facial pose of the unlearned person robustly, $d_r{}^{(i)}$ must be a function of r independent of i. Therefore, we define this $d_r{}^{(i)}$ to be an evaluation value.

2.1.3 CLAFIC Method

Projective vector $\hat{y}$ into subspace i of an input vector y is shown as
covariance matrix

$$\hat{y} = \sum_{j=1}^{k} \{(y - c^{(i)})^T e_j{}^{(i)}\} e_j{}^{(i)} \tag{13}$$

correlation matrix

$$\hat{y} = \sum_{j=1}^{k} \{y^T e_j{}^{(i)}\} e_j{}^{(i)} \tag{14}$$

We use the squared norm of $\hat{y}$ as the evaluation value of similarity to the subspace constructed in the section 2.1.1.
covariance matrix

$$\|\hat{y}\|^2 = \sum_{j=1}^{k} \{(y - c^{(i)})^T e_j{}^{(i)}\}^2 \tag{15}$$

correlation matrix

$$\|\hat{y}\|^2 = \sum_{j=1}^{k} \{y^T e_j{}^{(i)}\}^2 \tag{16}$$

And then, we search the area that maximizes $\|\hat{y}\|^2$ to extract the face area.

2.2 Genetic Algorithm

To search the faces of which both position and size are unknown in an image is just the maximum searching problem of $\|\hat{y}\|^2$ in the 4-dimensional space of (x_c, y_c, M_x, M_y) that are the center of gravity and magnifying power of the extracted area. We apply the Genetic Algorithm [8] to this searching problem.

1. Individual
   The genotype $G_k(k = 1, 2, \cdots, N)$ of the individual B_k is defined as follows

$$G_k = (x_{ck}, y_{ck}, M_{xk}, M_{yk}) \quad (17)$$

$$(x_{ck}, y_{ck}) \ : \ \text{the center of gravity}$$
$$(M_{xk}, M_{yk}) \ : \ \text{magnifying power}$$

2. Fitness Value
   The fitness value F is defined as follows

$$F = \|\hat{y}\|^2 \quad (18)$$

   Here, $\|\hat{y}\|^2$ is the evaluation value considered in the section 2.1.3.

3. Selection
   N individuals are generated in one generation. $R\%$ individuals of which the fitness values are lower die out. Let this number of individuals be $M = N \times \frac{R}{100}$.

4. Multiplication
   $(100 - R)\%$ individuals of which the fitness values are upper survive in the next generation. Then from this $(N - M)$ individuals new M individuals are generated, the total number is fixed in every generation. $S\%$ individuals within a range die out however big the fitness values are, while they are less than a threshold. This prevents all individuals from gathering in one place in the early generation.

5. Heredity Rule
   When from an individual B_i with genotype G_i a new individual $B_i{}'$ is generated, its genotype $G_i{}'$ defined as follows

$$G_i{}' = (x_{ci} + \Delta x_i, y_{ci} + \Delta y_i, M_{xi} + \Delta M_{xi}, M_{yi} + \Delta M_{yi}) \quad (19)$$

$$\Delta x_i \ : \ \text{rand num in } [-\Delta x_{max}, \Delta x_{max}]$$
$$\Delta y_i \ : \ \text{rand num in } [-\Delta y_{max}, \Delta y_{max}]$$
$$\Delta M_{xi} \ : \ \text{rand num in } [-\Delta M_{xmax}, \Delta M_{xmax}]$$
$$\Delta M_{yi} \ : \ \text{rand num in } [-\Delta M_{ymax}, \Delta M_{ymax}]$$

   Here, Δx_{max},Δy_{max},ΔM_{xmax},ΔM_{ymax} are functions of F_i, they are decided by learning as follows

   (a) We extract a face area from an image by the visual observation and let the position be (x_0, y_0), the magnifying power be (M_{x0}, M_{y0}). Then we calculate the fitness value F of the $(x_0 + \Delta x, y_0 + \Delta y, M_{x0} + \Delta M_x, M_{y0} + \Delta M_y)$.

   (b) We let Δx,Δy,ΔM_x,ΔM_y be the functions of fitness value F, and record the Δx,Δy, ΔM_x,ΔM_y of various fitness values. When the fitness values are equal, we record the maximum Δx_{max}, Δy_{max}, ΔM_{xmax}, ΔM_{ymax}. And when some fitness values don't appear in this process, we define $\Delta x_{max}, \Delta y_{max}, \Delta M_{xmax}, \Delta M_{ymax}$ as follows

$$\Delta x_{max} = xmax(F - 1)^2 \quad (20)$$
$$\Delta y_{max} = ymax(F - 1)^2 \quad (21)$$
$$\Delta M_{xmax} = M_x max(F - 1)^2 \quad (22)$$
$$\Delta M_{ymax} = M_y max(F - 1)^2 \quad (23)$$

   Here, $xmax$,$ymax$,$M_x max$,$M_y max$ is the maximum of Δx_{max},Δy_{max},ΔM_{xmax}, ΔM_{xmax} in the neighborhood of face.

6. Mutation
   M individuals are generated at random once in every L generations, not following the heredity rule defined above.

3 Experiment 1: Examination of accuracy

3.1 Outline

In this experiment, R equals to 36, the number of categories is 3. So we prepare 108 images that consist of three persons of which horizontal rotation steps are 36. We cut these images in two ways. One is the face image including the region of hair, and the other is the image of the inner region of face, as shown in Fig.2 respectively. The size of them is normalized to 64×64 pixels. Then we construct the subspace of every person from various extracted features as Fig.2 indicate some of them.

original image smoothing edge edge+smoothing

Training Data 1

original image smoothing edge edge+smoothing

Training Data 2

Figure 2. Training Data

We examine the best feature vector for the facial pose detection by an evaluation value $d_r^{(i)}$ defined in the section 2.1.2. To extract face area we compare the search space in two methods (using covariance matrix and correlation matrix). In general, to extract face area, the feature vector is considered in two points that is 1) searching process can be early and easily, and 2) background area can be rejected and face area can be decided exactly. But to search the face area of which the location and size are unknown is the optimum value searching problem in the 4-dimensional space, so convergence is difficult. Therefore, in this experiment we examine if it is able to reject the background area and decided the exact face area mainly.

3.2 Results and Considerations

3.2.1 Pose Detection

The results are shown in Fig.3 and Fig.4. The vertical axis indicates the distance $d_r^{(i)}$ defined in the section 2.1.2. In this experiment the number of category is 3, so three parts of the horizontal axis indicate the person A, B, C classified as category $i = 1, 2, 3$ respectively. The horizontal axis in every part indicates the horizontal face rotation step $r = 1, 2, \cdots, 36$.

Fig.3 and Fig.4 indicate some results when the various features extracted from the some images of the person A are projected into every subspace constructed from the images of the person A, B, C. The dotted line in every part indicates the facial pose of the person A projected into every subspace. Fig.3 indicates the results when the feature vector extracted from images including the region of hair are projected. This line graph suggests us that the facial pose is detected only when the features are projected into the subspace constructed from the images of person A. It is clear that we must prepare the facial images of the person of which we intend to detect the pose. On the other hand, Fig.4 indicates the results when the various features extracted from the images of the inner region of face are projected. These figures show that the facial pose is detected when the features are projected into not only the subspace A but also the subspace B, C. Therefore it is possible that facial pose of any person is detected without many sample data.

Then, Fig.5 and Fig.6 indicate the difference from the correct pose. The horizontal axis in the graphs shows the difference. The white region in the bar chart shows the number for the learned person. The shading shows the number for the unlearned person. The dotted line in the line graph shows the ratio for the learned person. The solid line shows the ratio for the unlearned person. The results for the learned person don't make much dif-

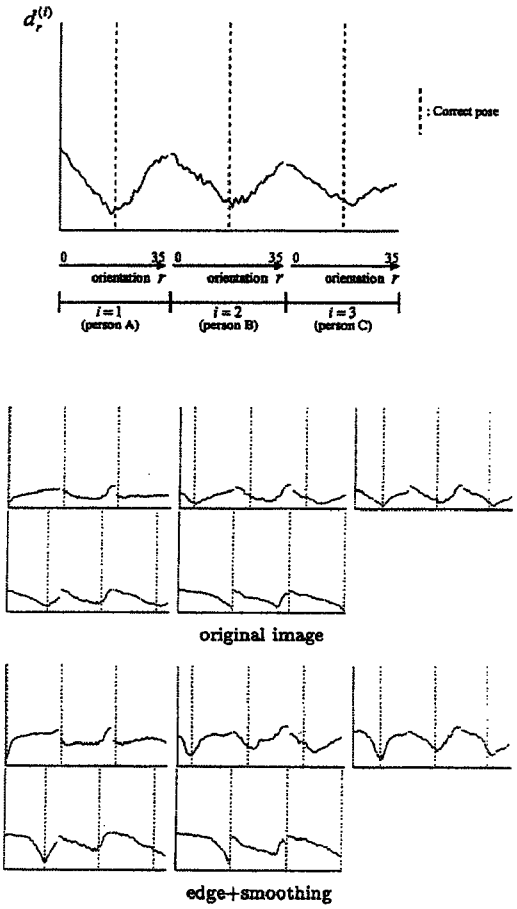


Figure 3. variation in $d_r^{(i)}$ (Training Data 1)

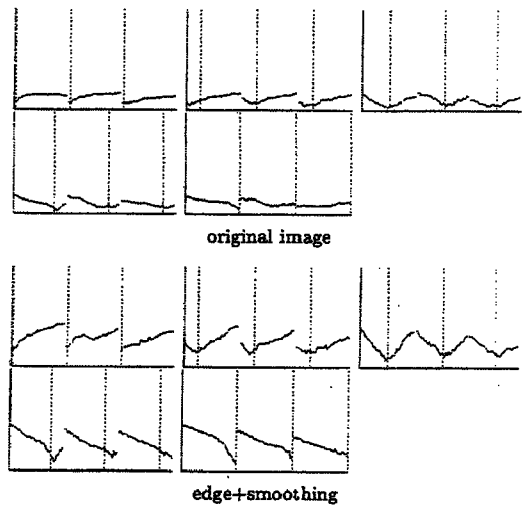


Figure 4. variation in $d_r^{(i)}$ (Training Data 2)

ference, but for the unlearned person the pose is detected correctly using the features extracted from images of the inner region of face. Particularly when smoothing edge features are used, the difference of the distance $d_r^{(i)}$ is big as Fig.4 indicates, and serious error is not made as Fig.6 indicates.

It follows from what has been said that we may detect the facial pose for the unlearned correctly if the features extracted from images of the inner region of face are used. Therefore our purpose in the next section is extracting inner region of face as Fig.2 shows.

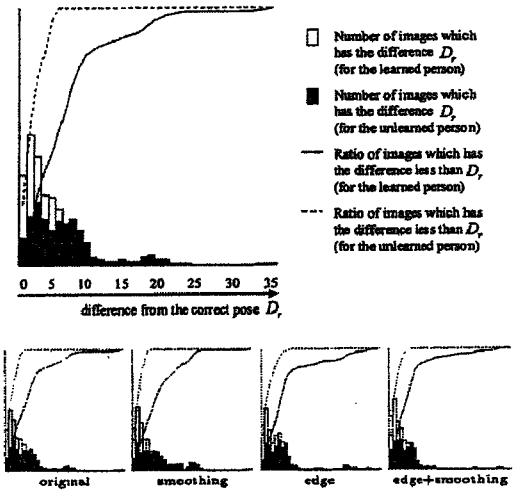


Figure 5. difference from the correct pose (Training Data 1)

Figure 6. difference from the correct pose (Training Data 2)

3.2.2 Face Extraction

From the results in section 3.2.1 we try to cut the face area of Training Data.2 shown in the Fig.2.

To extract the face area is just to reject the background area. But the various background images are not trained, so the curved surface in the search space is unpredictable. Fig.7 and Fig.8 indicate noticeable examples of the search space on the assumption that face size is fixed for the images taken on the various conditions. In Pattern Matching the vertical axis in the graphs indicates the Euclid Distance, and in CLAFIC Method it is the $\|\hat{y}\|^2$ defined in the section 2.1.3, so extracting the face area is equal to looking for the minimum and maximum in this search space respectively. The black region in the graphs shows the face area with which the template overlaps more than half. Here, these two examples shown in Fig.7 and Fig.8 are noticeable examples of images taken on the various conditions. Data.1 is the image to make convergence difficult, which has different background from face color near the face area and similar color to face in the other region. Data.2 is the image has complicated background. Data.1 and Data.2 are the images taken on different illumination condition from the training data.

The search space using the original images is shown in Fig.7, but the face area is not the optimum value in the search space. Then, to reduce the influence of illumination change we have done an experiment using edge feature. The results are shown in Fig.8.

In the Fig.8 the face area is the optimum value in the search space though the graph is discontinuous curved space. As the graphs indicate, in the method using correlation matrix the score in the background is not restrained. So it is possible to converge on the points in the background. On the other hand, using covariance matrix the score in the face area is obviously higher than in the background, so it is possible to decide the face area exactly.

Therefore in the next section we experiment with the CLAFIC method using covariance matrix for edge feature and evaluate if it is possible to extract the face area accurately.

4 Experiment 2:Face Extraction using Genetic Algorithm

4.1 Outline

We construct the subspace from edge features as Fig.2 shows. And we experiment with two methods. One is the method of constructing one subspace from 36 images. The other is the method of constructing three subspaces from three sets of 12 images, that are sets of images face to left, right, and center respectively. And in the latter method the fitness value F is defined as follows

$$F = \sum_{i=0}^{2} \|\hat{y}_i\|^2 \tag{24}$$

We use the Genetic Algorithm considered in the section 2.2 to extract face area. The number of individuals in one generation is 36, 18 individuals

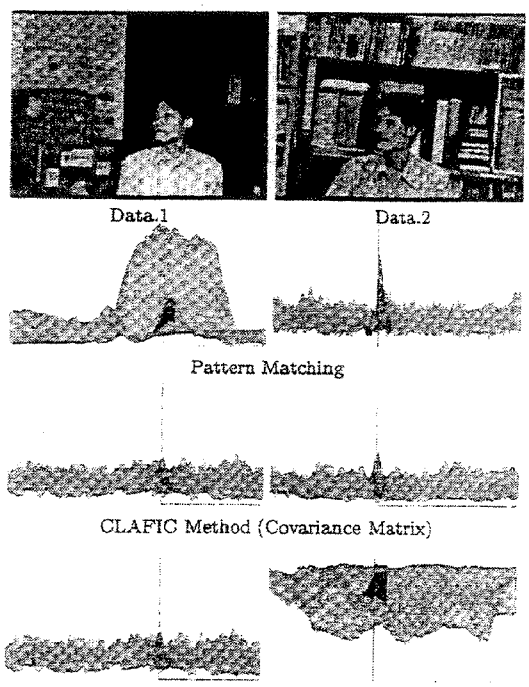


Figure 7. Search Space(original image)

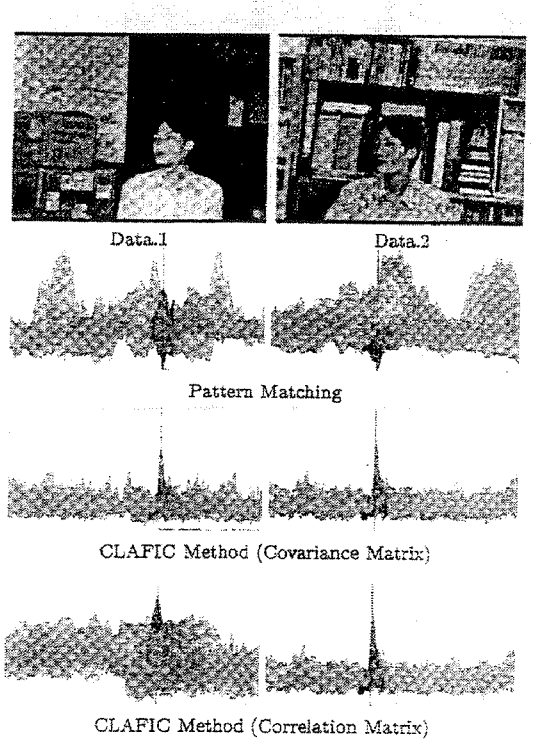


Figure 8. Search Space(edge)

of which the fitness values are lower die out, and then from the survived 18 individuals new 18 individuals are generated. The new 18 individuals are mutated once in every 20 generations. Then, to prevent searching space from being huge, we don't generate the individuals at random in the first generation, but generate in the area that can be face area by using hue.

Then we consider the evaluation value of extraction accuracy. Let the area extracted by the visual observation be C, the area extracted by GA be H. Then we calculate the size S_D of the area D as follows

$$D = (C \cap \bar{H}) \cup (\bar{C} \cap H) \tag{25}$$

And we define an evaluation value R as follows

$$R = \frac{S_D}{S_C} \tag{26}$$

Then, we define another evaluation value as the differences D_r between the correct pose and the pose r minimizes d_r defined the section 2.1.2.

In this experiment we use 250 images with various backgrounds.

4.2 Results

The results are shown in Fig.9- Fig.12. The top image in Fig.9-Fig.11 is one of the images we try to search. The horizontal axis in the graph indicates the generation of GA. The vertical axis in the upper part of the graph indicates F, and it in the lower part indicates R. The next images are face areas extracted in various generations and the training data of which the pose is detected. And Fig.12 indicates some of the results of face extraction and the pose detection in the various images. In Fig.9 and Fig.10, as generations go by, F gradually increases, R decreases. So it follows that the face area is extracted accurately.

Fig.13 indicates the extraction accuracy. The horizontal axis in these graphs shows R, the vertical axis shows the number of the extracted images of which the accuracy is R. And the line graph shows the accumulated rate. Here, one step of the horizontal axis is 0.05.

Fig.14 indicates the extraction accuracy by another evaluation value D_r. The horizontal axis in these graphs shows $|D_r|$.

As Fig.13 and Fig.14 indicate, when we use the method of constructing three subspaces, the accuracy is improved. Then we consider the difference of accuracy in generations. As generations go by, the accuracy is gradually improved, but doesn't make much difference. So it seems that we can stop the searching in the early generation as the case may be.

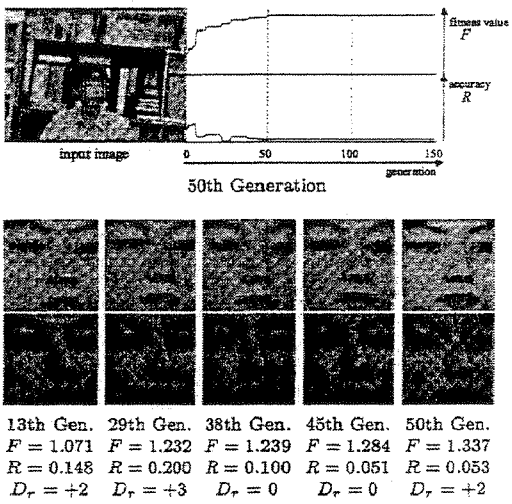


Figure 9. Result 1 (success)

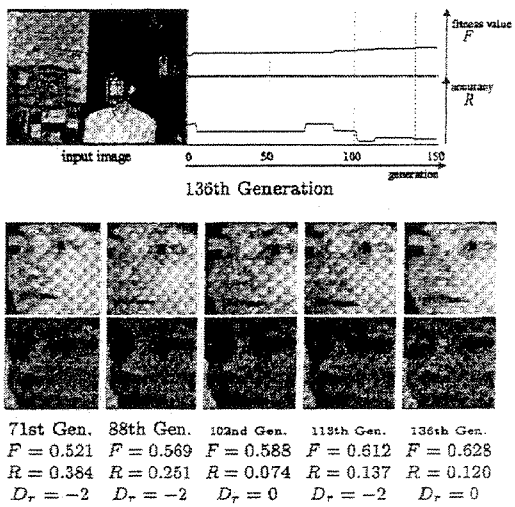


Figure 10. Result 2 (success)

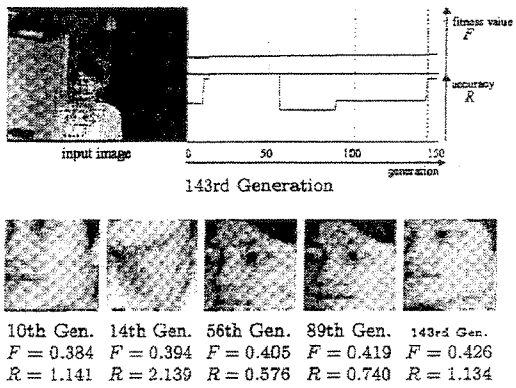


Figure 11. Result 3 (failure)

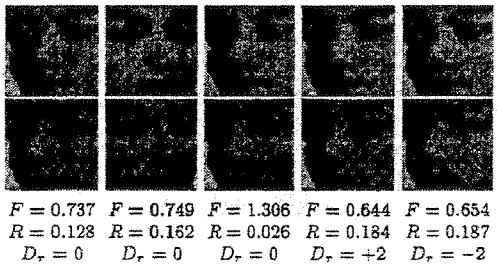


Figure 12. Some Results of Face Extraction and the Pose Detection

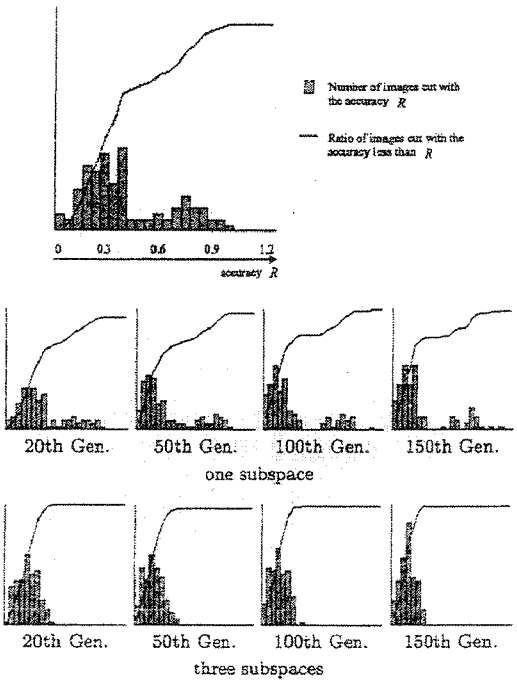


Figure 13. Extraction Accuracy 1

5 Conclusion

In this paper our purpose is to extract the face area and detect the pose. To detect the facial pose of anyone with few sample data, it is effective to use the features extracted from the images of the inner region of face. Particularly using edge features is the most effective.

And then, to extract the face area of which the location and size are unknown from the images with various backgrounds is just to search in the 4-dimensional space, so it is difficult to converge on the optimum value. So we considered the feature vector to be used for construction of subspace and the construction method. Consequently us-

ing covariance matrix the score in the face area is obviously higher than in the background, so it is possible to decide the face area exactly.

Furthermore, we could extract the human faces from images with various backgrounds using Genetic Algorithm actually. It follows that this method is effective for the discontinuous search space. And we could see the extraction accuracy is improved using the method of constructing three subspaces.

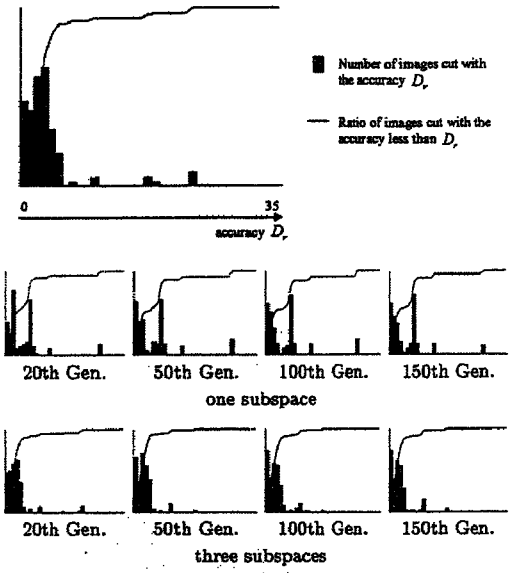


Figure 14. Extraction Accuracy 2

References

[1] J. Ahang, Y.Yan and M. Lades, "Face Recognition: Eigenface, Elastic Matchin, and Neural Nets," Proc. IEEE, Vol.85, No.9, pp.1422-1435, 1997.

[2] A. Hara and T. Nagao, "Extraction of facial regions of arbitrary directions from still images with a genetic algorithm," Technical Report of IEICE, HCS97-12, pp.37-44, 1997.

[3] N. Ishikawa and Y. Ariki, "Integration of Face and Speaker Recognition by Normalized Multiple Feature Subspace Method," Technical Report of IEICE, PRMU96-41, pp.31-38, 1996.

[4] Y. Komatsu and Y. Ariki, "Orientation Invariant Face Extraction and Recognition based on Subspace Method," Technical Report of IEICE, PRMU95-191, pp.7-14, 1996.

[5] M. Lades, J.C. Vorbruggen, J. Buhmann, J. Lange, C.v.d. Malsburg, R.P. Wurts, and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," IEEE Trans on Computers, Vol.42, No.3, pp.300-311, 1993.

[6] S. Masunaga and T. Nagao, "Extraction of human facial regions in still images using a genetic algorithm," Technical Report of IEICE, PRMU95-160, pp.13-18, 1995.

[7] H. Murase and S. K. Nayar, "3D Object Recognition from Appearance -Parametric Eigenspace Method-," Transaction of IEICE, J77-D-II, No.11, pp.2179-2187, 1994.

[8] T. Nagao and T. Agui and H. Nagahashi, "Pattern Matching of Binary Shapes Using a Genetic Method," Transaction of IEICE, J76-D-II, No.3, pp.557-565, 1993.

[9] E. Oja "Subspace Methods of Pattern Recognition," Research Studies Press, England, 1983.

[10] Y. Sugiyama and Y. Ariki, "Facial Region Tracking and Training Method by Subspace Projection," Technical Report of IEICE, PRMU97-162, pp.77-82, 1997.

[11] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, Vol.3, No.1, pp.71-86, 1991.

論　文

Labeled Graph Matchingを用いた動画像に対する人物頭部及び表情変化を伴う部位の抽出

村上　真[†*]　米山　正秀[††]　白井　克彦[†]

Extraction of Human Face and Transformable Region by Facial Expression Based on Extended Labeled Graph Matching

Makoto MURAKAMI[†*], Masahide YONEYAMA[††], and Katsuhiko SHIRAI[†]

あらまし　本論文では，動画像認識に対し，局所的特徴量と大局的構造及び運動の情報を同時に統合的に扱う新しい枠組みを提案し，人物頭部の表情変化等を抽出する問題に適用しその有効性を示す．従来の動画像認識法は，各フレームに対する認識結果系列から認識する手法がほとんどであるが，各フレームの認識に対しても積極的に動きの特徴を扱うべきである．しかし，静止画の認識に時系列データを組み込むことは，情報量が膨大になり，処理も複雑になる困難性がある．ここで提案する方法は，従来静止画に用いられてきたLabeled Graph Matching法の概念を拡張して適用するものである．この方法はSparse Graphを扱うため計算量の増大を防ぐことができ，様々な処理の段階に応じて扱う特徴量を動的に変化させることで，動画像認識における複雑な処理を簡潔に統合化することを可能にする．本論文では，具体例として，人物頭部及び部位の抽出を行い，本手法の有効性を示す．

キーワード　顔領域抽出，Labeled Graph Matching，動画像認識，ヒューマンインタフェース，人物顔認識

1. まえがき

人間と機械が円滑なコミュニケーションを行うためには，音声情報だけではなく，人物頭部から得られる様々な情報を認識することが有効である．人間の顔から得られる情報としては，個人性・表情・頭部動作（うなずき等）が挙げられ，人間同士のコミュニケーションにおいて，それらは重要な役割を果たしている．これまでにも，個人認証・表情認識に関する様々な研究[1], [2]が行われてきたが，静止画に対しての研究が主であり，あくまでも形状に関するパターン認識の問題に帰着できる性質のものであった．しかし，人間の顔から得られる情報に限って考えても，顔の動き，表情の変化など動的な情報も極めて大きな意味をもっている．このような考えに基づき主に表情認識の分野で顔の動きを用いた研究[3]が行われている．本研究は，動画像の動きの中に含まれる情報の抽出法について新しい方法を提案し，その有効性を調べたものである．

ここで用いる手法は，従来静止画に対して使われてきたLabeled Graph Matching法[4]〜[6]が基礎となっている．静止画によって形状情報のみから人物顔を認識するには，少なくともいくつかの特徴点を抽出する必要があり，一方で，いくつかの特徴点を決定するには顔全体の構造情報を用いなければならない．したがって，形状を認識するには，局所的特徴と大局的特徴を同時に，しかも様々なレベルにおいて扱う必要がある．Labeled Graph Matching法は，これを簡潔にモデル化する一つの方法である．

本研究は，この手法の概念を動画像認識に拡張するものである．具体的には，人物頭部の構造とその動き方の両方をモデルベースの手法によって記述して，時間変化も含む大局的モデルを用意する．これに局所的特徴量を合わせて，Labeled Graph Matchingの方法に基づいて処理することによって，人物頭部の動画像

† 早稲田大学理工学部，東京都
School of Science and Engineering, Waseda University, Tokyo, 169-8555 Japan
†† 東洋大学工学部，川越市
School of Engineering, Toyo University, Kawagoe-shi, 350-8585 Japan
* 現在，東洋大学工学部

に対して，局所的及び大局的空間特徴量だけでなく時間変化に対する特徴量を同時に扱うことが可能になる．

2. Labeled Graph による時間軸の表現

本章では静止画に対する Labeled Graph による空間的表現を時間軸情報を含む形に拡張する方法に関して述べる．

2.1 空間的特徴量

まず，Labeled Graph による空間的特徴の記述について示す．大局的構造と局所的特徴とを同時に扱うことを簡潔にモデル化するために，物体の空間的特徴を Sparse Graph で表現する．図 1 に入力画像と学習データセットからなるモデルに対する Labeled Graph 表現の概念図を示す．入力画像及び学習データはそれぞれ Image Domain，Model Domain と呼ばれる層に，“Vertex” と “Edge” からなる Sparse Graph として表現される．各 “Vertex” には局所的空間特徴量が対応し，“Edge” には大局的な構造情報が対応することとなる．

2.1.1 空間的局所特徴量

図 1 に示す各 Vertex には，GWT（Gabor Wavelet 変換）により様々なレベル・方向のウェーブレット係数として抽出された局所特徴量を保存する．入力画像の 256 階調濃度ベクトルを $L(\vec{x})$ とすると，画像の座標 $\vec{x}_0$ における，方向とサイズ $\vec{k}$ のウェーブレット係数は，マザーウェーブレット $\psi_{\vec{k}}$ と $L(\vec{x})$ の畳込み積分として以下のように表される．

$$(\mathcal{W}L)(\vec{k},\vec{x}_0) := \int \psi_{\vec{k}}(\vec{x}_0-\vec{x})L(\vec{x})d^2x = (\psi_{\vec{k}} * L)(\vec{x}_0) \quad (1)$$

ここでは，マザーウェーブレットは以下のように定義される Gabor Wavelet を用いる．

$$\psi_{\vec{k}}(\vec{x}) = \frac{\vec{k}^2}{\sigma^2}\exp\left(-\frac{\vec{k}^2\vec{x}^2}{2\sigma^2}\right) \cdot \left[\exp(i\vec{k}\vec{x}) - \exp\left(-\frac{\sigma^2}{2}\right)\right] \quad (2)$$

ここで，マザーウェーブレットのレベル，方向は以下のように定義される．

$$\vec{k}_{\nu\mu} = k_\nu e^{i\phi\mu}$$
$$k_\nu = \frac{k_{max}}{f^\nu}, \phi_\mu = \frac{\pi\mu}{8} \quad (3)$$

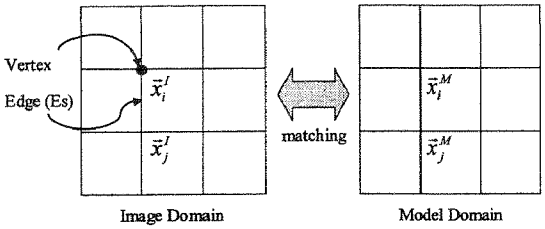


図 1 Labeled graph 表現（空間的特徴量）
Fig. 1 Labeled graph expression. (spatial feature)

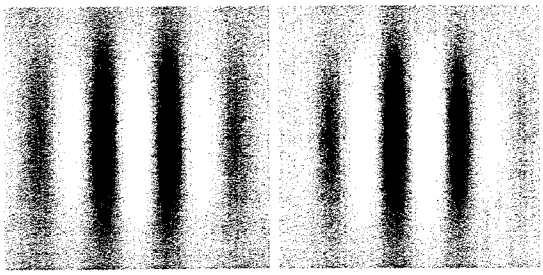


図 2 Gabor wavelet
Fig. 2 Gabor wavelet.

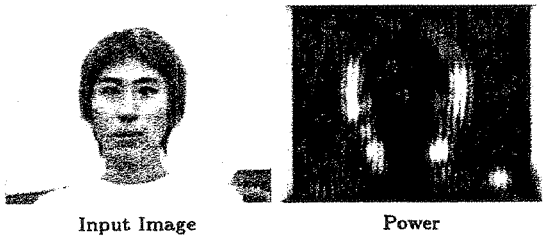


図 3 GWT $\nu = 0, \mu = 0$
Fig. 3 GWT $\nu = 0, \mu = 0$.

図 2 及び図 3 にマザーウェーブレットと人物顔画像から得られるウェーブレット係数のパワーの例をそれぞれ示す．なお，本論文での周波数解析は，4 レベル $\nu \in \{0,\cdots,3\}$，8 方向 $\mu \in \{0,\cdots,7\}$ とし，ウェーブレット係数のパワーを各 Vertex $\vec{x}_i$ の特徴量として用いる．

$$\mathcal{J}_{\nu\mu,i} := |(\mathcal{W}L)(\vec{k}_{\nu\mu},\vec{x}_i)|, i \in V \quad (4)$$

従来法では，学習データが静止画 1 枚であるため，入力画像 I と学習データ M との空間的局所特徴ベクトルの照合には，類似性の尺度として以下の関数を用いていた．

$$S_V(\mathcal{J}^I,\mathcal{J}^M) := \frac{\mathcal{J}^I \cdot \mathcal{J}^M}{||\mathcal{J}^I||||\mathcal{J}^M||} \quad (5)$$

本論文では，時系列データを扱うため，情報量が膨

大になることを考慮し，学習データセットを Eigenface 法 [1] により学習し，固有空間上で入力画像 I とモデル M の照合を行う．なお，一般的な Eigenface 法では共分散行列を用い平均顔を原点とする固有空間を構成するが，本論文では頭部領域抽出を目的としているため相関行列を用いて固有空間を構成する．

学習データセットの特徴ベクトルの集合から行列 X

$$X = [\mathcal{J}_1^M, \cdots, \mathcal{J}_p^M, \cdots, \mathcal{J}_P^M] \tag{6}$$

及び，相関行列 R

$$R = XX^T \tag{7}$$

を求め，固有値 λ_q，固有ベクトル $e_q, (q = 1, \cdots, Q, Q \leqq P)$ を得る．得られた固有ベクトルの張る空間上でモデル M は以下のように表現される．

$$\mathcal{J}'^M_p = [e_1, \cdots e_q, \cdots e_Q]^T \mathcal{J}_p^M \tag{8}$$

入力画像 I とモデル M との照合には，固有空間上でのユークリッド距離の最小値 S_{Vs} を用いる．

$$S_{Vs}(\mathcal{J}'^I - \mathcal{J}'^M_p) := \min_p \|\mathcal{J}'^I - \mathcal{J}'^M_p\| \tag{9}$$

ここで，$\mathcal{J}'^I$ は入力画像 I の固有空間上への射影である．

2.1.2 構造的特徴量

図 1 の両 Domain において，Vertex $\vec{x}_i$ と $\vec{x}_j$ とを空間的に結ぶ Edge の集合を Es とすれば，次に定義する隣り合う Vertex 間の相対的位置として構造的情報が表される．

$$\Delta_{ij} := \|\vec{x}_j - \vec{x}_i\|, (i,j) \in Es \tag{10}$$

Image Domain と Model Domain 間の構造的照合には以下の関数を用いる．

$$S_{Es}(\Delta^I, \Delta^M) := \min_p \sum_{(i,j)\in Es} \|\Delta^I_{ij} - \Delta^M_{ij,p}\| \tag{11}$$

照合の過程で，構造的特徴量 $S_{Es}(\Delta^I, \Delta^M)$ は空間的な拘束条件として考えられる．学習データに現れない変形に対してはこの拘束を緩めることにより柔軟な認識が可能になる．1 枚の学習データから個人認証を行う場合には，Δ^M_{ij} は確定的となるため，回転・変形を伴う人物を認識する際に S_{Es} の拘束は重要な役割を果たす．本研究では学習データセットとして回転・変形を含む動画像を用いるため，データセットに現れない変形に対しても対処できる柔軟な制限を与えることとなる．

2.2 時間的特徴量

時間的特徴量に関しては，空間的特徴量の概念 "Vertex" と "Edge" をそれぞれ時間方向に拡張するもので，時間的局所特徴量及び大局的な動き方の特徴量が与えられる．その概念図を図 4 に示す．空間的特徴量と大きく異なる点は，データ表現が空間的特徴量では Sparse であるが，時間的にはそうではない点である．したがって，時間的特徴量について Image Domain と Model Domain で照合を行うと，大局的な動きの特徴のみが有効となるので，局所的特徴量の時間変化について，Model Domain との照合は行わない．しかし，一般的な物体追跡の考え方から Image Domain 内での処理において各 Vertex における速度ベクトルを使用しているので，ここでは時間的局所特徴量として各 Vertex における速度ベクトルを，大局的特徴としては速度ベクトルのシーケンスをそれぞれ対応させて用いている．

物体の動きを認識することは，Model Doamin に表現された動作との照合を行うことであり，Edge における Image Doamin と Model Domain との照合に他ならない．したがって，物体動作を認識するには，認識対象となる動作を学習データセットとして用意しモデルを構成すればよい．本論文の実験はさしあたり頭部領域抽出とした．したがって，可能な頭部動作すべてを学習データセットとして取得しなければならないこととなる．頭部動作として考えられるものとして，剛体としての平行移動・拡大縮小・3 次元空間の 3 軸に対する回転と，非剛体としての表情変化等による動作が挙げられる．実際にはこれらが複雑に絡み合うため，すべての動作を学習データとして用意するのは不可能である．本論文では，剛体としての運動は大局的特徴照合の際には扱わないこととした．

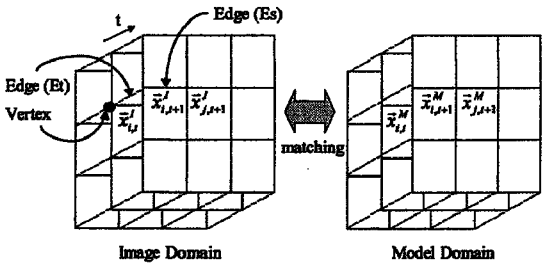


図 4 Labeled graph 表現（時間的特徴量）
Fig. 4 Labeled graph expression. (temporal feature)

2.2.1 時間的局所特徴量

時間的局所特徴量としては，Optical Flow [7] から得られた特徴量を用いる．Optical Flow による手法は，こう配法 [8]〜[10] とブロックマッチング法 [11], [12] に大別されるが，照合の際に形状によるマッチングを行うことと，計算量の点を考慮して 1 次の導関数によるこう配法を用いる．また本論文では，画像全体の速度場ではなく，各 Vertex における特徴量のみを抽出することを目的としているので，Lucas and Kanade の手法 [13], [14] を用いる．局所領域 Ω を 5×5 ピクセルとし，時刻 t における Vertex $\vec{x}_{i,t}$ の速度ベクトル $\Phi_{i,t}$ を求め，時間的局所特徴量とする．

本論文において学習データセットとして用いるのは，水平方向の回転・垂直方向の回転・発話による変形の 3 種類であるが，人物頭部動作すべてを学習データとして用意することは不可能である．したがって，時間的局所特徴量として，Φ^I と Φ^M とを照合することはせず，一般的な物体追跡処理を行うこととなる．

（1） 空間的関連性

時刻 t における Image Domain で，Edge Es によって空間的に結ばれた Vertex を $\vec{x}^I_{i,t}, \vec{x}^I_{j,t}$，それぞれの速度ベクトルを $\Phi^I_{i,t}, \Phi^I_{j,t}$ とする．追跡対象物体が剛体と仮定すると，各頂点は一定の構造を保ちながら運動することから，Image Domain 中での空間的な連続性として以下の関数を定義する．

$$S_{Vts}(\Phi^I_t) := \sum_{(i,j)\in Es} \frac{\Phi^I_{j,t}\cdot\Phi^I_{i,t}}{\|\Phi^I_{j,t}\|\|\Phi^I_{i,t}\|} \tag{12}$$

これは，概念的には **2.1.1** の空間的局所特徴量としてとらえられるが，上述したとおり Image Domain と Model Domain との照合ではないため，本論文では独立して扱うこととする．

（2） 時間的関連性

Image Domain における各 Vertex の速度ベクトルは時間方向に滑らかに変化すると仮定し，時間的連続性による拘束条件として，以下の関数を定義する．

$$S_{Vtt}(\Phi^I_t) := \sum_{i\in V} \frac{\Phi^I_{i,t+1}\cdot\Phi^I_{i,t}}{\|\Phi^I_{i,t+1}\|\|\Phi^I_{i,t}\|} \tag{13}$$

これにより，前後のフレームからの制限を加えることを可能にする．

2.2.2 時間的特徴照合

人物顔特有の変化に対しては，DP マッチングにより学習データセットに現れる人物顔の動き方との照合を行い，より精度の高い領域抽出を目指す．

時間的情報を大局的にモデルと照合する処理は，時間方向に伸縮させながらマッチングすることと同意であるから，本論文では，DP マッチングを用いる．時刻 t における速度ベクトルの平均を C_t とし，入力画像に対する相対的速度ベクトル系列 $\Psi^I = \{\Phi^I_1 - C^I_1, \cdots, \Phi^I_s - C^I_s, \cdots, \Phi^I_S - C^I_S\}$ と学習画像に対する相対的速度ベクトル系列 $\Psi^M = \{\Phi^M_1 - C^M_1, \cdots, \Phi^M_t - C^M_t, \cdots, \Phi^M_T - C^M_T\}$ に対して，(Ψ^I, Ψ^M) からなる平面を考えると，この平面上の格子点 $g = (s,t)$ の系列 F は

$$F = g_1, \cdots, g_u, \cdots, g_U, (g_u = (s_u, t_u)) \tag{14}$$

と表される．二つの特徴ベクトル Φ^I_s と Φ^M_t との距離を $d(g) = d(s,t)$ とすると，F に沿った距離の総和は

$$S_{Et}(\Psi^I, \Psi^M) := H(F) := \frac{1}{S+T}\sum_{u=1}^{U} d(g_u)\cdot w_u \tag{15}$$

$$w_u = (s_u - s_{u-1}) + (t_u - t_{u-1}) \tag{16}$$

で表され，この値が小さいほど Ψ^I と Ψ^M の対応付けがよいことを示す．

3. 統合的照合法

静止画に対する Labeled Graph Matching 法の概念を時系列データに拡張して，動画像認識に適用することを提案したが，この処理では局所的特徴量と大局的な構造及び時間的な変化の特徴を様々なレベルで同時に扱う必要がある．しかし，すべての情報を様々な処理の段階にかかわらず用いることは，問題を複雑にするだけではなく解決を困難にすることとなる．本章では，処理の段階に応じてそれぞれの情報を適切に取り扱い解を得るための仕組みに関して述べる．

3.1 Labeled Graph Matching における処理

Labeled Graph Matching では，空間的局所特徴量と大局的な空間構造を線形結合することにより Cost Function を以下のように定義し，最急降下法により C_{Total} が最小になるように Image Domain と Model Domain を照合する．

$$C_{Total} := C_V + \lambda C_E \tag{17}$$

ここで，

$$C_V = -S_V(\mathcal{J}^I, \mathcal{J}^M) \tag{18}$$

$$C_E = S_E(\Delta^I, \Delta^M) \tag{19}$$

である．

実際の処理においては，領域を抽出する段階では，$\lambda \to \infty$ と発散させることにより，構造的拘束条件を最大にし，回転による見かけの変化や物体自体の変形に伴う影響を無視しておおよその位置を検出する．次に，それが収束した時点で λ を定数値に設定することにより，Model Domain に提示される構造からの一定の制限を受けつつ，物体の変形に対して柔軟な照合を可能にしている．

3.2 Labeled Graph Matching の動画像への拡張

3.2.1 Cost Function

静止画に対する Labeled Graph Matching では，局所特徴量と構造的特徴の二つしか扱っておらず，Cost Function はそれらの線形結合によって表現され，照合の際に変化させる相対的パラメータは λ のみである．これに **2.2** で定義した時間的特徴量を付加し，同様に線形結合により統合すると以下のようになる．

$$C_{Total}(\{x_i^I\}) := C_{Vs} + \lambda C_{Es} + \alpha C_{Vts} + \beta C_{Vtt} + \gamma C_{Et} \tag{20}$$

$$C_{Vs} = S_{Vs}(\mathcal{J}'^I, \mathcal{J}'^M) \tag{21}$$

$$C_{Es} = S_{Es}(\Delta^I, \Delta^M) \tag{22}$$

$$C_{Vts} = -S_{Vts}(\Phi^I, \Phi^I) \tag{23}$$

$$C_{Vtt} = -S_{Vtt}(\Phi^I, \Phi^I) \tag{24}$$

$$C_{Et} = S_{Et}(\Psi^I, \Psi^M) \tag{25}$$

第 3 項，第 4 項が表すのは，運動する物体における時間的空間的連続性による拘束である．第 3 項は認識対象物体が剛体と仮定できる場合の各 Vertex の空間的連続性を表し，α が大きいほど空間的連続性による拘束が大きくなる．第 4 項は各 Vertex の前後のフレームに対する時間的連続性を表し，β が大きいほど前後のフレームによる拘束が大きくなる．第 5 項は時間的特徴量を時間軸方向に伸縮させながら大局的に照合する場合の Cost であり，γ を大きく設定するほど Model Domain に提示される動作による影響を強く受けることとなる．なお本論文では，**2.1.1** で述べたとおり，空間的局所特徴量の照合には固有空間上でのユークリッド距離を用いているため式 (20) の第 1 項が従来法と異なるが，同様の効果を及ぼすものである．

表 1 処理過程によるパラメータ設定
Table 1 Parameter set in each process.

処理過程	λ	α	β	γ
1	∞	α_{max}	β_{min}	γ_{min}
2	∞	α_{max}	β_{max}	γ_{min}
3	λ_{min}	α_{min}	β_{max}	γ_{min}
4	λ_{min}	α_{min}	β_{max}	γ_{max}

3.2.2 制御方法

以上の四つのパラメータを動的に変化させることにより，処理過程に応じた特徴量を用いることが可能となる．本論文では領域抽出を目的としているため，表 1 のようにパラメータを変更することで処理過程の制御を行う．簡単のため各パラメータ ρ は ρ_{max} と ρ_{min} の 2 値とする．

（1） 処理過程 1

$\lambda \to \infty$ とすることにより構造的拘束を最大にし，Model Domain における構造を不変とする．更に，α を α_{max} に設定することにより Image Domain における速度ベクトルの空間的連続性による一定の拘束をかける．また，β を β_{min} とすることで時間的連続性による拘束を小さく設定し，各フレームで独立して処理を行う．

（2） 処理過程 2

β を β_{max} に設定することにより速度ベクトルの時間的連続性による拘束をかけ，前後のフレームによる効果を与える．

（3） 処理過程 3

α を α_{min} とし，速度ベクトルの空間的拘束条件を解放するとともに λ を λ_{min} とすることで構造的拘束を解放する．この過程は剛体の運動であるという仮定を取り去り，非剛体の運動として物体をとらえる過程である．

（4） 処理過程 4

γ を γ_{max} とすることで動作特徴の照合を行い，Model Domain に提示される動作による効果を付加する．

4. 実　　験

4.1 実験条件

本論文では **2.** で述べた Labeled Graph 表現を用いるが，非剛体としての動作のみを大局的特徴照合の際に用いるため，表情変化等に伴い変化する部位を，剛体として構造を保ちながら動作する頭部領域とは別に表現する．具体的には，図 5 に示すように，特徴量を

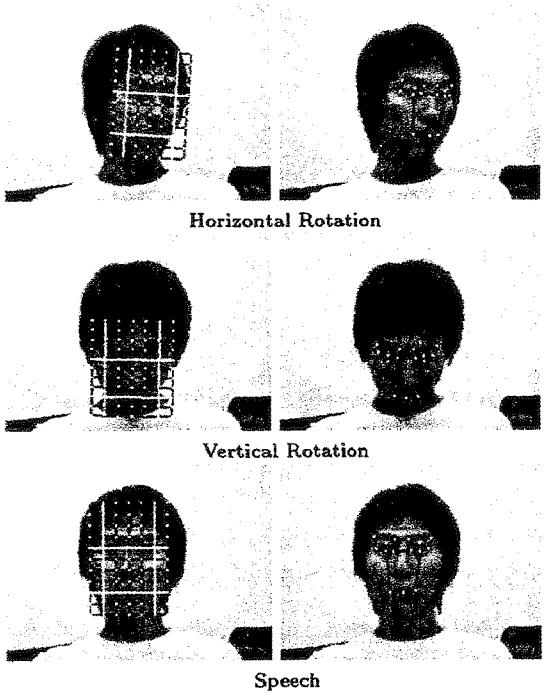


図 5　学習データ
Fig. 5　Training data.

明示的に二つに分割して考える．したがって，**3.2.2** で述べた処理過程 1 及び 2 は頭部領域を抽出することに対応し，3 及び 4 は部位抽出過程に対応する．部位に関しては長方形の格子ではないが，学習データセットにおける部位の空間的構造と各部位の位置関係の情報を含むため，Labeled Graph Matching の概念を覆すものではない．よって，長方形格子における言葉である "Vertex" と "Edge" を，それぞれ各頂点の局所的特徴量，大局的な構造及び運動という意味で使用する．

本実験に用いる学習データセットは図 5 に示した水平方向の回転・垂直方向の回転・発話による変形の 3 種類のデータである．図中の白い点が Vertex を表し，黒い線が Edge を表す．頭部領域に関しては，図でグレーの線で示した目の高さ・口の高さ及び両目尻をそれぞれ縦方向・横方向の基準として領域を定義している．なお，頭部領域を表現する Vertex は 7×11 点，部位を表現する Vertex は 16 点とし，それぞれを図のように連結することにより Edge を定義する．認識用データは，学習データセットと同じ水平方向の回転（Data 1）・垂直方向の回転（Data 2）・発話による変形（Data 3）に加えて，頭部動作に制限をつけないで撮影したデータ（Data 4）の 4 種類を用いる．実験試行数は，フレーム長 30 のデータで，Data 1 が 55 セット，Data 2 が 65 セット，Data 3 が 69 セット，Data 4 が 57 セットである．学習及び認識に使用した動画像データの詳細を表 2 に示す．このデータに対し，時間的空間的特徴量を設定し，Image Domain と Model Domain の照合を行う．処理過程 1 では，β を β_{min} とすることで，各フレームで独立して処理を行うこととなる．したがって初期状態は各フレームにおいて乱数により設定し，式 (20) で定義した Cost Function が最小になるように，最急降下法により顔領域を探策する．このとき，変更するパラメータ ρ の上限 ρ_{max} 及び下限 ρ_{min} には表 3 に示す値を設定した．なお本論文では，式 (20) の第 1 項に固有空間上でのユークリッド距離を用いているため，特徴量の異なる処理過程 1，2 と処理過程 3，4 では第 1 項の値が大きく異なる．そのため，実際には頭部領域抽出過程では $\rho \times 10^6$ とし部位抽出過程では $\rho \times 10^4$ として値の正規化を行っている．

表 2　動画像データ詳細
Table 2　Data format.

	Data Format
frame rate	30 frames/sec
image size	320 × 240 pixels
resolution	256 gray scale

表 3　実験におけるパラメータ設定値
Table 3　Parameter set in experiment.

	λ	α	β	γ
max	∞	0.8	0.7	0.7
min	0.0	0.0	0.0	0.0

4.2 実験結果

図 6 に抽出結果の例を，表 4〜表 7 に処理過程別の抽出精度の平均を示す．表中の精度は，各 Vertex $\vec{x}_i$ の，目視により指定した正解位置 $\vec{x}_i^C, (i = 1, \cdots, N)$ からのユークリッド距離の平均として，以下のように定義する．

$$D = \frac{1}{N} \sum_{i=1}^{N} \|\vec{x}_i^C - \vec{x}_i^I\| \tag{26}$$

また，精度 D が 4.0 以下のときに抽出成功としたときの成功率と成功時における計算量の平均も表に示す．ここで扱う計算量は，最急降下法により探索した回数であり，処理時間そのものは各過程によって異なる．また，時間的特徴量を用いない場合との比較のため，

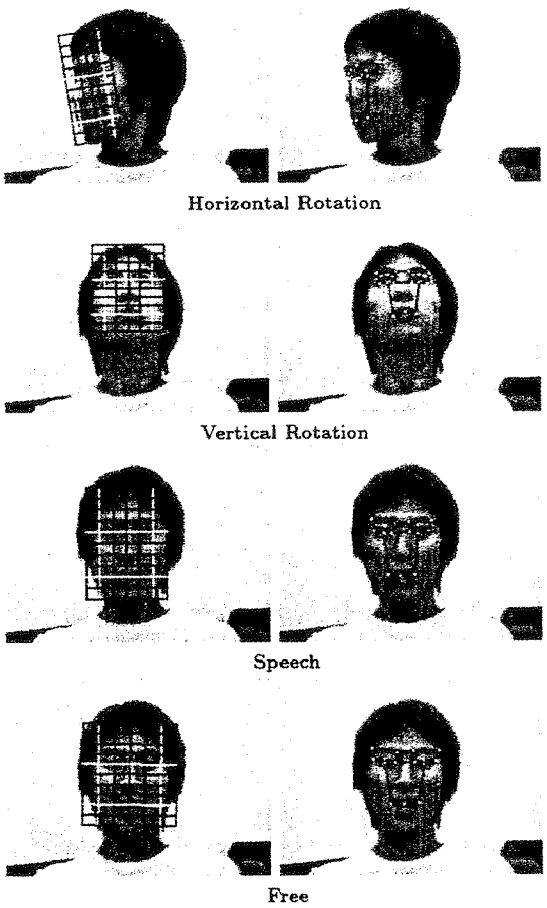


図 6 顔領域抽出結果
Fig. 6 Results of human face extraction.

表 4 処理過程別結果（Data 1）
Table 4 Results of extraction. (Data 1)

処理過程	抽出精度	成功率	計算量
0（頭部領域）	43.5	60.7	358.3
1（頭部領域）	16.7	82.5	128.2
2（頭部領域）	13.8	88.3	190.1
3（部位）	8.9	72.2	245.0
4（部位）	7.6	78.8	208.4

処理過程 0 として，式 (20) の第 1 項のみを用いた結果を併記しておく．

まず処理過程 1 の時間的特徴量の空間的拘束条件による影響を考察すると，回転を伴う頭部全体の動きが激しい Data 1 及び Data 2 に関して精度の向上が見られ，計算時間の大幅な短縮に成功している．頭部全体の動きがない Data 3 に関しては効果は見られないが，拘束のない Data 4 に関しても精度・計算量ともに向上しており，十分有効であることが示された．

表 5 処理過程別結果（Data 2）
Table 5 Results of extraction. (Data 2)

処理過程	抽出精度	成功率	計算量
0（頭部領域）	34.4	65.5	312.2
1（頭部領域）	15.3	80.5	118.0
2（頭部領域）	7.2	88.6	212.5
3（部位）	8.2	67.5	228.2
4（部位）	6.5	75.9	211.7

表 6 処理過程別結果（Data 3）
Table 6 Results of extraction. (Data 3)

処理過程	抽出精度	成功率	計算量
0（頭部領域）	14.6	70.6	272.8
1（頭部領域）	17.3	72.5	285.3
2（頭部領域）	10.2	85.5	112.9
3（部位）	7.1	75.9	319.8
4（部位）	5.3	80.5	192.3

表 7 処理過程別結果（Data 4）
Table 7 Results of extraction. (Data 4)

処理過程	抽出精度	成功率	計算量
0（頭部領域）	31.2	67.2	285.0
1（頭部領域）	18.5	77.8	193.8
2（頭部領域）	13.2	86.5	138.3
3（部位）	7.2	73.5	273.6
4（部位）	7.0	76.2	220.4

次に処理過程 2 における時間的連続性による効果であるが，これは Data 3 の方が Data 1 及び Data 2 よりも精度・計算量の両面で有効であることがわかる．これは，Data 1，Data 2 の頭部動作に対して，前後のフレームによる拘束が大きすぎたためであると考えられる．また，この過程においても Data 4 に関して精度の向上が見られ，86.5%の頭部領域抽出率が得られた．

処理過程 3 以降は，部位抽出過程であるが，大まかな頭部領域抽出に成功していれば必然的に抽出が容易になる．しかし，頭部動作の回転により，見かけの画像として遮へいが生じる Data 1 及び Data 2 においては Data 3 と比較して部位抽出が困難であることがわかる．

最後に処理過程 4 であるが，Data 1，Data 2，Data 3 において，部位自体の変形あるいは見かけ上での変形を学習しているため，過程 3 において得られた結果よりも向上が見られる．また，撮影条件に特別な拘束のない Data 4 に関してはわずかな向上しか確認できなかったが，76.2%の部位領域抽出率が得られた．

5. むすび

本論文では，従来の静止画に対する Labeled Graph

Matching 法の概念を拡張し，新しい動画像認識の枠組みに関して述べた．物体は様々なレベルで大局的に安定した構造を保ちある関連性をもちながら，局所的にはそれとは独立に運動するという観点から，我々は物体の構造と運動を局所的大局的にとらえながら認識する手法を提案した．様々な情報を様々なレベルで柔軟に扱うためには，処理の段階に応じて扱う情報を効果的に制御する必要があり，本論文では人物頭部領域抽出に本手法を適用することでその効果に関する考察を行った．その結果，特別な制限を設けないで撮影したデータに対しても頭部領域抽出で 86.5%，部位領域抽出で 76.2%という高い成功率が得られることを確認した．

本手法の応用範囲は広く，今後は，個人認証及び表情認識に対して本手法の効果を確認していきたい．また，本論文では問題を簡単にするため扱う情報を制御するパラメータを 2 値としたが，より複雑な認識過程の場合はより柔軟に処理をする必要がある．よって，パラメータ制御法も今後の課題として挙げられる．

文　献

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," J. Cognitive Neuroscience, vol.3, no.1, pp.71–86, 1991.

[2] J. Ahang, Y. Yan, and M. Lades, "Face recognition: Eigenface, elastic matching, and neural nets," Proc. IEEE, vol.85, no.9, pp.1422–1435, 1997.

[3] I.A. Essa and A.P. Pentland, "Facial expression recognition using a dynamic model and motion energy," Proc. ICCV, pp.360–367, 1995.

[4] M. Lades, J.C. Vorbruggen, J. Buhmann, J. Lange, C.v.d. Malsburg, R.P. Wurts, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," IEEE Trans. Comput., vol.42, no.3, pp.300–311, 1993.

[5] J. Bhumann, J. Lange, and C.v.d. Malsburg, "Distortion invariant object recognition by matching hierarchically labeled graphs," Proc. IJCNN Int. Conf. Neural Networks, pp.155–159, IEEE, 1989.

[6] W. Konen, T. Maurer, and C.v.d. Malsburg, "A fast dynamic link matching algorithm for invariant pattern recognition," Neural Network, vol.7, pp.1019–1030, 1994.

[7] J.L. Barron, D.J. Fleet, and S.S. Beauchemin, "Performance of optical flow techniques," IJCV, 1994.

[8] C. Fennema and W. Thompson, "Velocity determination in scenes containing several moving objects," CGIP9, pp.301–315, 1979.

[9] B.K.P. Horn and B.G. Schunck, "Determining optical flow," AI17, pp.185–204, 1981.

[10] H.H. Nagel, "Displacement vectors derived from second-order intensity variations in image sequences," CGIP21, pp.85–117, 1983.

[11] F. Glazer, G. Reynolds, and P. Anandan, "Dcene matching through hierarchical correlation," Proc. IEEE CVPR, pp.432–441, 1983.

[12] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," Int. J. Comp., Vision2, pp.283–310, 1989.

[13] B.D. Lucas, Generalized image matching by the method of differences, Ph.D. Dissertation, Dept. of Computer Science, Carnegie-Mellon University, 1984.

[14] B. Lucas and T. Kanade, "An iterative image registration technique with an application too stereo vision," Proc. DARPA IU Workshop, pp.121–130, 1981.

（平成 13 年 11 月 8 日受付，14 年 4 月 1 日再受付）

村上　真　（正員）

1997 早大・理工・情報卒．1999 同大大学院修士課程了．2002 東洋大・工・情報・専任講師．画像認識，ヒューマンインタフェース研究に従事．情報処理学会，IEEE 各会員．

米山　正秀　（正員）

1962 慶大・工・電気卒．同年，日本コロムビア（株）入社．画像工学，音響工学の研究に従事．1982（株）リコー入社．音声情報処理，超音波イメージング，ニューラルネットワーク応用研究を手掛ける．1995 東洋大・工・情報・教授．この間，1979〜1980 カリフォルニア大（サンタ・バーバラ）客員研究員．1979〜1990 慶大・工・非常勤講師．工博．1993 海洋音響学会論文賞受賞．日本音響学会，情報処理学会，人工知能学会，ASA，IEEE 各会員．

白井　克彦　（正員）

1963 早大・理工・電気卒．1968 同大大学院博士課程了．同年，電気工学科専任講師．1970 同助教授．1975 同教授．1991 情報学科教授．1998 常任理事・副総長．工博．音声認識・合成技術，自然言語処理，信号処理向けアーキテクチャ設計，ヒューマンインタフェースの研究に従事．情報処理学会，日本音響学会，IEEE 等各会員．

論　文

感情語による歩行動作制御を目的としたディジタルアクタシステムの構築

村上　真[†]　森野　岳宏[††]　長船　大毅[††]　誉田　雅彰[††]
白井　克彦[††]

Costruction of Virtual Actor System to Control its Walking Motion with Emotion from Language Instruction

Makoto MURAKAMI[†], Takehiro MORINO[††], Daiki OSAFUNE[††], Masaaki HONDA[††], and Katsuhiko SHIRAI[††]

あらまし　個人レベルでキャラクタアニメーションを制作するには，直感的で容易なインタフェースにより多様な動作が生成可能なシステムが必要となる．本研究では，監督が俳優に言語で指示を与えるような自然な方法で動作制御を行うことが有効であると考え，感情語による指示により歩行動作を制御可能なディジタルアクタシステムを提案する．感情語による指示から動作を生成・制御するためには，様々な感情語に対応した動作データを収集し，感情語—動作間の関係を記述した感情表出モデルを構築する必要がある．本研究では，感情語の多様性を保証する形でデータ収集を行い，得られたデータから，多様な感情語相互の関係を木構造として表現することで感情表出モデルを構築する．また，多様性を持たせてモデルを構築したとしても，未知語による指示が出現することは避けられない．そのため，未知語が入力された場合に新たに感情を学習し，感情表出モデルを更新する感情学習の機能についても検討する．最後に，感情表出部と感情学習部から成るディジタルアクタシステムを実装し，評価を行う．

キーワード　キャラクタアニメーション，感情語，感情表出，動作生成，動作制御

1. まえがき

近年，ディジタルコンテンツの流通に伴い，映像制作における3DCGアニメーションの使用頻度が増加している．また，家庭用PCの性能向上や利用者層の拡大により，個人レベルでの3DCGアニメーション制作の需要も高まっている．しかし，3DCGアニメーションを用いた映像コンテンツの中でも，主要な位置を占めるキャラクタアニメーションにおいては，まだ個人レベルの制作は困難である．例えば，従来からキャラクタアニメーション制作に利用されてきたモーションキャプチャは，機材の設置場所の確保やコストの面から個人で利用できるものではない．また，キーフレームアニメーションや，数値入力やプログラミングによるアニメーションなども，利用にあたり熟練した技能が必要であり，直感的にキャラクタアニメーションを制作することは困難である．

そこで，多様な動作を生成・制御可能な技術に関する研究が活発に行われている．多様な動作を生成する手法としては，身体の末端部のいくつかを制御することで，その他の関節を自動で制御することが可能な逆運動学を用いたもの，人間の筋肉を考慮した解剖学的に正当なモデルであるアクションユニットを用いたもの[1][2]，ラバノテーションという人物動作の記述法を用いて動作を記号によって表現するもの[3]等がある．また，基準となる動作を制御することで多様な動作を生成する手法としては，歩行動作のような周期的な運動を三角関数により近似し，その関数のパラメータ修飾により制御を行うもの[4]等がある．さらに，動作を感情に関する要素によって制御する手法としては，歩行動作を基準歩行動作と感情つき動作の特徴を表す

† 東洋大学工学部，埼玉県
Faculty of Engineering, Toyo University, Saitama, 350-8585 Japan
†† 早稲田大学理工学部，東京都
School of Science and Engineering, Waseda University, Tokyo, 169-8555 Japan

特徴量成分に分離し，様々な特徴量成分を重みつき線形結合することにより制御を行うもの [5] [6] 等がある．しかし，監督が俳優に言語で指示を与えるような自然な方法で動作制御を可能とするシステムはない．そこで本研究では，コンピュータ内のキャラクタ (アクタ) に対してユーザが言語で指示を与えることで，動作制御を行うことのできるディジタルアクタシステムを提案する．

言語を用いた指示から動作を生成・制御するためには，言語 (概念)—動作間の関係を明らかにしなければならない．言語による指示には,「手を上げて」といった動作生成を要求するもの,「大きく手を振って」といった物理的な変動を要求するもの,「楽しそうに」といった感情に関係する変動を要求するもの等，様々ななものが存在する．本研究では，対象を歩行動作に限定し，主に感情に関する変動を要求する動作指示 (感情語) により動作制御を行うことを目的とする．したがって，感情語—動作間の関係を表現する感情表出モデルの構築を行う．

本研究では，感情を，喜び・悲しみ・怒り・驚きといったいくつかのカテゴリで表現されるものとしては扱わず，より多様なものとして捉える．したがって，いくつかのカテゴリの感情動作を収録して分析することによりモデルを構築するという手法はとらず，以下に示す方法でモデル構築用データの収集を行う．まず，被験者に目的動作を想像してもらい，自由に感情語による指示を行ってもらう．次に，多様な歩行動作を生成可能な歩行動作制御システムを使用して，被験者に想像した通りの動作を生成してもらう．得られた感情語と生成された動作の対応関係から，多次元動作空間中に感情語を表現する．さらに，この多次元空間から適切な動作を表出させるためには，空間中に分布する多様な感情語相互の関係を明らかにする必要がある．本研究では，それらを木構造として表現することで感情表出モデルを構築する．

また，多様性を持たせてモデルを構築したとしても，未知語による指示が出現することは避けられない．そのため，未知語が入力された場合に新たに感情を学習し，感情表出モデルを更新する感情学習の機能についても検討を行う．

最後に，感情表出モデルに従い動作を制御する感情表出部と，未知語が入力された場合に感情表出モデルを更新する感情学習部を実装したディジタルアクタシステムを構築し，評価を行う．

2. 歩行動作制御システム

本章では，感情表出モデルの構築に使用する動作データを収集するための歩行動作制御システムについて述べる．

歩行動作制御システムの要求定義は,「普通に歩く」を初期状態とし，ユーザが想像している目的動作を正確かつ容易に作成可能なことである．そこで，少ない操作で多様な歩行動作を生成・制御可能なモデル (歩行動作制御モデル) を構築する必要がある．本研究で提案する歩行動作制御システムは，関節構造モデルと動作制御モデルの 2 つのモデルからなる．関節構造モデルとは，歩行動作において個々の関節の動作がどのように連動しているのかをグループとして構造化したものである．このモデルにより，同時に制御する部位を少なくすることができる．動作制御モデルとは，物理的な特徴を表す複数の特徴量成分を保持し，それらを関節構造モデルに基づいて様々に付加する事で，多様な歩行動作を制御可能にしたモデルである．歩行動作を生成・制御する際には，動作制御モデルが保持する特徴量成分を選択し，関節構造モデルにおける対応制御部位に，その特徴量成分に対して重みを掛けて付加する事になる．

2.1 モデル構築用動作データ

関節構造モデル及び動作制御モデルの構築に使用する歩行動作データをモーションキャプチャ装置で収録する．収録歩行動作は以下に示す 19 種類 20 人分である．

- 速く歩く
- 遅く歩く
- 俯いて歩く
- 胸を張って歩く
- 腕を大きく振って歩く
- 細かく歩く
- 大股で歩く
- 細かく歩く
- 腕を外向きにして歩く
- 腕を内向きにして歩く
- 蟹股で歩く
- 内股で歩く
- 縦に揺れながら歩く
- 横に揺れながら歩く
- 捻りながら歩く
- 肘を曲げながら歩く

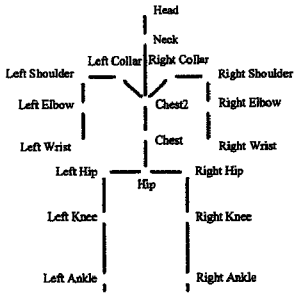


図 1　人体モデル
Fig. 1　Human body model

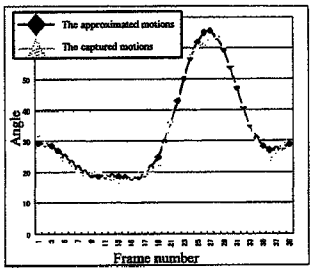


図 2　第 3 高調波までを用いて生成した動作データ
Fig. 2　Approximated motion by up to the third harmonic

- 頭を前に落として歩く
- 上を向きながら歩く
- 腿を大きく振りながら歩く

尚，本研究では，図 1 に示す階層化された 19 の関節からなる人体モデルを使用している．

2.2　関節構造モデルの構築

人体モデルは 19 関節で定義されているため，分析対象データは 19 関節× 3 軸方向＝ 57 関節動作から構成される．本節では，重回帰分析を用いて連動性を基に，この 57 関節動作を構造化する方法について述べる．

（1）　スペクトル情報への集約

モーションキャプチャデータから得られる離散時間関節動作信号 Θ_{mn} を式 (1) に示す離散フーリエ変換により周波数領域に変換する．

$$C_{mk} = \sum_{n=0}^{N-1} \Theta_{mn} e^{-i\frac{2\pi nk}{N}} \qquad (1)$$

ここで，m は関節動作番号 ($0 \leqq m \leqq 56$) を，k はスペクトルの高調波次数を表している．よって，$|C_{mk}|$ は関節動作番号 m の関節動作信号の第 k 高調波のスペクトルを表す．

過去の研究より，人物動作における関節動作は全て第 3 高調波までのスペクトル情報で十分再現可能であることが確認されている [4]．図 2 に，右膝の前後運動における第 3 高調波までの情報によって生成した動作データの波形と元動作データの波形を示す．図 2 より，本研究で使用するデータに対しても，第 3 高調波までの情報でほぼ元動作データが再現できていることがわかる．

時間領域において各関節動作を制御する場合，1 周期のフレーム数に相当する次元数のパラメータを制御する必要があるが，第 3 高調波までのスペクトルに情報を集約することで，制御する情報を簡略化することができる．そこで本研究では，全関節動作の時系列データを第 3 高調波までのスペクトル情報に集約しモデルを構築する．

（2）　連動性のルールセットの導出

関節動作間の連動性を明らかにするために，重回帰分析を用いて連動性のルールセットを導出する．まず，他の関節動作と連動性が高いと推測される関節動作をキー関節動作とする．次に，説明変数にキー関節動作のスペクトル情報を，目的変数にキー関節動作以外の関節動作のスペクトル情報を使用して，前者のデータから後者のデータを生成する関数群を重回帰分析によって導出する．この関数群を関節動作間の連動性のルールセットとする．

（3）　関節動作の構造化

LeftHip, RightHip, LeftShoulder, RightShoulder の関節動作をキー関節動作 I として，他の関節動作との連動性のルールセットを導出した．次に，この連動性のルールセットの導出に使用していない歩行動作データのキー関節動作 I のスペクトル情報から，連動性のルールセットによりキー関節動作以外の関節動作データを導出する．この動作データと元動作データとの相関関係から，各関節ごとの動作の再現度がわかる．15 人分の収録歩行動作に対して，関節動作ごとの相関係数の平均値を再現度とし，キー関節動作 I との連動性の評価を行った．

表 1 に，分析結果の 1 つである右脚部位における再現度 (相関係数の平均) を示す．表 1 から，再現度の高い関節動作と再現度の低い関節動作があることがわかる．

ここで，再現度が 0.5 以上の関節動作を従属関節動作 I と設定する．従属関節動作 I は，連動性のルール

表 1 再現度表 - 右脚部位
Table 1 The averages of the correlation coefficients of the right leg joint motions

関節動作名	再現度
RightKnee-Z	0.8553
RightKnee-X	0.9410
RightKnee-Y	0.7048
RightAnkle-Z	0.6801
RightAnkle-X	0.8541
RightAnkle-Y	0.3531

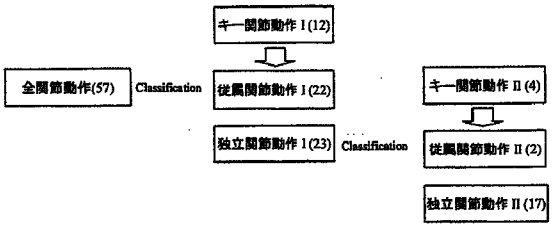


図 3 関節動作の分類図
Fig. 3 Classification of the joint motions

セットから生成可能であることから，キー関節動作 I との連動性が高いと推測される．つまり，従属関節動作 I はキー関節動作 I の変化に伴って規則的に変化するものとして扱うことができる．

また，再現度が 0.5 より小さい関節動作を独立関節動作 I と設定する．独立関節動作 I は，連動性のルールセットから生成困難であることから，キー関節動作 I との連動性が低いと推測される．つまり，独立関節動作 I はキー関節動作 I の変化に伴って規則的に変化しないと考えられる．

次に，独立関節動作 I の中で，また新たにキー関節動作 II を設定し，同様の手法で関節動作の分類を行う．これにより，独立関節動作 I を，キー関節動作 II，従属関節動作 II 及び独立関節動作 II に分類する．

これまでに分類した関節動作を図 3 に示す．全関節動作 57 動作の中で，歩行動作において連動して制御される関節動作は 40 動作 (キー関節動作 I・II，従属関節動作 I・II) あり，独立に変化する関節動作は 17 動作 (独立関節動作 II) あった．

次に，各関節動作を連動して制御される関節群に分類し，歩行動作における関節動作の構造を明確化した．分類された結果を図 4 に示す．

関節動作は大きく分けて，互いに高い連動性を保持しながら運動している関節動作群と，そうでない関節動作群 (グループ 5) の 2 種類に分類される．そして，連動性の高い関節動作は更に 4 つのグループ (グループ 1, 2, 3, 4) に分類される．

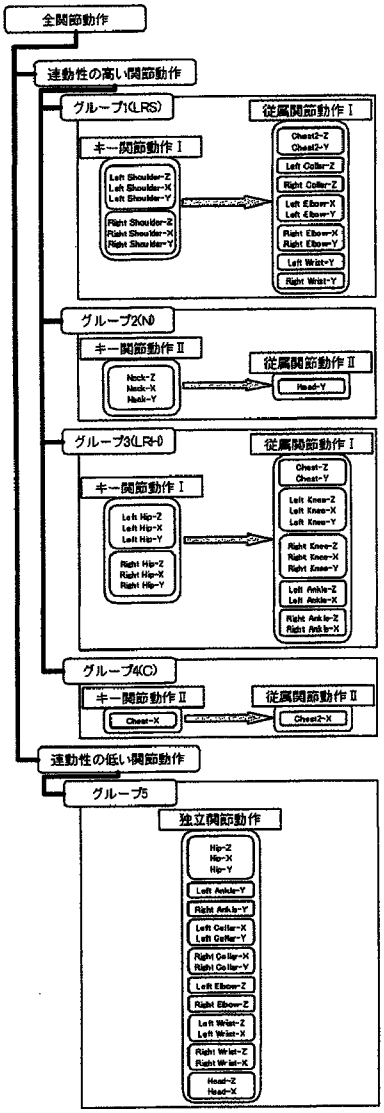


図 4 関節動作グループ構造
Fig. 4 Structure of the joint motions

2.3 動作制御モデルの構築

（1） 特徴量成分の抽出

本研究では，歩行動作を制御する際に重要な動作制御項目として，「速さ」，「姿勢」，「腕の振り」，「脚の振り」，「腕の回転」，「脚の回転」，「縦揺れ」，「横揺れ」，「捻り」，「肘の曲げ」，「頭の姿勢」，「腿の振り」の 12 項目を取り上げ，各項目に対応した物理的な特徴を保持している収録歩行動作から，その特徴を表す特徴量成分を抽出する．表 2 に動作制御項目と抽出する特徴量成

表 2　動作制御項目と特徴量成分
Table 2　Walking motion data for every feature

動作制御項目	特徴量成分　(抽出対象歩行動作)	式
速さ	速く歩く	$\delta^{+}_{mk0}, \lambda^{+}_{mk0}$
	遅く歩く	$\delta^{-}_{mk0}, \lambda^{-}_{mk0}$
姿勢	俯いて歩く	$\delta^{+}_{mk1}, \lambda^{+}_{mk1}$
	胸を張って歩く	$\delta^{-}_{mk1}, \lambda^{-}_{mk1}$
腕の振り	腕を大きく振って歩く	$\delta^{+}_{mk2}, \lambda^{+}_{mk2}$
	細かく歩く	$\delta^{-}_{mk2}, \lambda^{-}_{mk2}$
脚の振り	大股で歩く	$\delta^{+}_{mk3}, \lambda^{+}_{mk3}$
	細かく歩く	$\delta^{-}_{mk3}, \lambda^{-}_{mk3}$
腕の回転	腕を外向きにして歩く	$\delta^{+}_{mk4}, \lambda^{+}_{mk4}$
	腕を内向きにして歩く	$\delta^{-}_{mk4}, \lambda^{-}_{mk4}$
脚の回転	蟹股で歩く	$\delta^{+}_{mk5}, \lambda^{+}_{mk5}$
	内股で歩く	$\delta^{-}_{mk5}, \lambda^{-}_{mk5}$
縦揺れ	縦に揺れながら歩く	$\delta^{+}_{mk6}, \lambda^{+}_{mk6}$
横揺れ	横に揺れながら歩く	$\delta^{+}_{mk7}, \lambda^{+}_{mk7}$
捻り	捻りながら歩く	$\delta^{+}_{mk8}, \lambda^{+}_{mk8}$
肘の曲げ	肘を曲げながら歩く	$\delta^{+}_{mk9}, \lambda^{+}_{mk9}$
頭の姿勢	頭を前に落として歩く	$\delta^{+}_{mk10}, \lambda^{+}_{mk10}$
	上を向きながら歩く	$\delta^{-}_{mk10}, \lambda^{-}_{mk10}$
腿の振り	腿を大きく振りながら歩く	$\delta^{+}_{mk11}, \lambda^{+}_{mk11}$

分を示す．

特徴量成分を抽出する手法としては，鵜沼ら [5] や Kraus [6] の手法が適用できる．これらは「楽しく歩く」などの感情を持った歩行動作データと「普通に歩く」の歩行動作データのスペクトルの差分を算出する事で，感情の成分を数量化したものである．本研究では，物理的な特徴を表す特徴量成分を抽出する為にこの手法を用いる．ただし，抽出した特徴量成分に多様性を付加する為に，対象歩行動作データの分散を考慮する．

以下に抽出方法を述べる．まず，基準歩行動作「普通に歩く」のスペクトルデータ $\mu^{(S)}_{mk}$ を，モーションキャプチャ装置で収録した 20 人分の「普通に歩く」のスペクトルデータを平均する事で算出する．また，その分散 $\left\{\sigma^{(S)}_{mk}\right\}^2$ を算出する．ここで，m は関節動作番号 $(m = 0, \cdots 56)$ を，k はスペクトルの高調波次数 $(k = 0, \cdots, 3)$ を表す．

同様に，物理的な特徴を保持している 19 の歩行動作のスペクトルデータ $\mu^{(P)}_{mkl}$ を，モーションキャプチャで収録した 20 人分の歩行動作のスペクトルデータの平均から算出する．また，その分散 $\left\{\sigma^{(P)}_{mkl}\right\}^2$ を算出する．ここで，l は特徴量成分の番号 $(l = 0, \cdots, 18)$ である．

次に，特徴量成分を式 (2) により算出する．

$$\delta_{mkl} \equiv \mu^{(P)}_{mkl} - \mu^{(S)}_{mk} \tag{2}$$

また，動作制御の多様化を図る為に，式 (2) で算出した特徴量成分に，実測データの分散からなる変動 λ_{mkl} を加える．これは，正規分布 $N\left(\mu^{(P)}_{mkl}, \sigma^{(P)}_{mkl}\right)$, $N\left(\mu^{(S)}_{mk}, \sigma^{(S)}_{mk}\right)$ に乱数を与えて算出される $\mu^{(P)'}_{mkl}$, $\mu^{(S)'}_{mk}$ を用いて，式 (3) により算出する．

$$\delta_{mkl} + \lambda_{mkl} \equiv \mu^{(P)'}_{mkl} - \mu^{(S)'}_{mk} \tag{3}$$

（2）　動作制御パラメータを用いた動作制御

動作を制御する際には，基準歩行動作「普通に歩く」のスペクトルデータに対し，特徴量成分に分散による変動を加えて重みを掛け合わせたものを付加する．

よって，この重み (動作制御パラメータ) を変化させる事で，多様な歩行動作を生成・制御する事が可能になる．動作制御パラメータは動作制御項目毎に設定する．なお，2 つの特徴量成分を保持する動作制御項目においては，パラメータを + に移動させた場合には表 2 の上の特徴量成分 $(\delta^{+}_{mkr}, \lambda^{+}_{mkr})$ を，− に移動させた場合は下の特徴量成分 $(\delta^{-}_{mkr}, \lambda^{-}_{mkr})$ を用いる．ここで，r は動作制御項目の番号 $(r = 0, \cdots, 11)$ を表す．

したがって，動作制御によって生成される歩行動作の各関節動作データは式 (4) によって算出される．

$$\Theta_{mn} = \sum_{k=0}^{3}\left[\mu^{(S)}_{mk} + \sum_{r=0}^{11}\left\{w_r \left(\delta^{+or-}_{mkr} + \lambda^{+or-}_{mkr}\right)\right\}\right] e^{i\frac{2\pi nk}{N}} \tag{4}$$

ここで，w_r は動作制御パラメータを表す．

ただし，関節構造モデルにより，歩行動作における関節動作は連動性を基に幾つかのグループに分類される事から，特徴量成分の付加を，その特徴量成分に影響する関節動作グループのみに対して行う．さらに，動作制御項目に応じて，付加する特徴量成分のスペクトルデータの高調波成分 $(f0, f1, f2, f3)$ を取捨選択する．これにより，動作制御項目間の独立性を保ち，より正確な歩行動作を生成・制御することが可能となる．例えば，「脚の振り」の動作制御項目は，脚の部位の動きに影響しているので，脚の部位に多くの関節動作が見られる関節動作グループ 3(LRH) のみに，また，スペクトル成分の内波形に影響する $(f1, f2, f3)$ 成分のみに特徴量成分を付加する事になる．

表 3 に動作制御項目とそれに対応する関節動作グループ及びスペクトル成分を示す．なお，グループ 5(独立関節動作) は他の関節動作と連動しない事から，

表 3 動作制御項目と関節動作グループ及びスペクトル成分
Table 3 Group of joint motions and harmonic component controled for every feature

動作制御項目	関節動作グループ	スペクトル成分
速さ	グループ 1, 2, 3, 4	$f0, f1, f2, f3$
姿勢	グループ 1, 2, 3, 4	$f0$
腕の振り	グループ 1, 2	$f1, f2, f3$
脚の振り	グループ 3	$f1, f2, f3$
腕の回転	グループ 1	$f0$
脚の回転	グループ 3	$f0$
縦揺れ	グループ 2, 4	$f1, f2, f3$
横揺れ	グループ 1, 2, 3, 4	$f1, f2, f3$
捻り	グループ 1, 2, 3, 4	$f1, f2, f3$
肘の曲げ	グループ 1	$f0$
頭の姿勢	グループ 2	$f0$
腿の振り	グループ 3	$f0, f1, f2, f3$

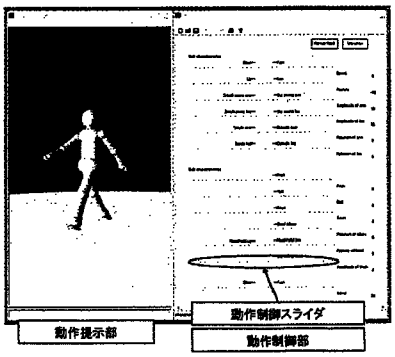


図 5 歩行動作制御システムのインタフェース
Fig. 5 Walking motion control system interface

どの特徴量成分を付加する際にも無視する.

2.4 歩行動作制御システムの実装

関節構造モデル及び動作制御モデルを基に歩行動作制御システムを実装した. インタフェースは図 5 に示すように, 動作提示部と動作制御部からなり, 動作制御部は 12 の動作制御パラメータと 1 周期のフレーム数を制御可能な 13 の動作制御スライダからなる.

2.5 歩行動作制御システム評価実験

本節では, 実装した歩行動作制御システムの評価実験について述べる.

2.5.1 実験方法

実験手法は, 被験者に感情歩行動作 (怒って歩く, 喜んで歩く, 悲しく歩く) を頭の中で想像してもらい, 想像された動作を目標動作として, 目標動作に近づけるようシステムを用いて歩行動作を生成してもらう. 次に, 生成した歩行動作を頭の中で想像した目標動作にどれだけ似せる事が出来たかという満足度としての動作再現度を, 5 段階評価のアンケート (5=非常に似ている, 4=似ている, 3=どちらでもない, 2=似ていない, 1=全く似ていない) で答えてもらう.

表 4 システム評価実験結果
Table 4 The result of the evaluation experiment

歩行動作	動作再現度
怒って歩く	4.2
喜んで歩く	4.6
悲しく歩く	4.6
平均	4.3

2.5.2 結果と考察

5 人の被験者に行ったシステム評価実験の結果を表 4 に示す.

評価実験結果によると, どの歩行動作においても 5 段階評価指標の 4(似ている) より高く, 本システムの動作再現度が高い事が解る.

3. 感情表出モデルの構築

本章では, 感情語—動作間の関係を表現する感情表出モデルの構築について述べる.

3.1 データ収集

感情語を用いた指示により動作を生成・制御するためには, 感情語—動作間の関係を記述した感情表出モデルを構築しなければならない. したがって, 感情表出モデルの構築には, 様々な感情語に対応した動作データを収集する必要がある. 直接的な方法としては, 感情語を指定し, それに従って被験者に実際に動作をしてもらうことで動作データを収集する方法が考えられるが, 指定した感情以外の動作データの収集が困難であるという欠点がある. そこで本研究では, 被験者が目的動作を想像し, それに適した感情語を指定し, 第 2. 章で示したシステムを用いて想像した通りの動作を生成するという方法で, 間接的にデータ収集を行う. このように被験者に自由な動作指示を提示してもらうことにより感情の多様性を保証することができる.

本研究で行った動作収集の手順を以下に示す.

(1) 状況提示

全く何も指定されない状態で目的動作を被験者に想像してもらうことは困難であると考え, 被験者の想像力が高まるような状況を予め用意し提示する. これは, 台本に与えられた状況から言語による指示により動作を生成・制御するディジタルアクタシステムの仕組みにも対応している. また, 本研究では歩行動作のみを対象にするため, 状況は全て歩行動作に適用可能なものを選んでいる.

表 5　収集した感情語
Table 5　Frequent instructions

出現数	感情語
16	とぼとぼ　ビクビクしながら
14	ルンルン
13	いらいらして　肩を落として
11	むかつきながら
10	うきうきして
9	がっくりしながら
7	がっかりして　逃げるように
6	怒って　胸を張って　ワクワク
5	うなだれながら　おそるおそる　おどおど こそこそ　慎重に
4	うつむいて　ドキドキしながら 早足で　へこんで

（2）　動作想像

被験者に (1) で与えられた状況に対して目的動作を想像してもらう．

（3）　感情選択

被験者に (2) で想像した目的動作を予め設定した基本 4 感情のいずれかに当てはめてもらう．基本 4 感情には，過去に心理学の研究から得られた 8 感情の内，歩行動作に付加し易い「歓喜」「悲嘆」「激怒」「恐怖・警戒」を選択した．このデータは，本研究で構築する感情表出モデルにおいて，従来の基本感情がどのように分布するかを考察するために使用するもので，感情表出モデルの構築には直接使用しない．

（4）　感情語の付与

被験者に (2) で想像した目的動作に対して感情語を付与してもらう．

（5）　動作生成

被験者に (2) で想像した目的動作を第 2. 章で示した歩行動作制御システムを用いて生成してもらう．

男子大学生 7 人を対象とし，各 60 状況についてデータ収集を行なった．今回の実験で特に多く出現した感情語は「びくびく」「とぼとぼ」「肩を落として」「ルンルン」「いらいらして」「うきうき」「がっくり」等であり，これらは，歩行動作に対してイメージの沸き易い感情 (概念・語彙) であると考えられる．収集した感情語のうち，出現頻度が高かったものを表 5 に示す．

3. 2　モデル構築方法

3. 1 節で収集したデータを元に，多次元動作空間に感情語を表現することができる．各感情語に対応した動作を生成するには，多次元動作空間中の 1 点に対応する動作データを生成すればよい．しかし，適切に動作を制御するためには，各感情語間の関係を明らかにする必要がある．本研究では，動作空間における距離関係に着目し，クラスター分析により各感情語を木構造として表現する．

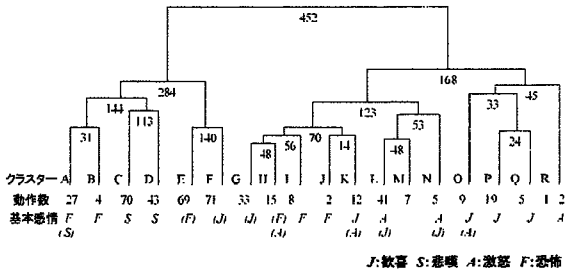


図 6　クラスター分析結果
Fig. 6　The result of cluster analysis

使用する動作データは，3. 1 節で収集した 420 動作である．各動作データから離散フーリエ変換によりパワースペクトルを求め，第 3 高調波までを特徴量として抽出する．また，全 57 関節動作のうち，2. 2 節で示した独立関節動作を除いた計 40 関節動作を使用する．したがって特徴量は，1 動作につきパワースペクトル 4× 関節動作数 40=160 次元からなる．

クラスター分析には Ward 法を使用し，距離の近い動作データをつなぎ続け最終的に 2 分木を構成する．また，距離尺度には標準化ユークリッド距離を用いる．

3. 3　モデル構築結果

本研究では，A～R までの計 18 個のクラスターになるように 2 分木を構成した．結果を図 6 に示す．ツリーの高さは各クラスターの非類似度に対応している．図中の数字は，そのクラスター内 (又はツリーの下位) に存在する動作数である．また，斜字体の記号はそれぞれ Joy・Sadness・Anger・Fear の頭文字であり，3. 1 節で述べた動作収集手順 (4) で被験者に選択してもらった基本感情の「歓喜」「悲嘆」「激怒」「恐怖・警戒」に対応している．各記号は，そのクラスター内において過半数を占める基本感情を示しており，3 分の 1 以上半数未満の場合には括弧書きで示している．

（1）　基本感情の分布

各クラスターにおける基本感情の割合を表 6 に示す．尚，括弧内の数字は動作数を表している．

図 6 及び表 6 から，まず，「歓喜」の感情を持つ動作が多く含まれるクラスターは K, O, P, Q を中心に多数存在することがわかる．したがって，「歓喜」の感情を持つ動作には多くのバリエーションが存在することが確認できる．しかし，全てのクラスターに満遍なく存在するのではなく，各クラスターにおける存在確率

表 6 各クラスター間における基本感情の割合
Table 6 The ratio of fundamental emotion in every cluster

クラスター	歓喜 [%]	悲嘆 [%]	激怒 [%]	恐怖・警戒 [%]
A	0.0 (0)	40.7 (11)	0.0 (0)	59.3 (16)
B	0.0 (0)	25.0 (1)	0.0 (0)	75.0 (3)
C	2.9 (2)	68.6 (48)	7.1 (5)	21.4 (15)
D	16.3 (7)	55.8 (24)	11.6 (5)	16.3 (7)
E	34.0 (8)	30.9 (14)	13.8 (15)	21.3 (32)
F	39.5 (28)	22.5 (16)	14.1 (10)	23.9 (17)
G	39.4 (13)	6.1 (2)	24.2 (8)	30.3 (10)
H	20.0 (3)	0.0 (0)	40.0 (6)	40.0 (6)
I	12.5 (1)	12.5 (1)	25.0 (2)	50.0 (4)
J	0.0 (0)	0.0 (0)	0.0 (0)	100.0 (2)
K	58.3 (7)	0.0 (0)	41.7 (5)	0.0 (0)
L	39.0 (16)	2.4 (1)	53.7 (22)	4.9 (2)
M	14.2 (1)	28.6 (2)	28.6 (2)	28.6 (2)
N	40.0 (2)	0.0 (0)	60.0 (3)	0.0 (0)
O	66.7 (6)	0.0 (0)	33.3 (3)	0.0 (0)
P	100.0 (19)	0.0 (0)	0.0 (0)	0.0 (0)
Q	100.0 (5)	0.0 (0)	0.0 (0)	0.0 (0)
R	16.7 (2)	0.0 (0)	75.0 (9)	8.3 (1)

に偏りがあることがわかる.

次に「悲嘆」についてはC, Dを中心に密集して存在し，さらに「悲嘆」の感情を持つ動作のほとんどは最初の分岐よりも左に存在していることがわかる．したがって,「悲嘆」の感情を持つ動作は「歓喜」ほど多くのバリエーションは持たないと考えられる.

「激怒」の感情を持つ動作はL, N, Rを中心にいくつか存在し，ばらつきも大きい．よって,「激怒」の感情を持つ動作は非常にバリエーションが大きいと考えられる.

最後に「恐怖・警戒」の感情を持つ動作が多く含まれるクラスターはA, B, I, Jを中心にいくつか存在する．また「恐怖・警戒」は「悲嘆」の近辺に多く存在する傾向があり，動作自体も「悲嘆」と近いと推測される．また，動作の分散は「悲嘆」よりもやや広く，多少のバリエーションがあると考えられる.

（2） 感情語の分布

各感情語における所属クラスターの割合を表7に示す．ただし，この表は同一感情語が特に多く出現したもののうち上位9種 (6個以上出現) のみに対するもので，クラスター欄には各感情語について2個以上出現したクラスターのみを記述している．例えば,「びくびく」という感情語を持つ動作は，全ての「びくびく」の動作の内，25.0%がクラスターFに，18.8%がクラスターC及びEのそれぞれに，12.5%がクラスターHに所属しているということである.

各感情語の内,「とぼとぼ」「うきうき」には，特定ないしは近隣のクラスターに集中する傾向が見られた．しかし,「びくびく」「いらいら」については，非常にバラツキが大きく，傾向はほとんど確認できなかった．以上より，感情語単位で見た場合，人によるイメージの違いが大きい感情語と小さい感情語が存在することがわかる.

表 7 各感情語における所属クラスターの割合
Table 7 The ratio of cluster to which every frequent instruction belong

感情 (概念) 語	クラスター (割合 [%])
とぼとぼ	C(62.5) D(18.8) E(12.5)
びくびくしながら	F(25.0) C,E(18.8) H(12.5)
ルンルン	P(28.6) F,L(21.4)
いらいらして	L(23.1) E,F,J,H(15.4)
肩を落として	C(30.8) D,F(23.1)
むかつきながら	F(27.3) L,R(18.2)
うきうきして	L(30.0) D,P(20.0)
がっくりしながら	D(44.4) C(33.3) E(22.2)
がっかりして	C(42.9) F(28.6)
逃げるように	E,F,G(28.6)

4. ディジタルアクタシステムの実装

本章では，第3.章で示した感情表出モデルを用いて動作を表出するディジタルアクタシステムについて述べる．本研究で提案するシステムは，木構造で表現されている感情表出モデルから目的動作を生成・制御する感情表出部と，未知語が入力された場合等に新たに感情語を学習し感情表出モデルを更新する感情学習部から成る.

4.1 感情表出部

提案システムの第1の目的は，ユーザの感情語による指示から目的動作を早く生成することである．したがって，木構造で表現されている感情表出モデルから目的動作を効率良く探索する必要がある．提案システムでは，入力された語が既に学習済みの語であった場合は，その語に対応する動作を表出する．ここで，対応する動作が複数存在し，それらが同一のクラスター内にある場合には，代表的な動作を1つ表出する．逆に，それらが複数のクラスターに分散して存在する場合には，各クラスターから代表的な動作をユーザに提示し選択してもらう.

提案システムの第2の目的は，目的動作を適切に生成することであるが，以上の方法で生成した動作は，予め学習されている動作そのものであるため，ユーザが意図したとおりの動作とは異なる場合がある．そこで，提案システムでは，表出された動作が満足である

か否かをユーザに判断してもらい，満足であれば動作生成を終了し，満足でなければ動作制御を行う．具体的には，提示した動作が存在しているクラスター内の動作において，分散の大きい特徴量成分を抽出し，その特徴量成分に対応する軸方向に変動を加えることで動作制御を行う．これをユーザに提示し，満足な動作が生成されるまで繰り返す．さらに不満が続く場合には，同一クラスター内では目的動作が生成不可能であると判断し，木構造において1つ上の階層に移動し，隣接するクラスター内から，同様の方法で新たに動作を生成・制御する．これを繰り返すことにより，感情表出モデルで表現される階層構造や包含関係を有効に利用し，より早く適切な動作を生成することが可能となる．

また，これを繰り返しても満足な動作が生成されない場合は，第2.章で示したシステムを使用して目的動作を生成してもらう．

4.2 感情学習部

本研究で提案する感情表出モデルは，3.1項で述べたように，感情語の多様性を保証する形で収集されたデータを元に構築されている．しかし，多様性を持たせてモデルを構築したとしても，未知語による指示が出現することは避けられない．このため，未知語が入力された場合に新たに感情語を学習し，感情表出モデルを更新する必要がある．

提案システムでは，ユーザから未知語による指示があった場合，既知語が入力されるまでユーザに語の変更を依頼する．既知語が入力された場合の動作生成は感情表出部に従い，最終的に生成された動作を最初に入力された未知語と対応させて学習を行う．また，既知語が入力された場合でも，感情表出部で動作制御を行い新しい動作が生成された場合には，既知語と生成動作を対応させて学習を行うこととなる．

学習方法は，既に学習されているデータに新たに収集されたデータを加え，改めてクラスター分析を行い，感情表出モデルを更新するという方法である．これを繰り返すことで，洗練された感情表出モデルを構築することが可能となる．

4.3 実装システム

感情表出部及び感情学習部から成るディジタルアクタシステムを実装した．インタフェースは図7に示すように，動作提示部と動作制御部からなり，動作制御部はさらに，第2.章で示した歩行動作制御システムと同様のスライダを用いたインタフェースと，言語による指示が可能なテキストボックスとボタンを用いたインタフェースからなる．

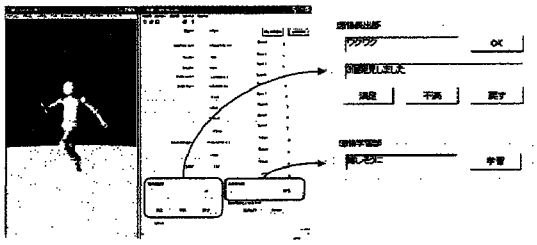


図7 ディジタルアクタシステムのインタフェース
Fig.7 Virtual actor system interface

ユーザがテキストボックスに言語による指示を入力すると，その結果が下のテキストボックスに表示される．図7は,「わくわく」と入力し「3個発見しました」と表示された例である．ユーザは「OK」ボタンを繰り返し押すことで結果の各動作を動作提示部に表示させることができる．

提示された動作がユーザの意図したとおりの動作かどうかを判断するために「満足」「不満」ボタンが用意されている.「不満」ボタンが押されると，提示された動作に対して動作制御が行われ，その結果生成された新しい動作が表示される．最終的にユーザの意図したとおりの動作が表示された場合には,「満足」ボタンを押すことにより動作の表出は終了し，必要であれば感情表出モデルが更新される．

また，ユーザが未知語を入力した場合には「Not Found!」と表示され，他の語に言い換えて入力することとなる．このとき，ユーザが入力した最初の未知語(図7の例では「嬉しそうに」)は感情学習部のテキストボックスに表示され，最終的に適切な動作が生成された後,「学習」ボタンを押すことで感情表出モデルを更新することができる．

5. ディジタルアクタシステム評価実験

本章では，第4.章で実装したシステムにより適切な動作が表出可能か否かを評価するための実験について説明する．

5.1 実験方法

評価実験の方法を以下に示す．まず被験者に目的動作を想像してもらい，実装システムを用いて目的動作を生成してもらう．次に，図8に示すように，生成した動作と感情表出モデルの各クラスターから選出した代表動作とを順次比較し，どちらが目的動作に近いか

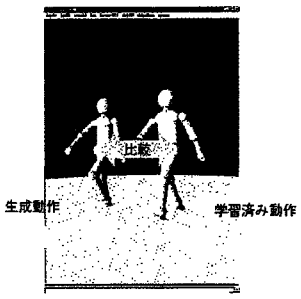


図 8 評価実験方法
Fig. 8 The evaluation experiment of virtual actor system

を評価してもらう．これにより，膨大な探索空間を全て探索する必要なく適切な動作が表出されているかどうか，あるいは，動作制御により不適切な動作が生成されていないかを評価することができる．

5.2 結果と考察

5 名の男子大学生に各 3 つの動作を生成してもらい評価を行った．各クラスターから選出した比較対象動作は 43 動作とした．その結果，生成動作より目的動作に近かった動作は 5 動作であり，表出誤り率は約 0.8%であった．

5.3 結　　論

本研究では，言語を用いた指示により，感情動作を生成・制御可能なディジタルアクタシステムの構築を目的とし，感情語—動作間の関係を木構造として表現した感情表出モデルの構築を行った．

また，感情表出モデルから目的動作を生成・制御する感情表出部と，未知語が入力された場合等に新たに感情語を学習し感情表出モデルを更新する感情学習部からなるディジタルアクタシステムを実装した．

さらに，実装システムを用いて，適切な動作が表出可能かどうかの評価実験を行った結果，表出誤り率は約 0.8%であった．

文　　献

[1] T. Komura, Y. Shinagawa, and TL. Kunii, "Muscle-based feed-forward controller of human body," Computer Graphics Forum, vol.16, no.3, pp.165-176, September 1997.

[2] T. Komura, Y. Shinagawa, and TL. Kunii, "Creating and retargetting motion by the musculoskeletal human body model," The Visual Computer, vol.5, no.16, pp.254-270, May 2000.

[3] Y. Yoshida, Y. Matsuoka, and K. Hachimura, "Processing of Human Body Movement based on Labanotation–LabanReader and LabanEditor–, IPSJ SIG-CH, vol.98, no.51, pp.61-68, 1998.

[4] 稲田和彦，佐藤潤一，中川雅通，浜田正弘，前原文雄，"関数近似と逆運動学を用いた人物動作の生成方法，" 画像電子学会研究予稿,94-01-02

[5] 鵜沼宗利，武内良三，"コンピュータアニメーションにおける感情を伴った人間の歩行動作の生成方法"，信学論 (D-II)，vol.J76-D-II，no.8，pp.1822-1831，August 1993.

[6] M. Kraus, "Human motion and emotion parameterization," CESCG 2004, April 2004.

Automatic phone segment alignment using statistical deviations from manual transcriptions

Toru Hayakawa

ATR Human Information Science Labs./Waseda University
toru@shirai.info.waseda.ac.jp

Hiroaki Kato

ATR Human Information Science Labs.
kato@atr.co.jp

Yoshinori Sagisaka

GITI Waseda University
sagisaka@giti.waseda.ac.jp

Katsuhiko Shirai

Waseda University
shirai@shirai.info.waseda.ac.jp

ABSTRACT

With the aim of achieving the automatic precise alignment of phone boundaries, the causes of automatic alignment errors were analyzed. It was quantitatively confirmed that various factors generated these alignment errors by comparing automatic alignment results with manually given boundaries. A remarkable reduction in errors was obtained by applying simple post-processing context-dependently. About 4 ms, eight times as much improvement as that achieved by speaker adaptation was obtained in RMS errors from manually given phonetic labels and about 5 ms, five times as much, was obtained in standard deviation from them.

1. INTRODUCTION

Precise temporal information is needed for the study of prosody and its application to speech technology. Although current speech recognition technologies enable quite reasonable alignment of phone segment boundaries in continuous speech with constituent phonetic sequences [2,3,4], it is also known that their boundaries are less reliable than those given manually by skilled transcribers. In this paper, with the aim of achieving the precise alignment of phone boundary segments in continuous speech, error factors in the results of automatic alignment are analyzed. As of HMM acoustic models used in automatic alignment, we adopted the standard ones used in current speech recognition technology because of its wide availability.

In section 2, we list the plausible factors producing automatic alignment errors. For quantitative analyses on automatic alignment errors, we carried out an alignment experiment using statistical speech recognition in section 3. In section 4, we analyzed alignment disagreements between automatically decided phone boundaries and manually given ones. Finally, in section 5, a simple alignment modification is applied to automatically determined boundaries to confirm the effectiveness of analyses on alignment disagreements over conventional simple improvement by speaker adapted HMM models.

2. FACTORS PRODUCING AUTOMATIC ALIGNMENT ERRORS

In automatic phone alignment using current statistical speech recognition techniques, there seem to be several factors producing disagreements in alignment between automatic and manually given phonetic labels. At the spectral feature extraction stage, windowing over 20 to 30 ms sometimes reduces time resolution as the extracted acoustic features are averaged over a given window size. In particular, at phone boundaries where the speech spectrum changes quite rapidly, the obtained spectrum provides no clear information on the spectral jump. On the other hand, as there is a big deviation in manual labels at phone boundaries where the spectrum changes slowly, this may cause big time alignment errors.

The structure of HMM acoustic models can also produce alignment errors. In standard HMMs, the minimum phone duration cannot be shorter than the frame increment period times the HMM state number. Furthermore, as most of the current statistical recognition techniques decide phone boundaries simply by maximizing the sum of likelihood value of acoustic

models rather than by detecting phone boundaries accurately, displacement of phone boundaries is not directly evaluated in the automatic alignment process. These factors seem to affect the allocation of phrase boundaries in automatic alignment differently.

3. EXPERIMENT ON AUTOMATIC ALIGNMENT

As a first step in analyzing disagreements in alignment between automatically and manually given phonetic labels, automatic alignment is carried out using standard HMM acoustic models. The experimental conditions are described in Table 1. Please note that spectral analysis (MFCC parameter extraction) was carried out over a 30 ms window with a 10 ms window shift. We used this long frame-shift simply by adopting the best practice of a current speech recognition technology in which this size of frame-shift is commonly used.

The automatic alignment results are compared to phonetic labels manually given by skilled transcribers by using spectrograms with 2.5-ms time resolutions [5]. The RMS (Root Mean Square) of alignment disagreement turned out to be 14.79 ms. Its standard deviation was 20.39 ms.

To improve alignment accuracy, HMM models were speaker-adapted by a VFS (Vector Field Smoothing) method [6] using the 503 test speech data of the target speaker. Applying speaker adaptation, the RMS value and its deviation were reduced to 14.30 ms and 19.41 ms, respectively.

These values are almost comparable to, or even smaller than those reported in previous studies [2,3]. However, they are still much bigger than the disagreements between transcribers. It has been found that disagreement between transcribers is within 5 ms for standard phone boundaries and exceeds 5 ms where spectral changes are unclear, such as vowel ends, transitions from semivowels, or palatalized consonants to vowels [5].

Table 1: Conditions for experiment on automatic alignment

Method	Forced alignment by maximizing the sum of likelihood scores of HMM models
HMM models	Gender-dependent tied state tri-phones (1400 states, 5 mixtures HMMNET [7]) trained using 503 phonetically balanced sentences uttered by 165 male speakers and 235 female speakers respectively.
Acoustic parameters	MFCC (dimension 12) and their time derivatives incremented every 10 ms
Test speech for alignment	503 sentence speech (ATR SPEECH DB Bset [1]) uttered by 6 males and 4 females who were not included in the trained speakers. Total 5030 sentence speech including about 250,000 phone boundaries.

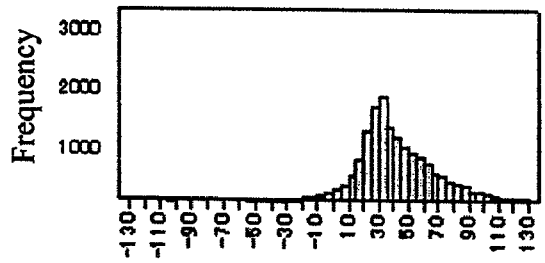
4. ANALYSIS OF AUTOMATIC ALIGNMENT ERRORS

As discussed in section 2, there are several factors that produce automatic alignment errors. In this section, alignment disagreements observed in automatic alignment experiments are analyzed based on the factors stated in section 2.

4.1. Errors caused by the extraction of spectral feature parameters

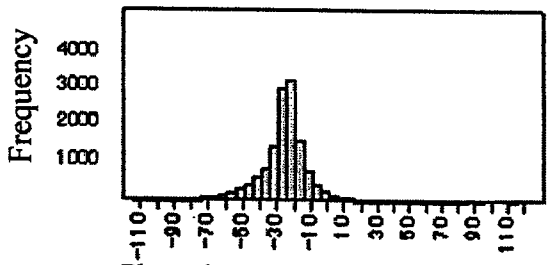
As spectral features are averaged over 30 ms windows, boundaries may be shifted within the window length. At phone boundaries that a human can easily identify from a raw waveform, the disagreements will be large. In particular, at the end of silent portions, this disagreement can be greatest. For an analysis frame including a silent portion, even a very short speech part provides the dominant contribution to the spectral characteristic of that frame, whilst the remaining silent portion has practically no contribution. Figure 1 shows this disagreement very clearly. As shown in the histogram of

(a) Vowel/consonant-to-silent boundary



Phone boundary disagreement (ms)
boundaries: 14125, average error: 41.59 ms, standard deviation: 26.79 ms

(b) Silent-to-vowel/consonant boundary



Phone boundary disagreement (ms)
boundaries: 14125, average error: -24.79 ms, standard deviation: 15.34 ms

Fig 1: Distribution of phone boundary disagreement before (a) and after (b) silent portion

the occurrence frequency of phone boundary disagreements, disagreement characteristics are totally different between before and after the silent portion. These differences, as expected, shortened the length of the estimated silent portions. Please note also that the peaks of the disagreement shifts are close to + or –30 ms, which coincides with the window length.

4.2. Errors caused by continuous spectral changes

It is difficult to determine phone boundaries where the spectrum is continuously slow-changing. As seen in the increase of boundary disagreements between human transcribers of more than 5 ms [5], such disagreements are also widely distributed in automatic alignment. Figure 2 shows the wide distribution of phone boundary disagreements at vowel successions. The standard deviation of disagreements at vowel successions reaches 21.16 ms, which is about two times larger than the distributions at boundaries where the spectrum changes quickly. This difference in distribution can also be seen in Figure 1. As the vowel offset is spectrally unclear compared to the vowel onset, the deviation in panel (a) of Fig. 1 is larger than that in panel (b).

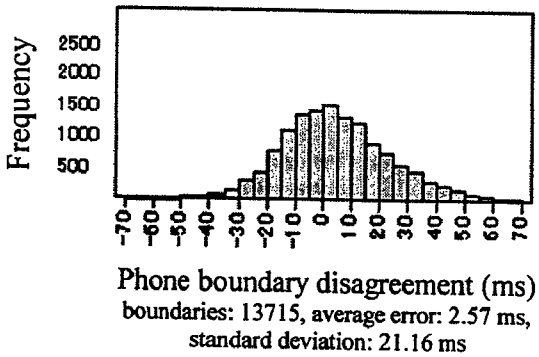


Fig 2: Distribution of phone boundary disagreement between successive vowels.

In addition to these disagreement analyses, we examined the relationship between the absolute error of phonemic length and the acoustic likelihood. We expected that the acoustic likelihood would flag the labels that had such large temporal errors that corrections by manual alignment would be required. The analyses were carried out for each speaker, label, and phonemic length because the scale of the likelihood measure depends on each of them. Figure 3 shows an example of the results under the condition that the label is /a/, the phonemic length by automatic alignment is 70 ms, and the speaker is one male. As shown in the figure, no correlation suggesting labels with extreme errors was observed.

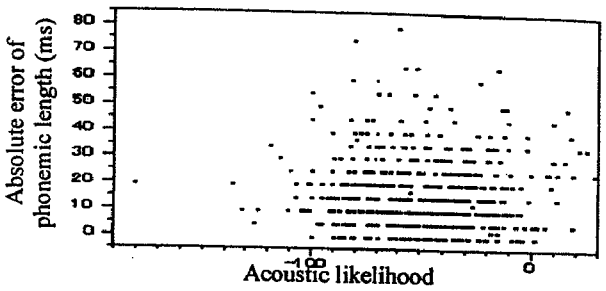


Fig 3: The relation between the absolute error of phonemic length and the acoustic likelihood (the label is /a/, the phonemic length by automatic alignment is 70 ms, the speaker is one male).

4.3. Errors caused by HMM model structure

We used tied state tri-phone HMM model that did not allow state skipping of null transitions. This model-dependent structure introduces a constraint of a 30 ms minimal phone length because we adopted a 10 ms frame-shift for input feature parameters. In continuous speech, however, phones exist that are shorter than 30 ms. In particular, the length of the Japanese /r/ segment is intrinsically short. Figure 4 (a) and (b) clearly show the apparent lengthening of /r/ segments due to the effect mentioned.

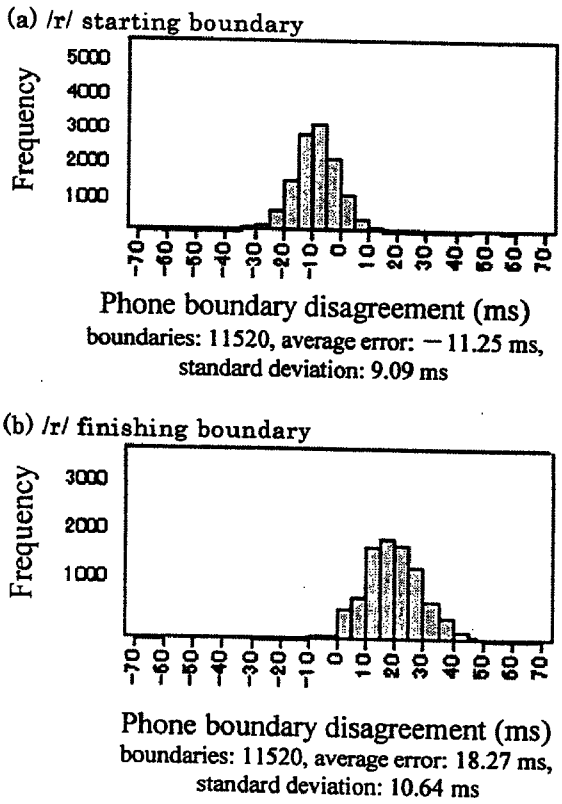


Fig 4: The distribution of phone boundary disagreement of label /r/.

4.4. Miscellaneous errors in automatic alignment

Besides the errors analyzed in the previous subsections, we found phone context dependent boundary shifts, as shown in Table 2. Although the main reasons for these shifts cannot be identified clearly, errors may exist that were caused by forced alignment of the HMM models. As most of the current statistical recognition techniques determine phone boundaries simply by maximizing the sum of likelihood value of acoustic models rather than by detecting phone boundaries accurately, displacement of phone boundaries is not directly evaluated in the automatic alignment process. This may affect these error disagreements.

Table 2: Other phone context dependent boundary shifts

Preceding phone	Succeeding phone	Average error (ms)
Voiceless consonant + voiceless vowel	Voiceless consonant	20.88
Voiceless consonant (except for palatal) + voiceless vowel	Voiced consonant	15.08
Voiced consonant	Vowel	12.55
Voiceless affricate	Vowel	-18.08
Syllabic nasal	Voiced consonant	-15.59
Vowel	Syllabic nasal	-15.56

5. MODIFICATION OF AUTOMATIC ALIGNMENT

As seen in the previous section, multiple factors are involved in automatic alignment errors. Although all of these factors need to be taken into account individually to reduce the amount of alignment errors, in practical terms, it seems quite difficult to do this for all of them. We need to conduct fairly large experiments to fix all the system parameter values to give the best recognition (alignment) performance. As a first step to reduce the amount of errors, we applied a post-processing modification to the automatic alignment results.

So far, in many cases, phone boundaries in the same phonetic context are shifted to one side for various reasons. The average duration shifts were calculated in all phonetic contexts and subtracted from the alignment results. To test the efficiency of this phonetic context dependent modification, 80% of all boundaries, which were randomly chosen, were used to calculate the context dependent shift and the remaining 20% were used in an open test. After repeating this jack-knife test five times, we got from 11.06 to 11.07 ms RMS errors and from 15.65 to 15.71 ms standard deviation. Comparing these values with the raw automatic alignment results, an improvement of about 4 ms was obtained in RMS errors. This improvement corresponds to about eight times that of the improvement obtained by speaker adaptation, as seen in section 3.

6. CONCLUSIONS

This paper is intended to achieve automatic precise alignment of phone boundaries through analysis of the causes of automatic alignment errors. For error analysis, automatic alignment results were compared with manually given boundaries. It was quantitatively confirmed that various factors generated these alignment errors. We proposed a post-processing procedure to reduce alignment errors. Alignment experiment showed that the proposed procedure could reduce alignment errors remarkably. About 4 ms, eight times as much improvements that achieved by speaker-adaptation, was obtained in RMS errors and about 5 ms, five times as much, was obtained in standard deviation.

Although we adopted the standard HMMs used in current speech recognition technology in this paper, as a second step to achieve reduction in errors, we are planning to fix all system parameter values to give the best recognition (alignment) performance based on the various error factors mentioned in this paper. Then we will compare the effects of this system modification with the current improvements of the post-processing modification. On the issue of the context dependent errors whose causes are not yet clear, we will first investigate whether automatic alignment under other conditions commonly has these errors in order to test their generality.

ACKNOWLEDGEMENTS

We would like to express our thanks to members of ATR-SLT and SLAP of ATR-HIS, in particular, H. Yamamoto, T. Hayashi, K. Okuda, R. A.-Yamada, and E. McDermott for all their support and suggestions for this study. This research was also supported in part by the Telecommunications Advancement Organization of Japan.

7. REFERENCES

[1] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano (1999). "ATR Japanese speech database as a tool of speech recognition and

synthesis," Speech Communication, Vol. 9, pp. 357-363.

[2] T. Osuga, Y. Horiuchi, and K. Ichikacwa (2001). "Investigation on the problems about automatic forced alignment for spontaneous speech," Proc. Fall Meet. Acoust. Soc. Jpn., pp. 107-108 (in Japanese).

[3] Y. Mera, A. Lee, H. Saruwatari, and K. Shikano (2001). "Statistical analysis of phoneme likelihoods in spontaneous speech using automatic segmentation by Julius," Proc. Fall Meet. Acoust. Soc. Jpn., pp. 57-58 (in Japanese).

[4] T. Fukada, S. Aveline, M. Schuster, and Y. Sagisaka (1997). "Segment boundary estimation using recurrent neural networks," Proc. Eurospeech'97, pp. 2839-2842.

[5] K. Takeda, Y. Sagisaka, S. Katagiri, and H. Kuwabara (1988). "A Japanese speech database for various kinds of research purposes," J. Acoust. Soc. Jpn., Vol. 44, pp.747-754.

[6] M. Tonomura, T. Kosaka, and S. Matsunaga (1994). "Unified speaker adaptation algorithm of maximum a posteriori estimation and transfer vector field smoothing," IEICE Technical Report, SP94-51, pp. 25-30.

[7] M. Ostendorf and H. Singer (1996). "Maximum Likelihood Successive State Splitting," Proc. Spring Meet. Acoust. Soc. Jpn., pp.131-132.

A. Abraham et al. (Eds.)
Soft Computing Systems:
Design, Management and Applications
IOS Press, 2002

Accurate Human Face Extraction using Genetic Algorithm and Subspace Method

Makoto Murakami, Masahide Yoneyama
Dept. of Information and Computer Sciences, Toyo University
2100 Nakanodai Kujirai, Kawagoe Saitama 350-8585, Japan

Katsuhiko Shirai
School of Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku Tokyo 169-8555, Japan

Abstract. Subspace method that can express facial images efficiently by linear translation into lower dimensional subspace has wide application for face recognition e.g. identification, facial pose detection etc. In the preprocess of this method the accurate extraction of human face area is required, but it is influenced by the light condition, various background, individual variation and so on, so it has not put into practical use yet. In this paper we examine the subspace method by comparison of the search space, and apply Genetic Algorithm to face extraction and show that the effective results was obtained.

1 Introduction

Lately multimodal interface has been noticed. It is expected as the latest interface that will take the place of the interface using keyboard or mouse. Most of us would accept that human face has important information. To construct the interface using visual information, the recognition of them is indispensable. So a considerable number of studies ever have been made on expression recognition, individual distinction and eye-gaze recognition etc.

Subspace method that can express facial images efficiently by linear translation into lower dimensional subspace has wide application for face recognition e.g. identification, facial pose detection etc. In the preprocess of this method the accurate extraction of human face area is required, but it is influenced by the light condition, various background, individual variation and so on, so it has not put into practical use yet. In this paper, we apply Genetic Algorithm to face extraction and show that the effective results was obtained.

2 Recognition Method

2.1 *Subspace Method*

2.1.1 Construction of Subspace

Let N features extracted from one image be the following N-dimensional vector x

$$x = [a_1, a_2, \cdots, a_N]^T, \tag{1}$$

where a vector x is normalized to $|x| = 1$.
The training set is shown as a set of the vectors

$$\{x_1, x_2, \cdots, x_R\}, \tag{2}$$

where R is the step number of rotation.

Then we calculate orthonormal bases in the subspace of the set of vectors. From the matrix X_{cov}

$$X_{cov} \equiv [x_1 - c, \cdots, x_R - c], \tag{3}$$

where c is the average vector of the training set

$$c = \frac{1}{R}\sum_{r=1}^{R} x_r, \tag{4}$$

we calculate the covariance matrix Q_{cov}

$$Q_{cov} \equiv X_{cov}{X_{cov}}^T. \tag{5}$$

In the same way from matrix X_{cor}

$$X_{cor} \equiv [x_1, \cdots, x_R], \tag{6}$$

we calculate the correlation matrix Q_{cor}

$$Q_{cor} \equiv X_{cor}{X_{cor}}^T. \tag{7}$$

Then, we solve the following characteristic equation

$$\lambda_j e_j = Q e_j \tag{8}$$

to obtain eigenvalues λ_j and eigenvectors e_j.
k eigenvectors corresponding to bigger eigenvalues

$$(e_1, \cdots, e_k), (\lambda_1 \geq \cdots \geq \lambda_k \cdots \geq \lambda_N) \tag{9}$$

are orthonormal bases of k-dimensional subspace.

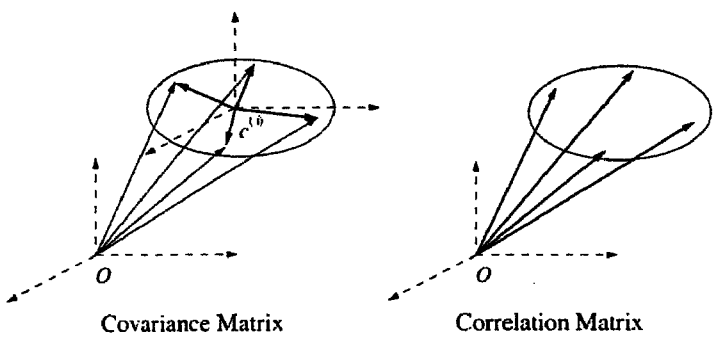


Figure 1: Conceptual Diagram of Subspace

2.1.2 Pose Detection

1. Projection into Subspace
   To detect the pose, we use the subspace constructed from the covariance matrix Q_{cov}, because it represents variance from the average, as the conceptual diagram in Fig.1 indicates.

   We project the feature vectors extracted from the images into the subspace. The projected vectors p_r

$$p_r = [e_1, e_2, \cdots, e_k]^T (x_r - c) \tag{10}$$

   are points in the subspace.

2. Pose Detection
   A normalized vector y extracted from an input image is projected into the subspace. The projected point z

$$z = [e_1, e_2, \cdots, e_k]^T (y - c) \tag{11}$$

   is one point in the subspace.
   We detect the pose from the distance d_r between the projected point z and the points p_r in the subspace.

$$d_r = |z - p_r| \tag{12}$$

   In general, to detect the pose we calculate r that minimizes d_r.

2.1.3 Face Extraction

Projective vector $\hat{y}$ into the subspace of an input vector y is shown as

$$\hat{y} = \sum_{j=1}^{k} \{(y - c)^T e_j\} e_j \tag{13}$$

when covariance matrix based subspace is used, or

$$\hat{y} = \sum_{j=1}^{k} \{y^T e_j\} e_j \tag{14}$$

when correlation matrix is used.

We use the squared norm of $\hat{y}$ as the evaluation value of similarity of an input vector to the subspace.

$$\|\hat{y}\|^2 = \sum_{j=1}^{k} \{(y - c)^T e_j\}^2 \tag{15}$$

$$\|\hat{y}\|^2 = \sum_{j=1}^{k} \{y^T e_j\}^2. \tag{16}$$

And then, we search the area that maximizes $\|\hat{y}\|^2$ to extract the face area.

2.2 Genetic Algorithm

To search the faces of which both position and size are unknown in an image is just the maximum searching problem of $\|\hat{y}\|^2$ in the 4-dimensional space of (x_c, y_c, M_x, M_y) that are the center of gravity and magnifying power of the extracted area. We apply the Genetic Algorithm to this searching problem.

1. Individual
   The genotype $G_k (k = 1, 2, \cdots, N)$ of the individual B_k is defined as follows

$$G_k = (x_{ck}, y_{ck}, M_{xk}, M_{yk}), \tag{17}$$
$$(x_{ck}, y_{ck}) \ : \ \text{the center of gravity}$$
$$(M_{xk}, M_{yk}) \ : \ \text{magnifying power.}$$

2. Fitness Value
   The fitness value F is defined as follows

$$F = \|\hat{y}\|^2, \tag{18}$$

   where $\|\hat{y}\|^2$ is the evaluation value considered in the section 2.1.3.

3. Selection
   Let N be the number of individuals in one generation. the ratio that When $R\%$ individuals of which the fitness values are lower die out, the number is $M = N \times \frac{R}{100}$.

4. Multiplication
   $(100 - R)\%$ individuals of which the fitness values are higher survive in the next generation. From this $(N - M)$ individuals new M individuals are generated, and then the total number is fixed in every generation. However, to prevent all individuals from gathering in the same location in the early generation, we make less than $S\%$ individuals survive in the neighbor, however big their fitness values are.

5. Heredity Rule
   When from an individual B_i with genotype G_i a new individual $B_i\prime$ is generated, its genotype $G_i\prime$ defined as follows

$$G_i\prime = (x_{ci} + \Delta x_i, y_{ci} + \Delta y_i, M_{xi} + \Delta M_{xi}, M_{yi} + \Delta M_{yi}), \tag{19}$$

$$\Delta x_i \ : \ \text{random number in } [-\Delta x_{max}, \Delta x_{max}]$$
$$\Delta y_i \ : \ \text{random number in } [-\Delta y_{max}, \Delta y_{max}]$$
$$\Delta M_{xi} \ : \ \text{random number in } [-\Delta M_{xmax}, \Delta M_{xmax}]$$
$$\Delta M_{yi} \ : \ \text{random number in } [-\Delta M_{ymax}, \Delta M_{ymax}].$$

   Δx_{max},Δy_{max},ΔM_{xmax},ΔM_{ymax} are functions of F_i, determined in the training process as follows

(a) Let (x_0, y_0) be the coordinate of a face area extracted from an image by the visual observation, and (M_{x0}, M_{y0}) be the magnifying power. We calculate the fitness value F of $(x_0 + \Delta x, y_0 + \Delta y, M_{x0} + \Delta M_x, M_{y0} + \Delta M_y)$ around the face area.

(b) Then we record the Δx, Δy, ΔM_x, ΔM_y as the functions of fitness value F. When the fitness values are equal, we record the maximum Δx_{max}, Δy_{max}, ΔM_{xmax}, ΔM_{ymax}. And when some fitness values can't appear in this process, we define Δx_{max}, Δy_{max}, ΔM_{xmax}, ΔM_{ymax} as follows

$$\Delta x_{max} = xmax(F-1)^2 \tag{20}$$

$$\Delta y_{max} = ymax(F-1)^2 \tag{21}$$

$$\Delta M_{xmax} = M_x max(F-1)^2 \tag{22}$$

$$\Delta M_{ymax} = M_y max(F-1)^2, \tag{23}$$

where $xmax$, $ymax$, $M_x max$, $M_y max$ is the maximum of Δx_{max}, Δy_{max}, ΔM_{xmax}, ΔM_{xmax} in the neighbor of face area respectively.

6. Mutation
M individuals are generated at random once in every L generations, not following the heredity rule defined above. This can prevent from convergence in the local maximum to lead the optimum solution.

3 Experiment

3.1 Outline

In this experiment we use face images taken from 36 horizontal directions as training data. The size of them is normalized to 64×64 pixels. Then we construct the subspace from edge and smoothing feature vector as Fig.2 shows, to scale back the influence of the light condition, individual variation and so on.

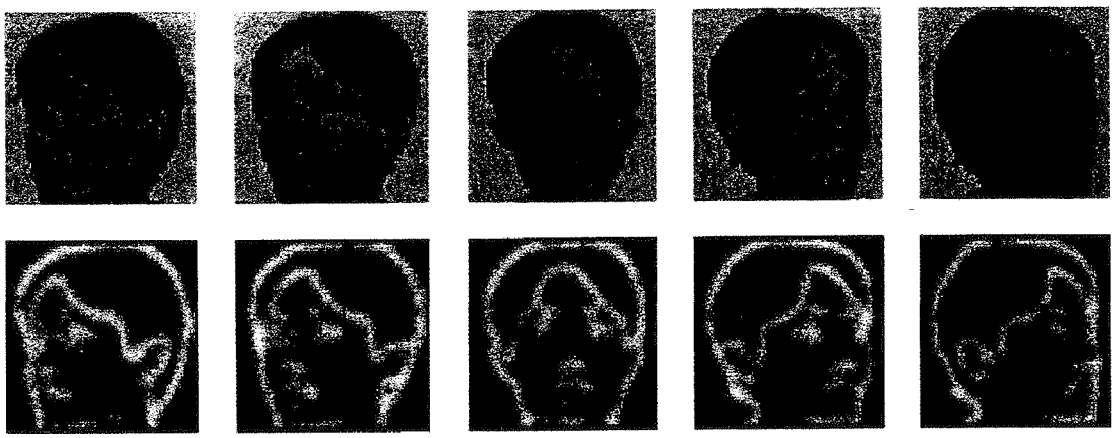


Figure 2: Training Data

Preparatory for the experiment of human face extraction, we compare the search spaces of $\|\hat{y}\|^2$ considered in the section 2.1.3 in two methods (using covariance matrix and correlation

matrix), and examine if it is able to reject the background area. Then we apply the Genetic Algorithm considered in the section 2.2 to human face extraction.

3.2 *Results and Considerations*

3.2.1 Comparison of Search Spaces

To extract the face area is just to reject the background area. But the various background images are not trained, so the curved surface in the search space is unpredictable. Fig.3 indicates noticeable examples of the search space on the assumption that face size is fixed for the images taken on the various conditions. The vertical axis in the graphs indicates the $\|\hat{g}\|^2$ defined in the section 2.1.3, so extracting the face area is equal to looking for the maximum in this space. The black region in the graphs shows the face area with which the template overlaps more than half.

As the graphs indicate, in the method using correlation matrix the score in the background is not restrained. So it is possible to converge on the points in the background. On the other hand, in the covariance matrix based method the score in the face area obviously higher than in the background though the graph is discontinuous curved space. Therefore in the next section we experiment with the method using covariance matrix.

Then we compare the inner product using some bases with all bases, to consider the number of bases constructing subspace. The pie charts in Fig.4 indicate the rates of bases of which inner product is maximized in the face area and the background of various images respectively. These graphs show that the important bases are e_1, e_2, e_3, and e_2, e_3 are unresponsive to the background, so possibly restrain the score in the background. Fig.3 shows also search spaces of the sum of the inner product of e_1, e_2, e_3. There is not a great difference between using all bases and using three bases (e_1, e_2, e_3) in search space itself. So we use e_1, e_2, e_3 in the experiment in the next section.

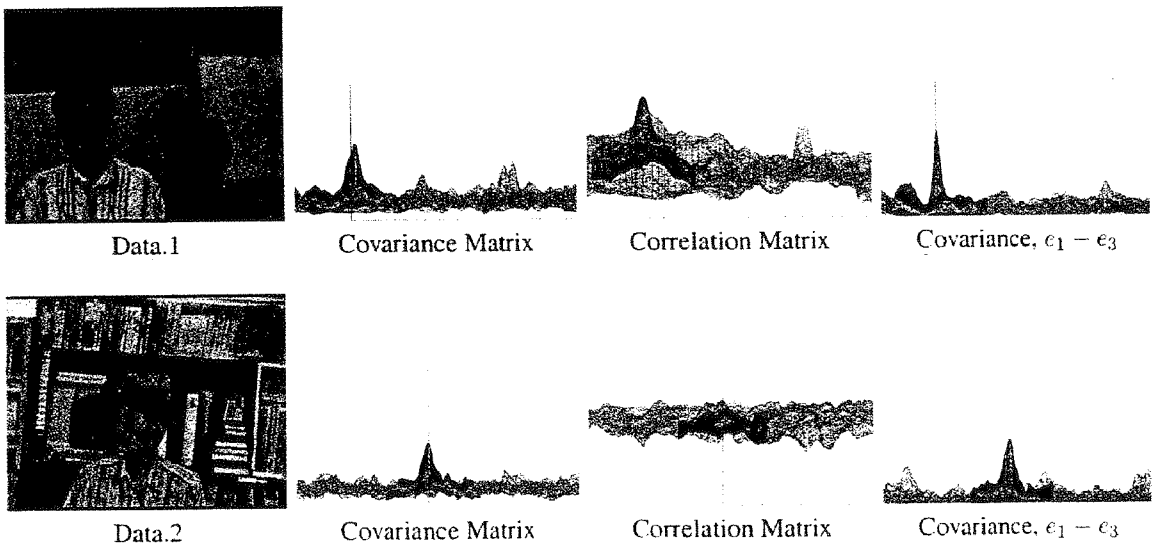
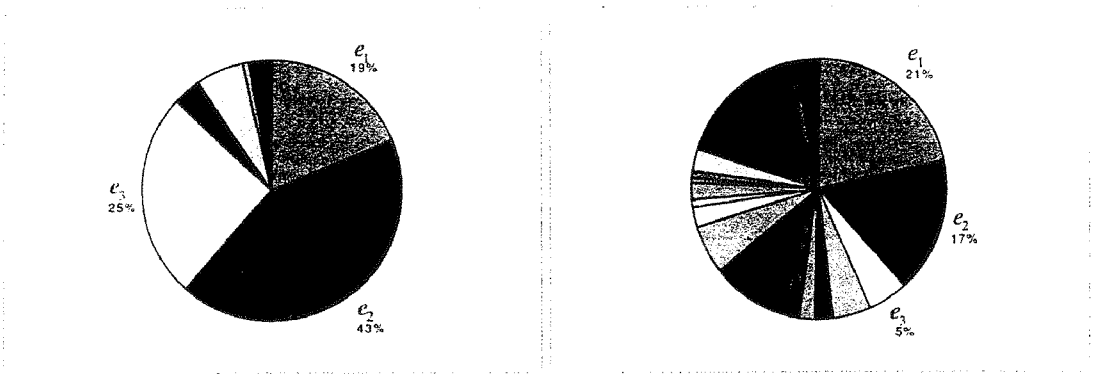


Figure 3: Search Space



Face Area Background

Figure 4: Rate of Bases of which inner product is maximized

3.2.2 Face Extraction using Genetic Algorithm

We use the Genetic Algorithm considered in the section 2.2 to extract face area. The number of individuals in one generation is 24, 12 individuals of which the fitness values are lower die out, and then from the survived 12 individuals new 12 individuals are generated. The new 12 individuals are mutated once in every 5 generations. Furthermore to prevent all individuals from gathering in the same location, we make not more than 3 individuals searching in the neighbor survive in the next generation. Therefore, four or more region can be searched in at the same time.

Then we consider the evaluation value of extraction accuracy. Let C be the area extracted by the visual observation, and H be the area extracted by GA. We calculate the size S_D of the area D defined as

$$D = (C \cap \bar{H}) \cup (\bar{C} \cap H). \tag{24}$$

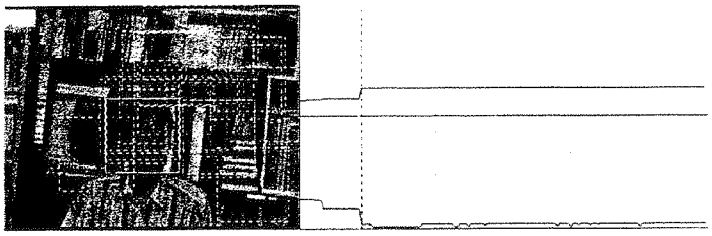
And we define an evaluation value R as

$$R = \frac{S_D}{S_C}. \tag{25}$$

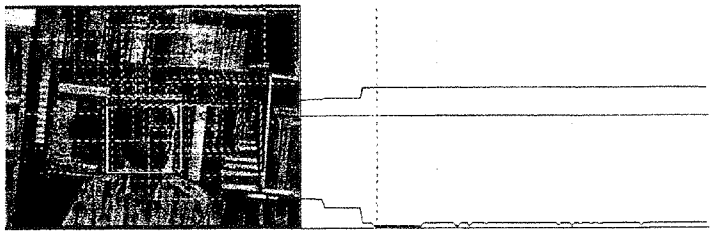
Then, we define another evaluation value as the differences D_r between the correct pose and the pose r minimizes d_r defined in the section 2.1.2. In this experiment we use 250 images with various backgrounds.

The results are shown in Fig.5-Fig.7. The horizontal axis in the graph indicates the generation of GA. The vertical axis in the upper part of the graph indicates fitness value F, and it in the lower part indicates accuracy R.

Fig.5 shows an example to converge in early generation; face area has been extracted in about the 24th generation. Fig.6 shows an example to converge getting out of local solutions; individuals begin gathering around the face area in the 64th generation, and the face area get extracted accurately by and by. Fig.7 shows an example of failure; no individual is generated near the face area during the search process.

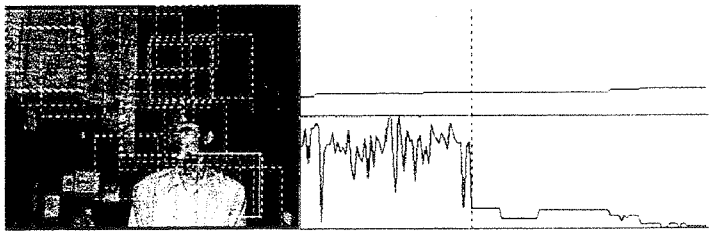


24th Gen.

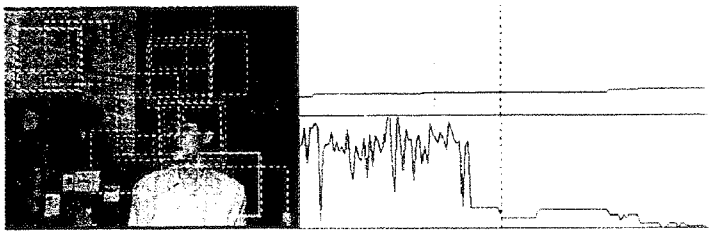


29th Gen.

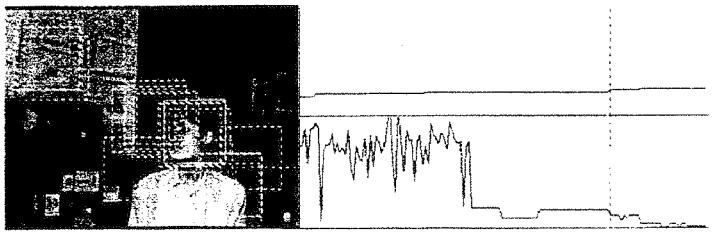
Figure 5: Result 1 (success)



64th Gen.

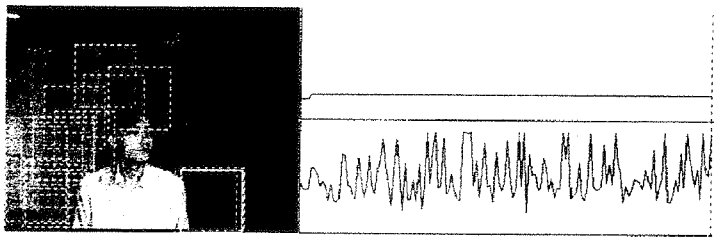


75th Gen.



115th Gen.

Figure 6: Result 2 (success)

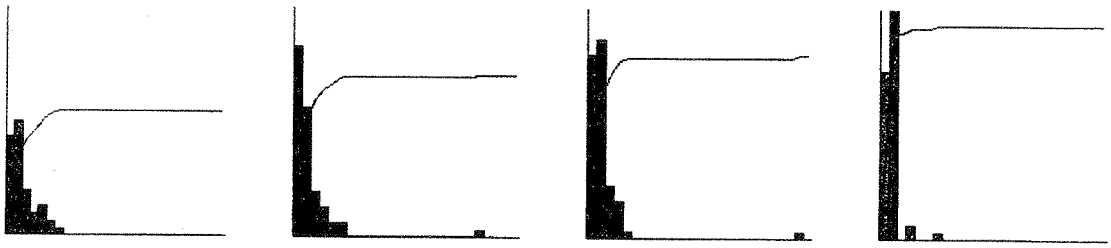


150th Gen.

Figure 7: Result 3 (failure)

Then we consider the change of accuracy in generations. Fig.8 indicates the extraction accuracy. The horizontal axis in these graphs shows R, the vertical axis shows the number of the extracted images of which the accuracy is R. And the line graph shows the accumulated rate. Here, one step of the horizontal axis is 0.05. Fig.9 indicates the extraction accuracy by another evaluation value D_r. The horizontal axis in these graphs shows $|D_r|$.

As generations go by, the accuracy of both R and D_r are gradually improved. It follows that our genetic algorithm is effective for the discontinuous search space.



10th Gen. 30th Gen. 50th Gen. 100th Gen.

Figure 8: Extraction Accuracy 1

10th Gen. 30th Gen. 50th Gen. 100th Gen.

Figure 9: Extraction Accuracy 2

4 Conclusion

In the method of constructing subspace using covariance matrix the score in the face area obviously higher than in the background, this method is effective to reject the background. But the search space is discontinuous curved space. So we apply Genetic Algorithm to this search space and the effective results was obtained.

Our method will be well fitted to the problem of face recognition. Subspace method or eigenface method is proposed for the identification or verification in the first place. It is important to note that the rejection rate for the unknown is very important for these applications such as security systems. Furthermore we are investigating the issues of robustness to changes in lighting, head orientation and the number of people.

For applications such as human interface, it is necessary to reduce computational effort, because of the processes for the image sequence. But once face location is fixed in a frame, we can track the face region for the next frames. Therefore we are investigating the accurate and robust tracking method for applications.

References

[1] Erkki Oja:"Subspace Methods of Pattern Recognition," Research Studies Press, England, 1983.

[2] Yoshie Komatsu and Yasuo Ariki:"Orientation Invariant Face Extraction and Recognition based on Subspace Method," Technical Report of IEICE, PRU95-191, pp7-14, 1996-01.

[3] Yoshiaki Sugiyama and Yasuo Ariki:"Facial Region Tracking and Training Method by Subspace Projection," Technical Report of IEICE, PRMU97-162, pp77-82, 1997-11.

[4] Tomoharu Nagao, Takeshi Agui and Hiroshi Nagahashi: "Pattern Matching of Binary Shapes Using a Genetic Method," Transaction of IEICE, J76-D-II, No.3, pp557-565, 1993-03.

[5] Akira Hara and Tomoharu Nagao:"Extraction of facial regions of arbitrary directions from still images with a genetic algorithm," Technical Report of IEICE, HCS97-12, pp37-44, 1997-09.

[6] Shinya Masunaga and Tomoharu Nagao:"Extraction of human facial regions in still images using a genetic algorithm," Technical Report of IEICE, PRMU95-160, pp13-18, 1995-11.

Corpus-based modeling of naturalness estimation in timing control for non-native speech

Makiko Muto†, Yoshinori Sagisaka†, Takuro Naito††, Daiju Maeki††, Aki Kondo††, Katsuhiko Shirai††

†GITI, Waseda University
1-3-10 Nishi-Waseda, Shinjuku-ku, Tokyo 169-0051, Japan

††Department of science and engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

makiko.muto@ruri.waseda.jp, sagisaka@giti.waseda.ac.jp, takuro@ruri.waseda.jp, daiju@toki.waseda.jp, 55aki3@toki.waseda.jp, shirai@shirai.info.waseda.ac.jp

Abstract

In this paper, aiming at automatic estimation of naturalness in timing control of non-native's speech, we have analyzed the timing characteristics of non-native's speech to correlate with the corresponding subjective naturalness evaluation scores given by native speakers. Through statistical analyses using English speech data spoken by Japanese with temporal naturalness scores ranging one to five given by natives, we found high correlation between their scores and the differences from native's speech. These analyses provided a linear regression model where naturalness in timing control is estimated by differences from native's speech in durations of overall sentences, individual content and function words and pauses. The proposed naturalness evaluation model was tested its estimation accuracy using open data. The root mean square errors 0.64 between scores predicted by the model and those given by the natives turned out to be comparable to the differences 0.85 of scores among native listeners. Good correlation between model prediction and native's judgments confirmed the appropriateness of the proposed model.

1. Introduction

For prosody control, control characteristics and generation mechanisms of segmental durations and fundamental frequency have been studied for many languages over many years. Recently with the ceaseless efforts of speech data collection and scientific computational modeling have enabled the quantitative control with high precision using statistical optimization techniques through the minimization of objective errors in the speech synthesis field [1].This technology known as corpus-based modeling has successfully been applied to both prosody and segmental characteristics control for synthesizing high quality speech with natural prosody. Though some of current corpus-based modeling put too much emphasis on optimization and often neglect underlying scientific principles in speech generation, in principle, they can be applied to the quantification of subjective evaluation where so many underlying mechanisms and principles have not yet been known.

On the other hand, in the study of foreign language learning, subjective evaluation by native speakers has been commonly used, which makes it difficult for learners to acquire language skill by themselves more efficiently. An automatic evaluation of non-native's speech is expected to highly reduce the teacher's efforts and promote the scientific understanding of speech generation control for a better design of CALL (Computer Assisted Language Learning) system. Up to now, little studies have been carried out on automatic evaluation based on objective criteria and scientific quantitative analyses are needed for measures for scoring.

As a non-native's speech is produced as a result of blending different control characteristics of the target language and the speaker's native one, the modeling of generation characteristics is difficult. It requires much knowledge of two languages on control factors and mechanisms in a common modeling framework. In addition, not only speech generation control characteristics, but also perceptual characteristics have to be taken into consideration for automatic evaluation [2]-[4].

In this paper, as a first step towards automatic evaluation of non-native's spoken language, we applied a corpus-based approach to evaluate the naturalness in the timing control of non-native speech using objective measures. Instead of modeling generation characteristics directly, we compare with native's speech and use the differences in timing characteristics as naturalness evaluation measures. For a spoken language database, we used English speech spoken by Japanese that had been collected for analyses of the acoustic characteristics of English speech spoken by Japanese learners. [5]. In Section 2, we will briefly look the timing control differences between Japanese and English. Statistical analyses are carried out using the above data to look for the objective factors highly correlating with native's subjective scores in Section 3. In Section 4, a linear regression model is proposed for automatic naturalness evaluation with experimental results applying to open data. Finally, the effectiveness of the automatic evaluation model is summarized and future problems are discussed.

2. The difference of timing control between English and Japanese

Comparing prosodic characteristics between English and Japanese, there is big difference in timing control. In English, stressed syllable and unstressed syllable appear alternately, which is called stress-timed rhythm. While in Japanese, moras are spoken in almost same length, which is called mora-timed rhythm. It is known that Japanese segmental durations are controlled by many factors [6]-[10]. Segmental durations are not characterized only by phoneme intrinsic attributes but also by constraints from longer units. In particular, a moraic constraint is clearly observed in the control of Japanese vowel durations. A negative correlation is found between vowel durations and adjacent consonant durations. The temporal compensation of vowel duration is more influenced by the preceding consonant duration than the following one, and this is considered to be an acoustic manifestation of mora-timing. This temporal compensation has been observed both in raw duration data and in normalized z-score data for each vowel [11]. Through such statistical analyses, it has been confirmed that the compensation takes place in mora units but not in syllable units. Mora-timed rhythm of Japanese is in contrast to the stress-timed rhythm observed for other languages such as English. It has also been observed that not only segmental durations but pause length is under moraic control [12]

Due to the presence of big difference in timing control, it is difficult for Japanese to speak English in stressed-timed. This causes one of the big unnaturalness of English speech spoken by Japanese. In fact, some native speakers have pointed out the unnaturalness of English speech spoken by Japanese in rhythm over whole utterance more seriously than incorrect pronunciation of individual words.

3. Analysis of timing control differences between natives and non-native learners

As it looks quite complicated to directly model non-native's timing characteristics, we decided to estimate naturalness of non-native's speech by comparing with native's speech. Not only the timing control difference between two languages, but also there exist other factors that are frequently observed in beginner's speech. In this section, we try to find remarkable durational differences in temporal control between natives and non-native learners by calculating correlation between subjective scores given by natives.

3.1. Speech data for analysis

For a speech database, English speech samples spoken by Japanese were selected from "English Learners' Speech Database" [5]. This database was designed to analyze the acoustic characteristics of English speech spoken by Japanese learners. Not only phonetic information but also some prosodic symbols indicating stress, phrase boundaries and pause locations had been pre-assigned in a text to purely measure learner's speech generation skill. Speakers were requested to practice the pronunciation of given sentences before recording. In recording, they were asked to read repeatedly until they could speak what they thought was the correct pronunciation. Most of subjects are Japanese university students with different English language skills. These data specifications, in particular the recording conditions to speakers are quite ideal for the purpose of our analyses to find timing control differences highly correlated with native speakers' evaluation in timing.

Furthermore, each speech sample has multiple scores evaluated by four native speakers having knowledge of English phonetics and plenty of experiences in teaching English to Japanese. For timing characteristics, they were asked to evaluate the rhythm of those utterances using a five-point rating scale ranging from"1" to "5" with large number corresponding to closer to native rhythm. Each utterance was evaluated by four persons. 104 speech samples were selected from the database for the analyses. They were uttered by sixty-three adult males. One subject uttered at least one sentence, but not more than four sentences.

3.2. Analysis factors for timing control differences

There seem to exist many timing differences between natives and non-native learners that native listeners pay attention to when they evaluate the rhythm of non-native speakers. Considering beginner's timing control characteristics and language dependent rhythmic differences observed in target language (English) and native's own (Japanese), we considered the following factors affecting temporal differences.

(1) Overall speech rate
Usually beginners cannot speak fast. Non-native beginners tend to utter slower than natives. This characteristic will be manifested in total speech duration.

(2) Local speech rate
Beginners cannot speak unfamiliar words as fast as natives. This characteristics will be measured by the difference of individual word durations.

(3) Stress timing
(Timing control characteristics of the target language)
In English, durations of function words are highly reduced to control time interval between stressed syllables. As strong mora-timing constraint of speaker's language makes it extremely difficult to reduce them, the duration difference in function words is expected to be a good cue for evaluation.

(4) Phrasing and pausing
It is well known that prosodic grouping including pausing affects temporal structure. We will be able to quantify these differences by comparing the position of phrase boundaries and pause insertion counts. In particular, as beginner tends to pronounce word by word, pause counts seem to be longer in the beginning stage.

(5) Mora timing
(Timing control characteristics of learners' language)
Though the quantification of moraic control of subject language itself is difficult to calculate without precise duration modeling, pause information can be used for this purpose. It is well known that pause length is also highly constrained by moraic units in Japanese read speech. Kaiki showed that the distribution of pauses in Japanese sentence has bimodal peaks at one mora length and three mora length [12]. Not only pause positions and counts but also their durations can also be informative.

Table 1: Correlation between timing control differences and subjective evaluation scores

timing control differences	correlation
sentence duration	-0.54
word duration	-0.53
function word duration	-0.48
pause count	-0.32
pause duration	-0.44
content word duration	-0.46
sentence duration except pauses	-0.56
ratio of word duration to sentence duration	-0.26
counts of incorrectly inserted pauses	-0.31
missing pause counts	0.13
sum of pause durations in a sentence	-0.48
duration of incorrectly inserted pauses	-0.42

To quantify the contributions of above factors, we selected the parameters expressing timing difference from native's speech for analysis. They were sentence durations, word durations, function word durations, pause counts, pause durations, content word durations, sentence durations except pauses, ratios of word durations to a sentence duration, counts of incorrectly inserted pauses, missing pause counts, sum of pause durations in a sentence and durations of incorrectly inserted pauses.

3.3. Correlations between timing control differences and subjective scores

The subjective naturalness scores given by native speakers on the rhythm of non-native utterances were compared with timing control differences from native's ones. As reference native speech samples, we newly recorded the same sentences of a native speaker. In recording, we provided a speaker the same information that non-native speakers had been given [5].

To calculate timing control differences, native's durations were subtracted from non-native speaker's ones for each parameter listed in the previous sub-section. Subjective scores for each utterance were pooled over four native judgments. The correlation coefficients between timing control differences and the subjective scores are shown in Table 1. As shown in the correlation values, the following characteristics were observed for analysis factors.

As expected, beginner's speech tends to be longer. A remarkable negative correlation was observed between sentence duration differences and subjective evaluation scores. As well as sentence duration differences, there exist a negative correlation between individual word or function word duration differences and subjective evaluation scores. Though the current analyses themselves cannot distinguish the differences between control factors, we believe that further detailed evaluation modeling can clarify the contributions of these factors more clearly. Pausing related duration differences also show negative correlations between subjective evaluation scores except pause counts. Pre-assignment of prosodic symbols in the text might have reduced this correlation.

4. A model for automatic estimation of naturalness in timing control

Using timing control differences from native's utterances highly correlated with subjective evaluation scores, we made a statistical model for naturalness evaluation in timing control for English utterances spoken by Japanese. Sixty new sentences were selected from the same database for open tests. These sentences were read by twenty-nine adult males that included twenty-two speakers of the analysis set.

We adopted a linear multiple regression model for computing an evaluation score. For control factors of a model, only highly correlated timing control differences from native's utterances were selected. We used the combination of timing control differences underlined in Table 1, which gives the highest multiple correlation coefficient 0.66. This linear multiple regression model is denoted as Model1 in Table 2.

To evaluate the accuracy a model for automatic estimation, estimated evaluation scores were compared with subjective evaluation scores given by native speakers. The root mean square errors turned out to be 0.46 for the analysis data set and 0.64 for the test data set. Both of these values are smaller than the standard deviations among four subjects 0.86 and 0.85 respectively. These results show the reasonable performance of the proposed evaluation model.

The correlation scores were also calculated between predicted evaluation scores and their subjective evaluation scores for both the analysis set and the test set. The subjective and objective rating score shows in Figure 1. The correlations between subject evaluations were shown in Table 2. In average, correlations were 0.66 for the analysis set and 0.55 for the test set. Except the correlation between the test set of the subject No.4 where only small score range was observed, the estimated scores have good correlation between subjective ones.

Throughout the analysis of 104 utterances, we observed different naturalness evaluation characteristics between two sets of sentences consisting of 76 ordinary sentences and 35 very short sentences without any function words. For each sentence set, we calculated correlation using the same timing control differences from native's speech and subjective evaluation scores. Table 3 (a) and (b) respectively show the correlation coefficients of timing control differences highly correlated with subjective evaluation scores for ordinary sentences and short sentences.

To improve the modeling accuracy, we adopted the linear multiple regression of computing an evaluation score for each set. For control factors of model, highly correlated timing control differences from native's utterances shown in Table 3 were selected. Those combinations give the highest multiple correlation coefficient 0.67 in the ordinary sentences set and 0.70 in the very short sentences set. These linear multiple regression model are denoted as Model2 in Table 2.

As shown in Table 2, the correlation scores between predicted evaluation scores and their subjective evaluation scores were better than Model1. However, the root mean square errors of estimated scores turned out to be 0.54 for the analysis data set and 0.77 for the test data set. These values are a little bit worse than Model1 though both of these values are smaller than the standard deviations among four subjects. These inconsistencies seem to result from data scarcity. Further analysis using more systematic data set are needed.

Table 2: Correlation coefficient between subjective and objective evaluation scores. "a" and "t" indicate the case of analysis and test respectively.

	Model1_a	Model1_t	Model2_a	Model2_t
subject #1	0.55	0.63	0.58	0.61
subject #2	0.43	0.41	0.42	0.40
subject #3	0.52	0.48	0.54	0.47
subject #4	0.53	0.16	0.60	0.32
ALL	0.66	0.55	0.70	0.58

Table 3: Correlation coefficient between timing control differences and subjective rating scores by natives speakers for subgroups.

(a) ordinary sentences having function words

timing control differences	correlation
sentence duration	-0.52
word duration	-0.53
function word duration	-0.48
pause duration	-0.46
sentence duration except pause	-0.56
sum of pause duration in sentence	-0.46

(b) short sentences without function words

timing control differences	correlation
sentence duration	-0.36
word duration	-0.32
pause count	-0.29
pause duration	-0.23
sentence duration except pause	-0.19
pause counts of incorrect insert	-0.36
sum of pause duration in sentence	-0.30
pause of incorrect insert duration	-0.23

5. Conclusion

Aiming at automatic evaluation of naturalness with objective measures, we analyzed the correlations between the timing control differences from native speaker's utterance and the native speakers' subjective scores. Using English speech data spoken by Japanese learners and the corresponding English native's speech, we observed the high correlation between native's subjective evaluation scores and the timing differences from native's speech in sentence durations, word durations, function word durations, pause durations. Next, using timing control differences correlated with subjective score, we estimated subjective scores using a statistical model. The estimation accuracy test using open data confirmed the effectiveness of the proposed model. In this paper, though we could have used only gross timing control characteristics and small data with different speaker characteristics, reasonably good estimation performance was obtained. We are planning to use more detailed timing characteristics for better estimation with more systematic data aiming at utterance contexts and speakers free naturalness evaluation.

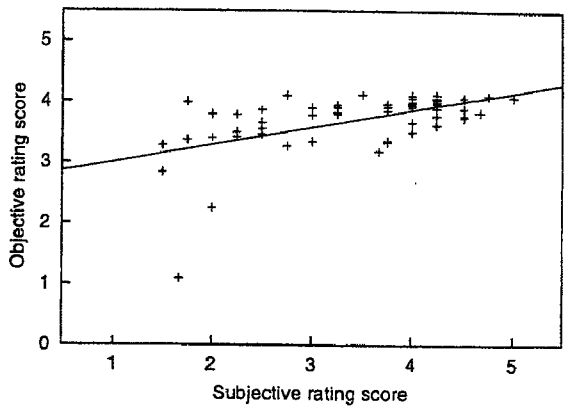


Figure 1: Subjective and objective rating score in the test data.

6. References

[1] Y. Sagisaka, "Modeling and perception of temporal characteristics in speech," Proc. ICPhS'03, 2003 (to appear).

[2] H. Kato, M. Tsuzaki and Y. Sagisaka, "Acceptability for temporal modification of consecutive segments in isolated words," J. Acoust. Soc. Am. 101: 2311-2321, 1997.

[3] H. Kato, M. Tsuzaki and Y. Sagisaka, "Acceptability for temporal modification of single vowel segments in isolated words," J. Acoust. Soc. Am. 104: 540-549, 1998.

[4] H. Kato, M. Tsuzaki and Y. Sagisaka, "Effects of phoneme class and duration on the acceptability of temporal modifications in speech," J. Acoust. Soc. Am. 111: 387-400, 2002.

[5] N. Minematsu, et al., "Development of English speech database spoken by Japanese learners," Preprint of the COCOSDA Workshop 2001, pp. 76-81, 2001.

[6] Y.Sagisaka and Y. Tohkura, "Phoneme duration control for speech synthesis by rule," (in Japanese) Trans. IEICE J67-A, No.7, pp.629-636, 1984.

[7] K. Takeda, Y. Sagisaka and H. Kuwabara, "On sentential effects in the control of segmental duration in Japanese," JASA Vol.86 (6) pp.2081-2087, 1989.

[8] N. Kaiki, K. Takeda and Y. Sagisaka, "Linguistic properties in the control of segmental duration for speech synthesis," pp.255-264 in "Talking Machines" edited by G.Bailly et al, North-Holland, 1992.

[9] N. Kaiki and Y. Sagisaka, "The control of segmental duration in speech synthesis using statistical models," pp.391-402 in "Speech perception, production and linguistic structure"edited by Y. Tohkura et al, Ohmsha IOS press, 1992.

[10] Y. Sagisaka, and N. Kaiki, "Prosody control for spontaneous speech synthesis," Proc. ICPhS'91, pp.506-509, 1991.

[11] N. Campbell, "A study of Japanese speech timing from the syllable perspective," (in Japanese) Trans. PSJ Vol.3 No.2, pp.29-39, 1999.

[12] N. Kaiki, and Y. Sagisaka, "Pause characteristics and local phrase-dependency structure in Japanese," Proc. ICSLP'92, pp. 357-360, 1991.

RECOGNITION OF PARA-LINGUISTIC INFORMATION AND ITS APPLICATION TO SPOKEN DIALOGUE SYSTEM

Shinya Fujie[1] *Yasushi Ejiri*[1] *Yosuke Matsusaka*[1] *Hideaki Kikuchi*[2] *Tetsunori Kobayashi*[1]

[1] School of Science and Engineering, Waseda University, Japan
[2] School of Human Sciences, Waseda University, Japan
{fujie,ejiri,yosuke,koba}@tk.elec.waseda.ac.jp, kikuchi@waseda.jp

ABSTRACT

The human-human interactions in a spoken dialogue seem to use not only linguistic information in the utterances but also some sorts of additional information supporting linguistic information. We call these sorts of additional information "para-linguistic information". In this paper, we present a recognition method of attitudes by prosodic information, and a recognition method of head gestures. In the former method, in order to recognize two attitudes, such as "positive" and "negative", F0 pattern and phoneme alignment are introduced as features. In the latter method, in order to recognize three gestures, such as "nod", "tilt" and "shake", left-to-right HMM is introduced as the probabilistic model as well as optical flow is introduced as features. Experiment results show that these methods are sufficient to recognize user's attitude as para-linguistic information. Finally, we show a proto-type spoken dialogue system using para-linguistic information and how these sorts of information contribute the efficient conversation.

1. INTRODUCTION

The human-human interactions in a spoken dialogue seem to use not only linguistic information in the utterances but also some sorts of additional information supporting linguistic information. We call these sorts of additional information "para-linguistic information."

In most of conventional spoken dialogue systems a user and a system exchanges utterances in order, and the system tries to understand user's intention with the linguistic-information in his/her utterance. In an actual spoken dialogue, however, humans exchange their intentions effectively with para-linguistic information along with linguistic information. We sometimes do not express our intention explicitly in words. However, hearers can recognize our intention through our face or voice expressions. It is indispensable to adopt the para-linguistic information, in order to develop a spoken dialogue system which can interact with a human more naturally.

In this paper, we present recognition methods of attitude in prosodic information and head gesture, as para-linguistic information. The spoken dialogue system that combines these information and talks with human effectively is also presented.

The rest of the paper is organized as follows. Section 2 shows para-linguistic information that occurs in a spoken dialogue and the studies focusing on these information. It also shows the task of the dialogue we focused in this study. Section 3 and 4 present the methods and the experimental results of the attitude recognition by prosody and the head gesture recognition respectively. Section 5 presents how these para-linguistic information are combined and show the implemented spoken dialogue system.

2. PARA-LINGUISTIC INFORMATION IN SPOKEN DIALOGUE

2.1. Para-linguistic information

Para-linguistic information in a spoken dialogue are mainly classified into acoustical information and visual information. Acoustical information includes, for example, prosodic information, such as pitch and speed, and linguistically needless words, such as fillers. Visual information includes facial expressions, eye gaze, head gestures and so on.

Among the works on prosodic information in a spoken dialogue, [1] uses it to advance the speed and accuracy of parsing process, [2] and [3] estimate Dialogue Act(DA) in a spoken dialogue with this kind of information. On recognition of head gestures and facial expressions, accurate recognition methods are developed[4][5] and a non-verbal communication using this kind of information has been reported[6]. The database for investigating relationship between spoken language and movement of a head has been collected[7].

0-7803-7980-2/03/$17.00 © 2003 IEEE

2.2. Task of the dialogue

In this study, we deal with the dialogue in which the system helps users to decide something. The system suggest some plans to satisfy the user's explicit or implicit demands through the conversation. We call this kind of dialogue "consulting dialogue".

In this paper, we deal with dialogue in which the system help users to decide a cuisine and a shop for lunch.

3. ATTITUDE RECOGNITION USING PROSODY

3.1. Target

The target is to recognize user's attitude from para-linguistic information. Let's think about the case that the advisor suggested a plan: for example, "How about hamburger." The users often do not express their evaluation explicitly. They simply repeat the advisors' words, say "Hamburger...". Linguistic information, the repeated words, itself do not include users' evaluation (positive/negative) against the plan. However, the evaluation often appears in the expression of face or voice that is para-linguistic information. If this evaluation is automatically recognized by the system, it contributes the efficient and smooth conversation. The aim of this section is to recognize the users' evaluation (positive/negative) of the suggested plan from the prosodic information of the repeated words.

3.2. Data

We recorded the utterances of the users' responses to the system's suggestions. In order to collect a large number of utterances that include the positive or negative attitudes, the users' responses in the single turn (the system's suggestion and user's response) were recorded. The variety of the recorded utterances are seen in Table 1. In the PHRASE column of this table, the underline part represents the repetition of the CATEGORY or SHOP in the system's suggestion. The words "ka" and "ne" are post-positional particles of Japanese, which don't mean positive or negative by themselves. The phrases "iinjanai" and "so-dane" are usually received as positive, but they can be received as negative depending on the way to say. The utterances of 20 students in our laboratory, total 2000 utterances, were recorded.

Table 1. The variety of the recorded utterances

CATEGORY/SHOP(10 kinds)		PHRASE(5 kinds)	ATTITUDE(2 kinds)
Hamburger	McDonald's	_____ ka	positive
Noodles	Ajigen	_____ ne	negative
Lunch Box	Mumin	_____	
Curry	Hoka-ben	iinjanai	
Refectory	Soba-no-mi	so-dane	

For example, in the case that the CATEGORY is "Hamburger," the PHRASE is "____ ka" and the ATTITUDE is negative. First, the user hears the utterance "Hamburger nan te do kana", which means "What about hamburger?". After that, the user utters "Hamburger ka" with negative attitude, and we record it.

3.3. Recognition method

We apply the fundamental frequency (F0) extraction and the phoneme alignment to the recorded utterances. As consideration of these results, we decided to use the following 3-dimensional feature $x = (x_1, x_2, x_3)$ to distinguish the attitude.

x_1 : the gradient of F0 at the vowel part of the first mora in the utterance

x_2 : the range of F0 throughout the utterance

x_3 : the time length of the last mora in the utterance

x_1, the gradient is calculated with the least square method. The filler phrase at the beginning of the utterance, for example "ah", "un" and so forth, was accepted, so we ignored it in the feature extraction.

In Fig.1, an example of feature extraction is seen. In this figure, the following tendency of features can be seen.

- x_1 is a positive number if the attitude is positive, or a negative number otherwise.
- x_2 is larger if the attitude is positive than otherwise.
- x_3 is smaller if the attitude is positive than otherwise.

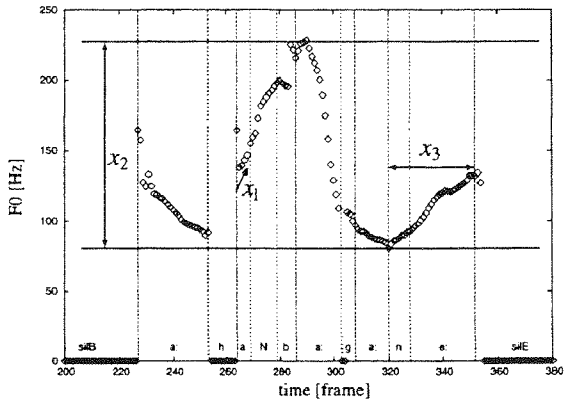
An utterance is recognized as positive or negative with Bayesian classifier. The normal distribution is calculated with the features for each category.

3.4. Experiment and results

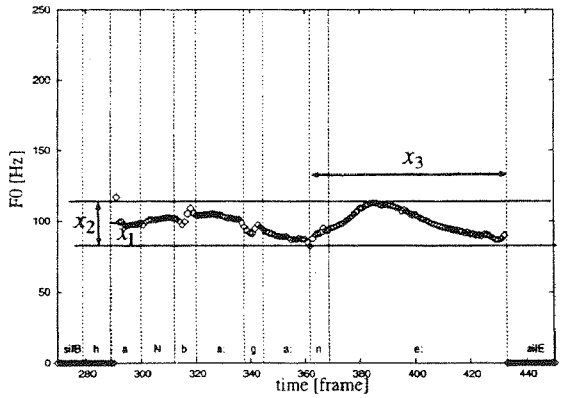
The recorded utterances are divided into 4 sets. Each set contains 5 person's utterances, total 500 utterances. In one experiment process, we use 3 sets as a training set and the remain as a test set. This process was performed for each set, total 4 times, and we calculated the average recognition rate of the all results. In Fig.2 the results are seen.

In order to capture differences between phrases, we experimented the data without the phrase "iinjanai" whose last mora is always short independently of the attitude, the data that contains only the phrase "____ ka" and "____ ne" which seem to contain similar features and the data that contains only the phrase "____ ka". These experiments results are also seen in Fig.2. In Fig.3, the experiment result of each person is seen.

In Fig.2, the rise of the recognition rate is acquired by limiting a phrase. However, if the limitation is too much,



(a) positive



(b) negative

Fig. 1. An example of prosodic feature extraction

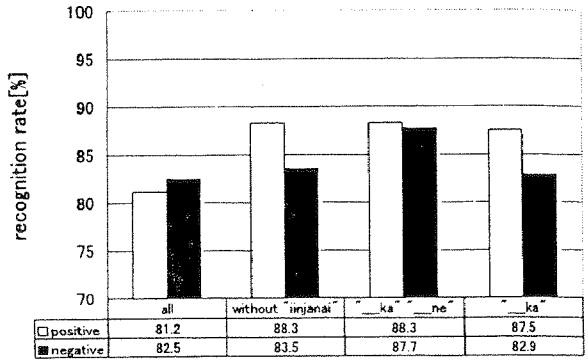


Fig. 2. Recognition results of the attitude recognition by prosody

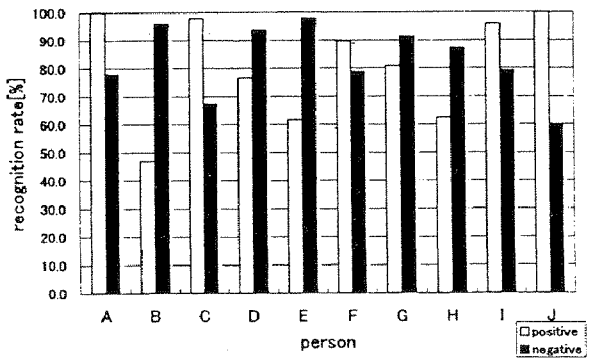


Fig. 3. Recognition results of selected 10 persons

the rate falls down. The cause of fall is insufficiency of the training data. In Fig.3, recognition rates are different for each person. While for some persons the positive result is better than the negative one, the negative result is better for others.

3.5. Comparison with recognitions by human

In the experiment in the previous section, the correct answer is the attitude which the speaker supposed when he spoke. In order to compare the recognition ability of our model with that of a human, we calculate a measure of agreement between results of humans' annotations and the experiment results.

Among the recorded utterances, we picked up 20 utterances from each category, positive and negative. 5 annotators (A–E) listened to the utterances in random sequence and decided the attitude, positive or negative, for each.

We introduce Cohen's κ [8] as a measure of agreement. It is often used as an evaluation measure of a dialogue corpus annotation. Since Cohen's κ is a value calculated from 2 annotations, we calculated it for each combination of the annotations including the experiment result(M). The minimum, maximum and average of the rates with others was calculated for each person. The result is seen in Table 2.

Although the machine(M) is ranked fifth in average, the maximum and minimum values are similar to humans'. As evaluating with Cohen's κ, the agreement is good if $0.60 < \kappa < 0.75$. In this point of view, this result is sufficient to say that the model has the recognition ability similar to human's.

4. HEAD GESTURE RECOGNITION

4.1. Target

Head gestures, for example, nod, tilt, and shake are often generated with rephrases. They represent users' evaluation as well as the prosodic information described in previous

Table 2. Calculation results of Cohen's κ

	min	max	average
M	**0.52**	**0.80**	**0.66**
A	0.52	0.85	0.72
B	0.52	0.79	0.62
C	0.65	0.80	0.72
D	0.65	0.85	0.76
E	0.61	0.79	0.72

section. The aim of this section is to recognize the users' head gestures.

4.2. Categories to recognize

At first, in order to observe what kind of head gesture occurs in a natural dialogue, face-to-face dialogues of humans were recorded. In these dialogues, the following three gestures often occur as a feedback to the dialogue partner.

nod representing a positive response to the question, a sympathy or comprehension to partner's utterance and a encouragement of partner's sympathy to his utterance.

tilt representing a negative response to partner's utterance, a lack of confidence to his utterance and a state that he is in thought.

shake representing a negative response to partner's utterance.

The recognition categories here are the three categories above.

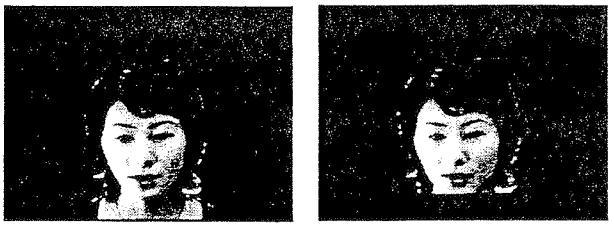
4.3. Recognition method

In this method, the movement of head is used as feature. The head region in an input image (in Fig.4(a)) is extracted by skin color and hair color information(in Fig.4(b)). Furthermore, the neck region is eliminated by introducing the ratio of width and height of head region(in Fig.4(c)). Optical-flows of the all pixels in the extracted head region are treated as values representing a head movement.

As in Fig.5, a head region is divided into 4 regions. The average of optical-flows in each region is a feature to recognize the head gesture. Therefore, total dimension of a feature vector is 8. It is expected to capture the feature of each head gesture, as in Fig.6.

In order to recognize the gesture, we introduce Viterbi algorithm in left to right HMM(Hidden Markov Model). In this study, in order to recognize head gestures during a spoken dialogue, spotting is required. 2 models, stillness model and garbage model, are introduced for spotting. Stillness model represents stopping of a head and garbage model represents other movements of a head than the three gestures.

(a) original image



(b) extraction with color model (c) elimination of neck region

Fig. 4. Extraction of head

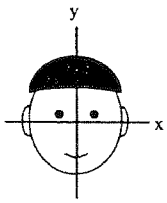


Fig. 5. Division of a head domain

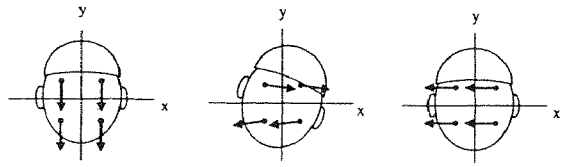
4.4. Experiment and result

The recorded data for the experiment consist of the natural dialogues described in Section 4.2(15 sets, the length of each set is 4 minutes), and additional new data(25 sets, the total length 144 minutes). Thus, the total amount of the data is 40 sets, 204 minutes. The 25 sets of additional data were recorded the gestures of the users which was directed to do the desired gestures. The recording continued during the direction, so gestures occurring naturally were recorded. The data contains totally 2148 gestures, except stillness and garbage.

The data were divided into 35 training sets and 5 test sets. Experiments were done by each test set, total 8 times. In Fig.3, results are seen in confusion matrix.

Deletion represents the number of times that a head gesture was recognized as a stillness or a garbage. Insertion represents the number of times that a stillness or a garbage was recognized as a head gesture.

In this experiment, the number of HMM states in each



(a) nod (b) tilt (c) shake

Fig. 6. Feature of each gesture

model is 10, the number of mixtures in head gesture model is 8, that in stillness model is 4 and that in garbage model is 16. These are the parameters that brought the best result in the experiments with various parameters.

Table 3. Experiment result of head gesture recognition

	Recognition result			
	nod	tilt	shake	deletion
nod	**1030**	2	0	293
tilt	16	**341**	14	148
shake	3	5	**263**	41
insertion	536	216	134	

Result shows that insertion errors often occurs in the recognition. Therefore, when we apply this result to a spoken dialogue system, these errors must be corrected. The simplest way may be rejecting the result while the user was not talking.

5. SYSTEM

5.1. Combining the information

In order to adopt the results of the attitude recognition by prosody described in Section 3 and the head gesture recognition described in Section 4 to a spoken dialogue system as para-linguistic information, the results must be combined. Particularly, when the results are contrary with each other, one is positive while the other is negative, it's serious problem to decide which result is plausible.

In this study, we introduce the combination as in Table. 4. For example, If the result of the head gesture recognition is "nod" and the result of the attitude recognition by prosody is "positive", the user's attitude is recognized as "strongly positive", while if the latter is "negative", it is recognized as "in thought", which represents the user cannot decide clearly.

Table 4. combining strategy of the recognition results (* indicates that this combination hardly occurs in a dialogue)

		head gesture		
		nod	tilt	shake
prosody	positive	strongly positive	weakly positive	strongly negative*
	negative	in thought	negative	strongly negative

5.2. Proto-type spoken dialogue system

We implemented a proto-type spoken dialog system with para-linguistic information on humanoid robot ROBISUKE (Fig. 7). We have developed humanoid robots that interact with humans with their perceptual and expressive abilities, such as a speech recognition, an individual identification using face images and facial expressions using its eyes, eyebrows and mouth[9][10]. ROBISUKE is the most recently developed humanoid robot in our group.

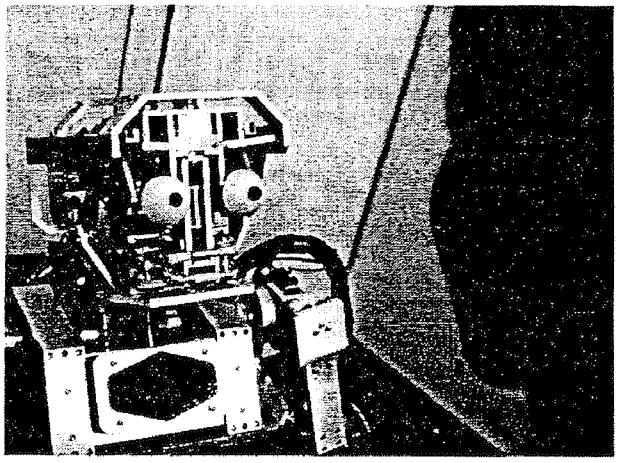


Fig. 7. ROBISUKE

In this system, user's responses to the system's suggestions are combined with the method described in section 5.1, and the system makes the dialogue proceed according to the decisions listed as follows.

- strongly positive : the suggestion is accepted, so that the system gives the details for sure.
- weekly positive : the suggestion is almost accepted, so that the system gives the details.
- in thought : the decision is delayed, so that the system

says "un(yes)" or just nods its head to receive another response by the user.

- negative or strongly negative : the suggestion is rejected, so that the system gives another suggestion if possible.

5.3. Example

An example dialogue is seen in Fig. 8.

The system suggests another idea when the user's response is negative to the previous suggestion. It suggests the store in the suggested category when the user's response is positive to the previous suggestion.

In Fig.8, the responses "Curry ka" and "Bento ne" which cannot be decided positive or negative with the only linguistic information, are appropriately interpreted using para-linguistic information.

U: Ohiru gohan nandakedo, doko ka ii tokoro naikana
(I'd like going to lunch. Do you have some idea?)
R: Curry nante do
(What about curry?)
U: Curry ka **(negative)**
(Curry)
R: Sorejaa, hamburger nante doukana
(Well then, what about hamburger)
U: Ah, hamburger ne **(strongly positive)**
(Oh, hamburger)
R: Hamburger nara chikaku ni McDonald ga aruyo
(For hamburger, there's McDonald's nearby)

Fig. 8. An example dialogue
User's utterances and ROBISUKE's utterances are preceded by U: and R: respectively.

6. CONCLUSION

In this paper, we implemented the attitude recognition by prosodic information in utterances, and the head-gesture recognition, as the para-linguistic information in spoken dialogue. Experimental results show that the recognition ability of the former method is similar to human's. We show a spoken dialogue system with para-linguistic information, and these sorts of information influence the progress of a dialogue effectively.

We aim to develop recognition methods of more information, for example, the attitudes recognition in more kinds of utterances and the facial expression and gaze direction recognition, which are accompanied as the para-linguistic information. Evaluation of the spoken dialogue system, particularly on the dialogue quality, is the most important issue.

7. REFERENCES

[1] Christian Lieske, Johan Bos, Martin Emele, Björn Gambäck, and CJ Rupp, "Giving prosody a meaning," in *Proceedings of ISCA EUROSPEECH'97*, 1997, vol. 3, pp. 1431–1434.

[2] E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann, "On the use of prosody in automatic dialogue understanding," *Speech Communication*, vol. 36, pp. 45–62, 2002.

[3] Helen Wright Hastie, Massimo Poesio, and Stephen Isard, "Automatically predicting dialogue structure using prosodic features," *Speech Communication*, vol. 36, pp. 63–79, 2002.

[4] Shinji Kawato and Jun Ohya, "Real-time detection of nodding and head-shaking by directly detecting and tracking the 'between-eyes'," in *Proceedings of Fourth IEEE international conference on automatic face and gesture recognition*, 2000, pp. 40–45.

[5] Ashish Kapoor and Rosalind W. Picard, "A real-time head nod and shake detector," Tech. Rep. 544, MIT Media Laboratory Affective Computing Group, 2002.

[6] Hiroshi Kobayashi and Fumio Hara, "Facial interaction between animated 3d face robot and human beings," in *Proceedings of 1997 IEEE International Conference on Systems, Man and Cyberntics(SMC97)*, 1997, vol. 4, pp. 3732–3737.

[7] S. Hayamizu, O. Hasegawa, K. Itou, K. Sakaue, K. Tanaka, S. Nagaya, M. Nakazawa, T. Endoh, F. Togawa, K. Sakamoto, and K. Yamamoto, "Rwc multimodal database for interactions by integration of spoken language and visual information," in *Proceedings of Fourth International Conference on Spoken Language (ICSLP96)*, 1996, vol. 4, pp. 2171–2174.

[8] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[9] Tsuyoshi Tojo, Yosuke Matsusaka, Tomotada Ishii, and Tetsunori Kobayashi, "A conversational robot utilizing facial and body expressions," in *Proceedings of 2000 IEEE International Conference on Systems, Man and Cyberntics(SMC2000)*, October 2000, vol. 2, pp. 858–863.

[10] Yosuke Matsusaka and Tetsunori Kobayashi, "System software for collaborative development of interactive robot," in *Proceedings of IEEE-Humanoids2001*, 2001, pp. 271–277.

A Method to Add Emotional Expressions to Text-to-Speech Synthesis

Kei Yazu, Makoto Murakami, Futoshi Sugimoto, and Masahide Yoneyama
Department of Information and Computer Sciences, Toyo University
2100 Kujirai Kawagoe, Saitama 350-8585 JAPAN
+81-49-239-1448
f_sugi@eng.toyo.ac.jp

ABSTRACT
Representing emotional expressions in text-to-speech synthesis has been a contemporary subject. We aim to add emotional expressions to synthetic speech. This system recognizes emotion based on the words in a sentence and synthesizes emotional speech.

Keywords
Prosody, text-to-speech synthesis, morphological analysis

INTRODUCTION

In recent years, since the technologies for text-to-speech synthesis have been developed [1], it has become easier to get the synthesized speech that is comparatively natural. However, there are many unsolved problems in the theme to add emotion to synthesized speech. In our research, we have been developing a system that judges emotion of a sentence based on the words included in the sentence, and read it aloud with emotion. The emotion expressed in a speech can be controlled by operating three factors: tempo, pitch and power. Although the relation between these factors and emotion is almost clear qualitatively, we attempt to make the relation clear quantitatively in this paper.

LIMITATION OF EMOTIONS AND WORDS

In this paper, we limit the emotions that are objects of the classification and the words to use. The emotions are joy, sorrow, anger and surprise that strongly characterize speech and neutral emotion that is located at the center of these emotions. In addition, we use only words contained in the IPA Lexicon of the Japanese Language for computers (IPAL). IPAL was compiled by the Information-technology Promotion Agency (IPA) in 1990. It contains 861 entries for basic Japanese verbs, 136 entries for basic Japanese adjectives, and 1,081 entries for basic Japanese nouns. Those are considered representative examples of Japanese words. Each entry includes information about semantics, morphology, grammatical categories, case frames and idiomatic usage [2].

EMOTIONAL SPEECH SYNTHESIS SYSTEM

The block diagram of the emotional speech synthesis system which we have been developing is shown in Figure 1. The system is organized into three parts, morpheme analysis, emotion classification and speech synthesis.

Morpheme Analysis

ChaSen is a morphological analyzer of Japanese sentences that wais released by the Nara Institute of Science and Technology (NAIST) [3]. We divide a sentence into words using the analyzer, and extract only adjectives from them

Emotion Classification

The final aim of the emotion classification in our research is

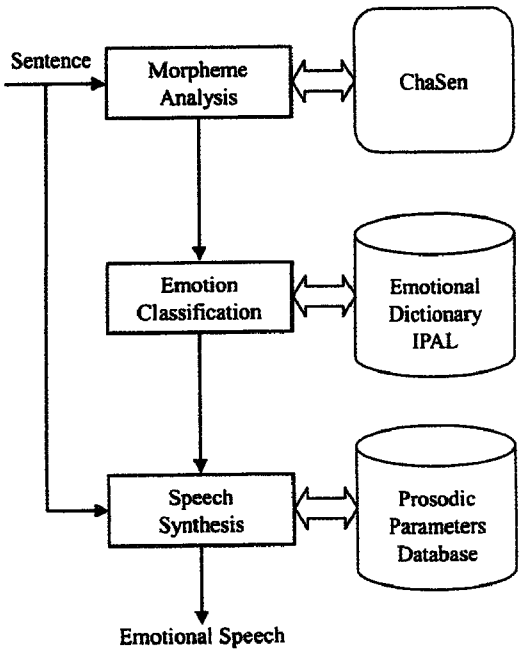


Figure 1 Emotional speech synthesis system

as follows. We classify the emotion of sentences based on the relation among all words constituting the sentence. These words are nouns, verbs, adjectives, and adverbs modifying them.

In the present phase, however, we focus on only adjectives in a sentence and decide the emotion of the sentence based on the emotion that those adjectives have. When each of the plural adjectives has different emotions, we adopt the emotion of the adjective which has the biggest degree of emotion as the emotion of a sentence. 31 adjectives related to the four emotions mentioned above among 136 adjectives contained in IPAL were used to classify emotions.

PROSODIC FACTORS

The emotion of speech can vary by controlling the following three parameters;

- Pitch; basic frequency
- Power; amplitude of spectrum
- Tempo; utterance time

The relation between emotions and parameters to control them has been studied already, and the abstract is shown in Table 1. The characteristics that are relative to neutral emotion are described. It is not clear about anger, but characteristics are estimated from a result of our study.

EVALUATION OF PROSODIC PARAMETERS

Experimental Method

An experiment to evaluate quantitatively the qualitative characteristics of the prosodic parameters, pitch, power and tempo, shown in Table 1, was conducted using a speech synthesis software, Visual Speech Creator [4]. While listening to a Japanese sentence in the speech synthesis software, subjects adjusted the prosodic parameters interactively so that the sentence could express each emotion to the greatest extent. Subjects were asked to adjust prosodic parameters in the whole sentence uniformly, but they were allowed to adjust them partially when they were not satisfied. Ten subjects were engaged in this experiment.

Experimental Results

The average of all subjects according to every prosodic parameter is shown in table 2. The numerical value in parenthesis is a magnification in comparison with the value of the parameter of natural emotion. Four subjects lifted pitch and power partially at the end of a sentence in order to express "Anger".

CONCLUSION

The values of parameters that were measured in the experiment shown in Table 2 are in almost the same correlation as the qualitative characteristics in Table 1. The synthesized speech expresses emotion well except that it is difficult to express "anger".

FUTURE WORKS

In this paper, the emotion of a sentence was decided based on only adjectives in the sentence, but we are attempting to classify the emotion of sentences based on the relations among all words: nouns, verbs, adjectives and adverbs, constituting the sentence. Furthermore, although the prosodic parameters were set uniformly in the whole sentence, it is necessary to consider that we control the parameters separately in every word and clause in order to get closer to human speech.

REFERENCES

1. Charpentier, F. and Moulines, E., Pitch Synchronous Wave form Processing Techniques for Text-to-Speech Synthesis using Diaphones, Proc. Eurospeech '89, 1989.
2. IPA Lexicon of the Japanese Language for computers, Information-technology Promotion Agency, http://www.ipa.go.jp/STC/NIHONGO/IPAL/ipal.html, 1990.
3. ChaSen, Nara Institute of Science and Technology (NAIST), http://chasen.aist-nara.ac.jp/index.html.en, 1999.
4. Visual Speech Creator, NTT-IT, http://www.ntt-it.co.jp/main.html.

Table 1 Relation between emotions and parameters

	Neutral Emotion	Joy	Sorrow	Anger	Surprise
Pitch	Normal	High	Slightly High	High	High
Power	Normal	Large	Small	Large	Normal
Tempo	Normal	Fast	Slow	Fast/Slow	Fast

Table 2 Result of experiment

	Neutral Emotion	Joy	Sorrow	Anger	Surprise
Pitch (Hz)	280	234 (0.8)	358 (1.2)	396 (1.4)	354 (1.2)
Power	50	33 (0.6)	80 (1.6)	80 (1.6)	65 (1.2)
Tempo (Sec)	3.765	4.314 (1.2)	3.381 (0.9)	3.021 (0.8)	3.381 (0.9)

Spoken Dialogue System Using Prosody as Para-Linguistic Information

Shinya Fujie[1], *Daizo Yagi*[1], *Yosuke Matsusaka*[1], *Hideaki Kikuchi*[2] & *Tetsunori Kobayashi*[1]

[1]School of Science and Engineering, Waseda University, Japan
[2]School of Human Sciences, Waseda University, Japan
{fujie, yagi, yosuke, koba}@tk.elec.waseda.ac.jp, kikuchi@waseda.jp

Abstract

An attitude recognizer of a speaker which uses prosodic features of speech is proposed and it is successfully applied to the dialogue system aiming at agreement formation. We use not only linguistic information but also some sorts of additional information supporting linguistic information in our human communication. In agreement formation dialogues, we are often required to express our attitude (positive or negative) to conversational partners' proposals. We sometimes reply explicitly in linguistic information. We sometimes reply information ambiguously. However, even in the ambiguous case, we implicitly express our attitude using prosodic information. By realizing the abilities of catching these nuances, the dialogue system can be more sophisticated. In this paper, we implemented an attitude recognizer based on the GMM using prosodic feature parameters. The performance of the system is comparable to the human ability. We also realized a proto-type of spoken dialogue system using the recognizer. We show how these abilities contribute to efficient conversation.

1. Introduction

In a spoken dialogue-based human communication, we use not only linguistic information but also some sorts of additional information supporting linguistic information. We call these sorts of additional information "para-linguistic information."

In most conventional spoken dialogue systems a user and a system exchanges utterances in order, and the system tries to understand the user's intention with the linguistic-information in his/her utterance. In an actual spoken dialogue, however, humans exchange their intentions effectively with para-linguistic information along with linguistic information. We sometimes do not express our intention explicitly in words. Nevertheless, hearers can recognize our intention through our face or voice expressions. Therefore, it is indispensable that a natural spoken dialogue system has the ability to recognize para-linguistic information.

Delleart et al[1] recognized four emotions, such as "happy", "sad", "anger", and "fear", by prosody. In this work, 17 pitch features divided into 5 groups are introduced. The only 5 features selected in order by the performance of the cross-validation experiment, achieved better performance than the original features. The majority voting of specialists method, where the original feature space is divided into the small sub spaces and the results of the sub spaces are combined, marked better as well. McGilloway et al[2] gave a benchmark of automatic emotion recognition using features by ASSESS system by Cowie. Lee et al[3] recognized negative and non-negative emotions using prosody, in order to improve the quality of the service in call center applications. In this work a total of 10 features representing F0 and energy information are introduced. The features reduced the dimension (6 or 7) by PCA and feature selection(the same method with [1]) marked better than the original base features. Ang et al[4] used prosody for the detection of frustration and annoyance in natural human-computer dialogue. They used the corpus of human-computer dialogue developed under the DARPA Communicator Project. The prosodic features used in this work are durations, speaking rate, pause, pitch and so on. They introduced language model feature as well. For more literature of the works on emotion recognition, refer [5].

The keyword of the categories to recognize in these works is "Emotion." The recognition results are important for estimating user's "real" emotional state, but we think that the behavior of the users and the intended message included implicitly as para-linguistic information in utterance are more important for effective and smooth conversation. In addition, the real emotion is not intuitively applicable for the dialogue strategy.

In order to make human-machine conversation more effective and smooth, we incorporate some methods to recognize para-linguistic information and treat those as user's implicit messages. In this paper, we introduce a method for recognizing a user's attitude (positive or negative) from prosody. This information can be used as the user's evaluation against the system's suggestion. The spoken dialogue system that combines these sorts of information, including head gesture, and talks with human effectively is also presented.

The rest of the paper is organized as follows. Section 2 presents the methods and the experimental results of the attitude recognition by prosody. Section 3 presents how some sorts of para-linguistic information are combined and show the implemented spoken dialogue system.

2. Attitude Recognition Using Prosody

2.1. Target

The target is to recognize a user's attitude using para-linguistic information in the agreement formation dialogue, in which a user decides a plan through interaction with an advisor who suggests some plans. Please imagine the case that the advisor suggests a plan; for example he proposes a hamburger for lunch, "How about hamburger." Often the users do not express their evaluation explicitly. They simply repeat the advisors' words, say "Hamburger...". Linguistic information, the repeated words, itself do not include the users' evaluation (positive/negative) against the plan. However, the evaluation often appears in the expression of face or voice, that is called para-linguistic information. If this evaluation is automatically recognized by the system, it contributes to efficient and smooth conversation. The aim of this section is to give a method to

recognize users' evaluation (positive/negative) to the suggested plan using the prosodic information of the repeated words.

2.2. Data

We recorded the utterances of the users' responses to the system's suggestions. In order to collect a large number of utterances that include the positive or negative attitudes, the users' responses in the single turn (the system's suggestion and user's response) were recorded. The variety of the recorded utterances are seen in Table 1. In the PHRASE column of this table, the underlined part represents the repetition of the CATEGORY or RESTAURANT in the system's suggestion. The words "ka" and "ne" are post-positional particles of Japanese, which do not define positive or negative by themselves. The phrases "iinjanai" and "so-dane" are usually received as positive, but they can be received as negative depending on the use. The utterances of 20 students in our laboratory, total 2000 utterances, were recorded.

Table 1: *The variety of the recorded utterances.*

CATEGORY/RESTAURANT		PHRASE	ATTITUDE
Hamburger	McDonald's	____ ka	positive
Noodles	Ajigen	____ ne	negative
Lunch Box	Mumin	____	
Curry	Hoka-ben	iinjanai	
Refectory	Soba-no-mi	so-dane	

For example, in the case that the CATEGORY is "Hamburger," the PHRASE is "____ ka" and the ATTITUDE is negative. First, the user hears the utterance "Hamburger nan te do kana", which means "What about a hamburger?". After that, the user utters "Hamburger ka" with negative attitude, and we record it.

2.3. Recognition method

We apply the fundamental frequency (F0) extraction and the phoneme alignment to the recorded utterances. As consideration of these results, we decided to use the following 3-dimensional feature $\boldsymbol{x} = (x_1, x_2, x_3)$ to distinguish the attitude.

x_1 : the gradient of F0 at the vowel part of the first mora in the utterance

x_2 : the range of F0 throughout the utterance

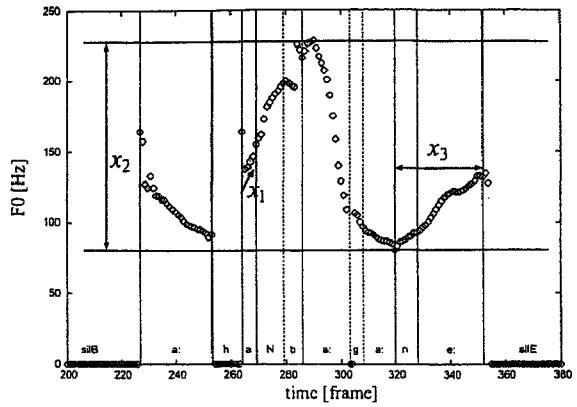
x_3 : the time length of the last mora in the utterance

x_1, the gradient is calculated with the least square method. The filler phrase at the beginning of the utterance, for example "ah", "un" and so forth, was accepted, so we ignored it in the feature extraction.

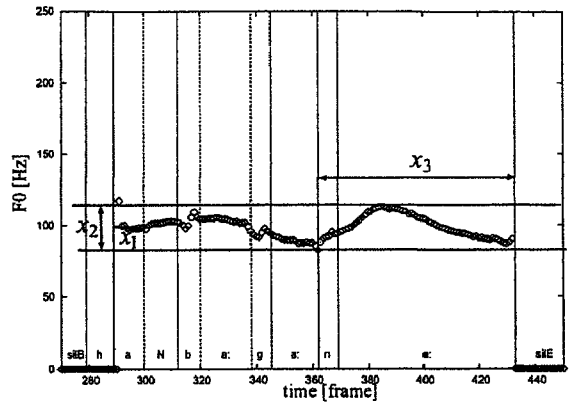
In Fig.1, examples of feature extraction are seen. In this figure, the following tendency of features can be seen.

- x_1 is a positive number if the attitude is positive, or a negative number otherwise.
- x_2 is larger if the attitude is positive than otherwise.
- x_3 is smaller if the attitude is positive than otherwise.

An utterance is recognized as positive or negative with Bayesian classifier.



(a) positive



(b) negative

Figure 1: *Examples of prosodic feature extraction.*

2.4. Experiment and results

The recorded utterances are divided into 4 sets. Each set contains 5 person's utterances, total 500 utterances. In one experiment process, we use 3 sets as a training set and the remainder as a test set. This process was performed for each set, total 4 times, and we calculated the average recognition rate of all the results.

In Fig.2, the experiment result of each person is shown.

We aimed to reduce the error caused by the differences between phrases and that between persons, by introducing Gaussian Mixture Model(GMM) as the probabilistic model. In Fig.3, the experiment result of GMM is seen.

In Fig.2, recognition rates are different for each person. While for some persons the positive result is better than the negative one, the negative result is better for others. In Fig.3, the maximum recognition rate (82.9%) is marked when the number of mixture is 16.

2.5. Comparison with recognitions by human

In the experiment in the previous section, the proper result is the attitude which the speaker implied when he spoke. In order to compare the recognition ability of our model with that of a human, we calculate a measure of agreement between results of humans' annotations and the experiment results.

Among the recorded utterances, we picked up 20 utterances from each category, positive and negative. 5 annotators (A–E)

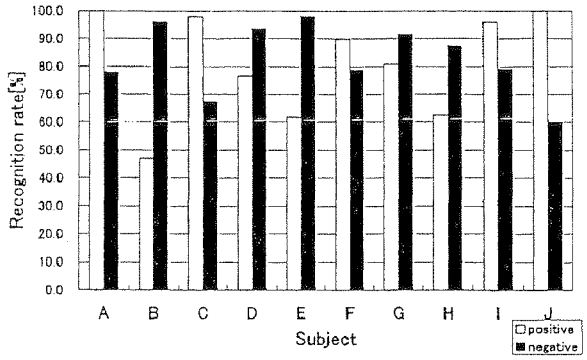


Figure 2: *Recognition results of selected 10 persons.*

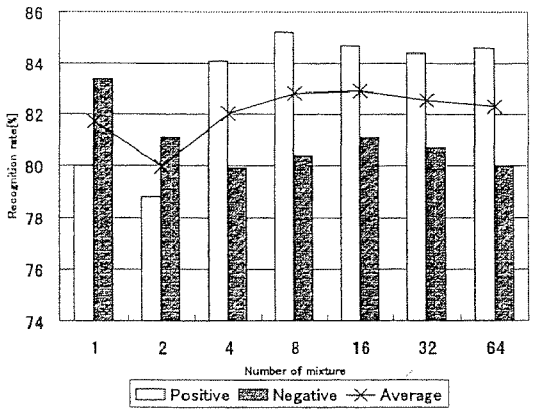


Figure 3: *Recognition results of GMM(Gaussian Mixture Model.)*

listened to the utterances in random sequence and decided the attitude, positive or negative, for each.

We introduce Cohen's κ [6] as a measure of agreement. It is often used as an evaluation measure of a dialogue corpus annotation. Since Cohen's κ is a value calculated from 2 annotations, we calculated it for each combination of the annotations including the experiment result(M). The minimum, maximum and average of the rates with others was calculated for each person. The result is seen in Table 2.

Table 2: *Calculation results of Cohen's κ.*

	min	max	average
M	**0.52**	**0.80**	**0.66**
A	0.52	0.85	0.72
B	0.52	0.79	0.62
C	0.65	0.80	0.72
D	0.65	0.85	0.76
E	0.61	0.79	0.72

Although the machine(M) is ranked fifth on average, the maximum and minimum values are similar to humans. As evaluating with Cohen's κ, the agreement is good if $0.60 < \kappa < 0.75$. In this point of view, this result is sufficient to say that the model has the recognition ability similar to human's.

3. SYSTEM

3.1. Combining the information

We aimed to adopt the results of the attitude recognition by prosody described in Section 2 to a spoken dialogue system as para-linguistic information. In our previous study, a head gesture recognition system has been implemented [7]. This system is able to recognize several types of head gestures, such as "nod", "tilt", and "shake," using optical flow as the feature and HMM as the probabilistic model. As well as prosodic information, these types of head gestures express the utterer's attitude well. We introduce results of a head gesture recognition to a spoken dialogue system as para-linguistic information. We now must combine both results to estimate the user's correct attitude. Particularly, when the results are contrary to each other, one is positive while the other is negative, it is a serious problem to decide which result is plausible.

In this study, we introduce the combination as in Table. 3. For example, If the result of the head gesture recognition is "nod" and the result of the attitude recognition by prosody is "positive", the user's attitude is recognized as "strongly positive", while if the latter is "negative", it is recognized as "in thought", which represents that the user cannot decide clearly.

Table 3: *Combining strategy of the recognition results.* (* indicates that this combination hardly occurs in a dialogue)

		head gesture		
		nod	tilt	shake
prosody	positive	strongly positive	weakly positive	strongly negative*
	negative	in thought	negative	strongly negative

3.2. Proto-type spoken dialogue system

We implemented a proto-type spoken dialog system with para-linguistic information on a humanoid robot ROBISUKE (Fig. 4). We have developed humanoid robots that interact with humans with their perceptual and expressive abilities; such as speech recognition, individual identification using face images, and facial expressions using its eyes, eyebrows and mouth[9]. ROBISUKE is the most recent humanoid robot developed in our group.

In this system, user's responses to the system's suggestions are combined with the method described in section 3.1, and the system makes the dialogue proceed according to the decisions listed as follows.

- strongly positive : the suggestion is accepted, so that the system gives the details promptly.
- weakly positive : the suggestion is almost accepted, so that the system gives the details.
- in thought : the decision is delayed, so that the system says "un(yes)" or just nods its head to receive another response by the user.
- negative or strongly negative : the suggestion is rejected, so that the system gives another suggestion if possible.

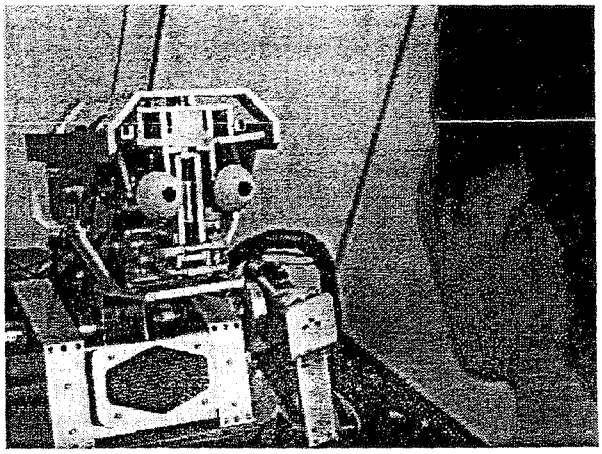


Figure 4: Conversation robot ROBISUKE.

3.3. Example

An example dialogue is seen in Fig. 5.

The system suggests another idea when the user's response is negative to the previous suggestion. It suggests the restaurant in the suggested category when the user's response is positive to the previous suggestion.

In Fig.5, the responses "Curry ka" and "Bento ne" which cannot be decided positive or negative with the only linguistic information, are appropriately interpreted using para-linguistic information.

U: Ohiru gohan nandakedo, doko ka ii tokoro naikana
(I'd like to go to lunch. Do you have an idea?)
R: Curry nante do
(What about curry?)
U: Curry ka **(negative)**
(Curry)
R: Sorejaa, hamburger nante doukana
(Well then, what about hamburger)
U: Ah, hamburger ne **(strongly positive)**
(Oh, hamburger)
R: Hamburger nara chikaku ni McDonald ga aruyo
(For a hamburger, there's a McDonald's nearby)

Figure 5: An example dialogue.
User's utterances and ROBISUKE's utterances are preceded by U: and R: respectively.

4. Conclusion

In this paper, we implemented the attitude recognition by prosodic information in utterances as the para-linguistic information in spoken dialogue. Experimental results show that the recognition ability is similar to human's. We show a spoken dialogue system with para-linguistic information, and these sorts of information influence the progress of a dialogue effectively.

We aim to develop recognition methods of more information, for example, the attitudes recognition in several kinds of utterances and the facial expression and gaze direction recognition, which are accompanied as para-linguistic information. Evaluation of the spoken dialogue system, particularly on the dialogue quality, is the most important issue.

j

5. References

[1] Dellaert, F.; Polzin, T.; Waibel, A., 1996. Recognizing emotion in speech: in *Proceedings of ICSLP '96*,

[2] McGilloway, S.; Cowie, R.; Cowie, E. D.; Gielen, S.; Westerdijk, M.; Stroeve, S., 2000. Approaching automatic recognition of emotion from voice: a rough benchmark: in *Proceedings of ISCA workshop on Speech and Emotion*.

[3] Lee, C. M.; Narayanan, S.; Pieraccini, R., 2001. Recognition of negative emotions from the speech signal: in *Proceedings of IEEE ASRU2001*, 240–243.

[4] Ang, J.; Dhillon, R.; Krupski, A.; Shriberg, E.; Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog: in *Proceedings of ICSLP2002*, 2037–2040.

[5] Cowie, R.; Cowie, E. D.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. G., 2001. Emotion recogntion in human-computer interaction: *IEEE Signal Processing Magazie*, 18(1), 32–80.

[6] Cohen, J., 1960. A coefficient of agreement for nominal scales: *Educational and Psychological Measurement*, 20(1), 37–46.

[7] Ejiri, Y.; Matsusaka, Y.; Kobayashi, T., 2002. Recognition of the head gesture under dialog: *Tehcnical Report of IEICE*, Vol.102, No.218, PRMU2002-61, 31–36. (*in Japanese*)

[8] Tojo, T.; Matsusaka, Y.; Ishii, T.; Kobayashi, T., 2000. A conversational robot utilizing facial and body expressions, in *Proceedings of 2000 IEEE International Conference on Systems, Man and Cyberntics(SMC2000)*, Vol. 2, 858–863.

[9] Matsusaka, Y.; Kobayashi, T., 2001. System software for collaborative development of interactive robot, in *Proceedings of IEEE-Humanoids2001*, 271–277.

A METHOD TO CLASSIFY EMOTIONAL EXPRESSIONS OF TEXT AND SYNTHESIZE SPEECH

Futoshi Sugimoto, Kei Yazu, Makoto Murakami, and Masahide Yoneyama

Department of Information and Computer Sciences, Toyo University
2100 Kujirai Kawagoe, Saitama 350-8585 JAPAN
+81-49-239-1448
f_sugi@eng.toyo.ac.jp

ABSTRACT

Representing emotional expressions on text-to-speech synthesis has been a interesting subject. We aim to add emotional expressions to synthetic speech. This system recognizes emotion based on the words in a sentence and synthesizes emotional speech.

1. INTRODUCTION

In recent years, since the technologies for text-to-speech synthesis have been developed [1], it has become easy to get the synthesized speech that is comparatively natural. However, there are many unsolved problems in the theme to add emotion to synthesized speech. In our research, we have been developing a system that judges the emotion of a sentence based on the words included in the sentence, and reads it aloud with the emotion. The emotion expressed in a speech can be controlled by operating three factors: tempo, pitch and power. Although the relation between these factors and emotion is almost clear qualitatively, we attempt to make the relation clear quantitatively in this paper.

2. LIMITATION OF EMOTIONS AND WORDS

In this paper, we limit the emotions that are objects of classification and words used. The emotions are joy, sorrow, anger and surprise that strongly characterize the speech and neutral emotion that is located at center of these emotions. In addition, we use only words contained in IPA Lexicon of the Japanese Language for computers (IPAL). IPAL has been compiled by the Information-technology Promotion Agency (IPA), in 1990. It contains 861 entries for basic Japanese verbs, 136 entries for basic Japanese adjectives, and 1,081 entries for basic Japanese nouns. Those are considered representative examples of Japanese words. Each entry includes information about semantics, morphology, grammatical categories, case frames and idiomatic usage [2].

3. EMOTIONAL SPEECH SYNTHESIS SYSTEM

The block diagram of the emotional speech synthesis system which we have been developing is shown in Figure 1. The system is organized into three parts, morpheme analysis, emotion classification and speech synthesis.

3.1. Morpheme Analysis

Since there are not blank spaces between characters in Japanese sentences, we need to divide a sentence into

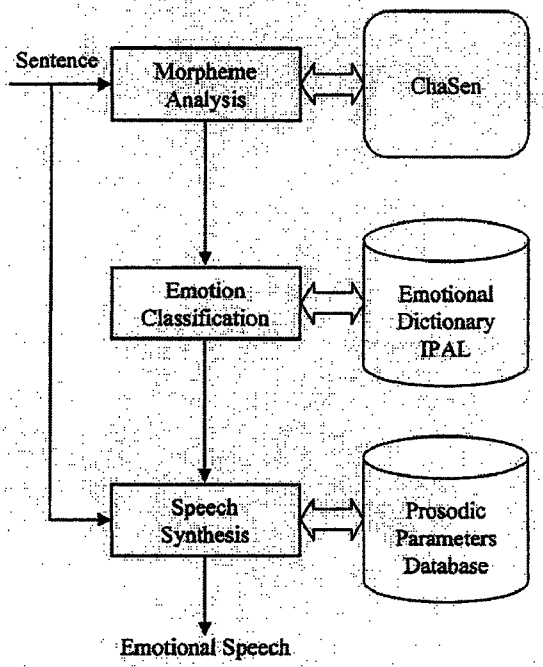


Figure 1 Emotional speech synthesis system

0-7803-8379-6/04/$20.00 ©2004 IEEE.

words (morphemes), and assign a part of speech to each word to analyze Japanese.

ChaSen is a morphological analyzer of Japanese sentences that has been released by the Nara Institute of Science and Technology (NAIST) [3]. We divide a sentence into words using the analyzer, and extract only nouns, adjectives, and verbs among them. An example is as follows:

我々は日本から来ました。(We came from Japan.)

⇩

我々 は 日本 から 来 まし た 。

3.2. Emotion Classification

The final aim of the emotion classification in our research is as follows. We classify the emotion of sentences based on the relation of all words constituting the sentence. These words are nouns, verbs, adjectives, and adverbs.

In the present phase, however, we focus on only adjectives and verbs in a sentence and decide the emotion of the sentence based on the emotion that those words have. Furthermore, the sentence patterns that we analyze are as follows:

S + C (S + V + C in English sentence)
S + V (Verbs are limited to intransitive verbs)

The emotions that are objects of the classification are four ones such as:

Joy, Sorrow, Anger, Surprise, and
Neutral emotion.

3.3. Speech Synthesis

The emotion of speech can vary by controlling the following three parameters:

- Pitch; basic frequency
- Power; amplitude of spectrum
- Tempo; utterance time

The relation between emotions and parameters to control them has been studied already, and the abstract is shown in Table 1. The characteristics that are relative to neutral emotion are described. It is not clear about anger, but they are estimated from a result of our study.

4. EMOTION OF WORDS AND SENTENCES

4.1. Emotion of words

We got the emotion that is impressed from each word by an experiment using fifteen subjects. Each subject classifies all words into five emotions mentioned above. 73 adjectives among 136 ones contained in IPAL were related to the four emotions except the neutral emotion. The emotional strength of each word is calculated by the following equation:

$$\frac{\text{the number of subjects who classified the word into the emotion}}{15}$$

Examples of words are shown in Table 1.

Table 1 Examples of the emotional strength of adjectives

Emotions / Words	Neutral	Joy	Sorrow	Anger	Surprise
thick	0.7	0.1	0.1	0.0	0.1
beautiful	0.1	0.9	0.0	0.0	0.0
arduous	0.1	0.0	0.9	0.0	0.0
annoying	0.0	0.0	0.0	1.0	0.0
rare	0.3	0.2	0.0	0.0	0.5

4.2. Emotion of sentences

The level of strength of the emotion in a sentence was gotten by the same experiment as in the previous section (4.1).

Sentence pattern S+C
The adjective in the part C controls emotion of a sentence in most cases.

Examples:
(1) この本はおもしろい。
(This book is interesting.) (Joy 1.0)
This+book+interesting (Joy 1.0)
(2) 面倒な仕事はつらい。
(Troublesome work is arduous.) (Sorrow 0.8)
Troublesome (Anger 0.7)+work+arduous (Sorrow 0.9)

Sentence pattern S+V
The emotion of a sentence is often determined by emotion of the verb, but there are many exceptions. The characteristics of the subject change the emotion of a sentence in many different ways.

Examples:
(1) 美しい婦人が嘆き悲しんでいる。
(A beautiful lady weeps.) (Sorrow 0.9)
beautiful (Joy 0.7)+lady+weep (Sorrow 0.9)
(2) 憎い敵が負ける。
(A hateful enemy loses.) (Joy 0.7)
hateful (Anger 0.9)+enemy+lose (Sorrow 0.8)

The relationship between subject and verb in sentence pattern S + V is shown in Table 2.

Table 2 Emotions of sentence pattern S + V

Subjects / Verbs	Neutral	Joy	Sorrow	Anger	Surprise

5. EVALUATION OF PROSODIC PARAMETERS

5.1. Experimental Method

An experiment to evaluate quantitatively the qualitative characteristics of the prosodic parameters, pitch, power and tempo, shown in Table 3, was conducted using speech synthesis software, Visual Speech Creator [4]. While listening to a Japanese sentence in the speech synthesis software, subjects adjusted the prosodic parameters interactively so that the sentence can express each emotion most. Subjects were asked to adjust prosodic parameters in the whole sentence uniformly, but they were allowed to adjust them partially when they were not satisfied. Ten subjects were engaged in this experiment.

5.2. Experimental Results

The average of all subjects in every prosodic parameter is

Table 3 Relationship between emotions and parameters

	Neutral Emotion	Joy	Sorrow	Anger	Surprise
Pitch	Normal	High	Slightly High	High	High
Power	Normal	Large	Small	Large	Normal
Tempo	Normal	Fast	Slow	Fast/Slow	Fast

Table 4 Results of experiment

	Neutral Emotion	Joy	Sorrow	Anger	Surprise
Pitch (Hz)	280	234 (0.8)	358 (1.2)	396 (1.4)	354 (1.2)
Power	50	33 (0.6)	80 (1.6)	80 (1.6)	65 (1.2)
Tempo (Sec)	3.765	4.314 (1.2)	3.381 (0.9)	3.021 (0.8)	3.381 (0.9)

shown in Table 4. The numerical value in parenthesis is magnification in comparison with the value of the parameter of natural emotion. Four subjects lifted pitch and power partially at the end of a sentence in order to express "Anger".

6. CONCLUSION

The values of parameters in the experiment shown in Table 4 are in almost the same correlation as the qualitative characteristics in Table 3. The synthesized speech expresses emotion well except that it is difficult to express "Anger".

7. FUTURE WORK

In this paper, the emotion of a sentence was decided based only on adjectives in a sentence, but we are attempting to classify the emotion of sentences based on the relations among all words, nouns, verbs, adjectives and adverbs, constituting the sentence. Furthermore, although the prosodic parameters were set uniformly in the whole sentence, it is necessary to consider that we control the parameters separately for every word and clause in order to get closer to human speech.

8. REFERENCES

[1] Charpentier, F. and Moulines, E., "Pitch Synchronous Wave form Processing Techniques for Text-to-Speech Synthesis using Diaphones", Proceedings of Euro speech '89, 1989.

[2] IPA Lexicon of the Japanese Language for computers, Information-technology Promotion Agency, http://www.ipa.go.jp/STC/NIHONGO/IPAL/ipal.html, 1990.

[3] ChaSen, Nara Institute of Science and Technology (NAIST), http://chasen.aist-nara.ac.jp/index.html.en, 1999.

[4] Visual Speech Creator, NTT-IT, http://www.ntt-it.co.jp/main.html.

Prosody based Attitude Recognition with Feature Selection and Its Application to Spoken Dialog System as Para-Linguistic Information

Shinya Fujie[1], *Daizo Yagi*[1], *Hideaki Kikuchi*[2], *Tetsunori Kobayashi*[1]

[1] School of Science and Engineering, Waseda University, Japan
[2] School of Human Sciences, Waseda University, Japan

Abstract

In this paper, prosody-based attitude recognition and its application to a spoken dialog system are proposed. Para-linguistic information plays a important role in the human communication. We aimed to recognize the user's attitude by prosody, and apply it to a spoken dialog system as para-linguistic information. In order to find important features to recognize the attitude from automatically extracted features, we applied some feature selection methods. Experimental results show the stepwise method, a combination of the forward selection method and the backward selection method, achieved the best recognition rate. Finally, the dialog system using the recognition results as para-linguistic information is shown.

1. Introduction

Humans exchange information in many ways. Particularly, in a spoken conversation, linguistic information carried by a speech is thought to be one of the most important information. However, we cannot communicate each other effectively enough without any sorts of information that helps the transmission of linguistic information. We call these sorts of information "para-linguistic information." In order to make a spoken dialog system talk with human more naturally, it is indispensable to introduce para-linguistic information. We focused on the speaker's attitude transmitted in the utterances.

Delleart et al[1] recognized four emotions, such as "happy", "sad", "anger", and "fear", by prosody. In this work, 17 pitch features divided into 5 groups are introduced. The only 5 features selected in order by the performance of the cross-validation experiment, achieved better performance than the original features. The majority voting of specialists method, where the original feature space is divided into the small sub spaces and the results of the sub spaces are combined, marked better as well. McGilloway et al[2] gave a benchmark of automatic emotion recognition using features by ASSESS system by Cowie. Lee et al[3] recognized negative and non-negative emotions using prosody. In this work a total of 10 features representing F0 and energy information are introduced. The features reduced the dimension (6 or 7) by PCA and feature selection(the same method with [1]) marked better than the original base features. Ang et al[4] used prosody for the detection of frustration and annoyance in natural human-computer dialog. They used the corpus of human-computer dialog developed under the DARPA Communicator Project. They introduced language model features as well as the prosodic features.

These works extracted the user's real emotion from utterances. The recognition results are important for estimating user's "real" emotional state. But the real emotion is not intuitively applicable for the dialog strategy, because people don't seem to estimate the partner's emotion for each utterance in the conversation. The important information for the dialog strategy is the partner's strategy and attitude, and they often appear as para-linguistic information in the conversation.

In this study, we recognize user's attitude using prosody as para-linguistic information. In order to find important features to recognize the attitude from automatically extracted features, we apply some feature selection methods. Finally, we show a spoken dialog system understanding para-linguistic information implemented on a humanoid robot.

2. Prosody-based Attitude Recognition

2.1. Data

We recorded the utterances of the users' responses to the system's suggestions. In order to collect a large number of utterances that include the positive or negative attitudes, the users' responses in the single turn (the system's suggestion and user's response) were recorded. The vari-

Table 1: The variety of the recorded utterances.

CATEGORY /RESTAURANT	Hamburger, Noodles, Packed Lunch, Curry, Refectory, McDonald's, Ajigen, Mumin, Hoka-ben Soba-no-mi
PHRASE	_____ ka, _____ ne _____, iinjanai, so-dane
ATTITUDE	positive, negative

ety of the recorded utterances are seen in Table 1. In the PHRASE, the underlined part represents the repetition of the CATEGORY or RESTAURANT in the system's suggestion. The words "ka" and "ne" are post-positional particles of Japanese, which do not define positive or negative by themselves. The phrases "iinjanai" and "so-dane" are usually received as positive, but they can be received as negative depending on the use. The utterances of 20 students in our laboratory, total 2000 utterances, were recorded.

For example, in the case that the CATEGORY is "Hamburger," the PHRASE is "____ ka" and the ATTITUDE is negative. First, the user hears the utterance "Hamburger nan te do kana", which means "What about a hamburger?". After that, the user utters "Hamburger ka" with negative attitude, and we record it.

2.2. Feature Extraction

The features are extracted from F0 pattern and phoneme alignment of the utterance. F0 pattern is extracted with *get_F0* module packaged in ESPS/Waves+, also provided as a component of an open source sound manipulation library *Snack*[5]. Phoneme alignment is obtained from the speech recognition result of a decoder, *Skood*, developed in our group.

13 features, F0 average, F0 median, F0 maximum, F0 minimum, F0 minimum of last mora, F0 range, first mora duration, last mora duration, mean of mora durations, F0 gradient of first mora, F0 gradient of last mora, F0 gradient maximum and F0 gradient minimum are chosen as basic features.

2.3. Feature Selection

In order to find important features to recognize the attitude, we applied some feature selection methods to our basic features. These methods select or eliminate the feature one by one according to the result of an experiment with previously selected feature set.

In each experiment, Gaussian Mixture Model(GMM) is learned for each category(positive/negative) and the cross validation test (750 utterances for learning set, 250 utterances for test set) is done by the Gaussian classifier with these GMMs.

Forward Selection Method(FS) is a method of incremental feature selection. It divides the basic features into one dimension features, and experiments are done for each feature. The feature that marks the most highest recognition rate, is selected as the first dimension of the feature vector. It requires experiments with the feature vectors, where the selected dimensions are fixed and the last dimension is one of the remaining features to determine the next dimension feature.

Backward Elimination Method(BE) does an experiment with basic feature vector at first. Then, the next experiments are done with feature sets, that each one dimension is eliminated from basic feature vector. The feature set that marks most highest recognition rate, is selected as the $N-1$ dimension feature vector, where N is the dimension of basic feature set. In the same way, features are eliminated in order.

Stepwise Method(S) is a combination method of FS and BE. At first, FS is done until the recognition rate is saturated. Then, BE is done with the selected feature vector by then until the rate is saturated. FS and BE are done by turns in this way.

3. Experiments and Results

3.1. Feature Selection

Table 2 4 show the results of forward selection, backward elimination and stepwise method respectively. The boldface features represent the features chosen for recognition with each method.

Table 2: Result of forward selection.

+/−	feature	recog. rate	mix.
+	**F0 grad. min.**	76.1%	8
+	**last mora dur.**	82.2%	64
+	**last mora F0 grad.**	84.2%	64
+	**F0 max.**	87.5%	64
+	F0 min.	87.0%	16
+	mora dur. mean	87.0%	16
+	first mora F0 grad.	87.4%	16
+	F0 grad. max.	87.4%	64
+	F0 mean	87.0%	16
+	F0 range	86.4%	8
+	last mora F0 min.	85.1%	8
+	F0 median	85.4%	32
+	first mora dur.	83.9%	32

Table 3: Result of backward elimination.

+/−	feature	recog. rate	mix.
		83.9%	32
−	F0 min.	85.6%	16
−	F0 median	86.3%	64
−	last mora F0 grad.	86.2%	64
−	F0 max.	86.8%	16
−	**last mora F0 min.**	86.2%	16
−	**first mora dur.**	86.0%	16
−	**first mora F0 grad.**	86.4%	64
−	**F0 grad. max.**	86.3%	16
−	**F0 grad. min.**	85.3%	8
−	**mora dur. mean**	84.2%	8
−	**F0 mean**	81.2%	32
−	**F0 range**	75.5%	32
−	**last mora dur.**		

Table 4: Result of stepwise.

+/−	feature	recog. rate	mix.
+	**F0 grad. min.**	76.1%	8
+	**last mora dur.**	82.2%	64
+	**last mora F0 grad.**	84.2%	64
+	**F0 max.**	87.5%	64
+	F0 min.	87.0%	16
+	**mora dur. mean**	87.0%	16
+	**first mora F0 grad.**	87.4%	16
−	F0 min.	87.9%	16
+	**F0 grad. min.**	88.3%	64
+	F0 median	87.3%	8
+	F0 grad. max.	86.8%	32
+	first mora dur.	86.2%	32
+	F0 range	85.5%	8
+	F0 mean	83.6%	16

3.2. Comparison with Human

In order to compare the recognition ability of our model with that of a human, we calculate a measure of agreement between results of humans' annotations and the experiment results.

Among the recorded utterances, we picked up 20 utterances from each category, positive and negative. 5 subjects (A–E) listened to the utterances in random sequence and determined the attitude, positive or negative, for each.

We introduce Cohen's κ [6] as a measure of agreement. It is often used as an evaluation measure of a dialog corpus annotation. Since Cohen's κ is a value calculated from 2 annotations, we calculated it for each combination of the annotations including the experiment results. The minimum, maximum and average of the rates with others was calculated for each person. The result is seen in Table 5. All three methods are the same value in the average. But the stepwise method is the best in order to obtain the highest minimum value.

Table 5: Calculation results of Cohen's κ

	min	max	average
FS	**0.48**	**0.74**	**0.64**
BE	**0.40**	**0.84**	**0.64**
S	**0.55**	**0.70**	**0.64**
A	0.52	0.85	0.71
B	0.65	0.85	0.76
C	0.52	0.79	0.65
D	0.65	0.80	0.74
E	0.70	0.79	0.75

FS: Forward Selection, BE: Backword Elimination, S:Stepwise, A–E: Subjects

4. Spoken Dialog System

4.1. Dialog Strategy

We aimed to adopt the results of the prosody based attitude recognition to a spoken dialog system as para-linguistic information. In our previous study, head gesture recognition system has been implemented [7]. This system is able to recognize several kinds of head gestures, such as "nod", "tilt", and "shake," using optical flow as the feature and HMM as the probabilistic model. As well as prosodic information, these kinds of head gestures express the utterer's attitude well. We introduce results of head gesture recognition to a spoken dialog system as para-linguistic information.

So we must combine the both results to estimate the user's correct attitude. Particularly, when the results are contrary with each other, one is positive while the other is negative, it's serious problem to decide which result is plausible. In such case, it is recognized as "in thought", which represents that the user cannot decide clearly.

The dialog strategy, when the system receives the user's response to its suggestion about a category, is shown in Figure 1. When it is positive, it can provide detail about the category. In our task, the system suggests the restaurant in the category. When it is negative, it can provide other candidates.

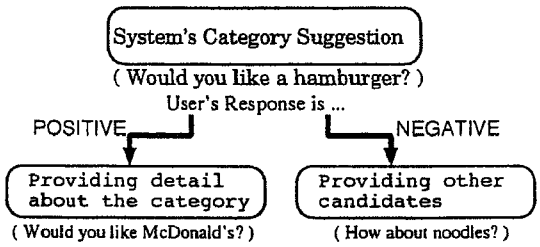


Figure 1: Dialog strategy

4.2. Proto-type spoken dialog system

We implemented a proto-type spoken dialog system with para-linguistic information on a humanoid robot ROBISUKE (Figure 2). We have developed humanoid robots that interact with humans with their perceptual and expressive abilities; such as speech recognition, individual identification using face images, and facial expressions using its eyes, eyebrows and mouth. ROBISUKE is the most recently developed humanoid robot in our group.

In order to recognize user's attitude in real-time, we adopted the F0 extraction to the sound input module, and phoneme alignment to the speech recognition module. The attitude recognizer can recognize user's attitude as his utterance finishes according to the inputs from these modules(Figure 3).

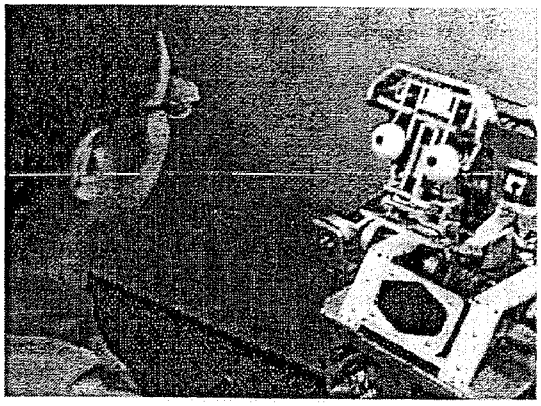


Figure 2: Humanoid robot ROBISUKE in conversation.

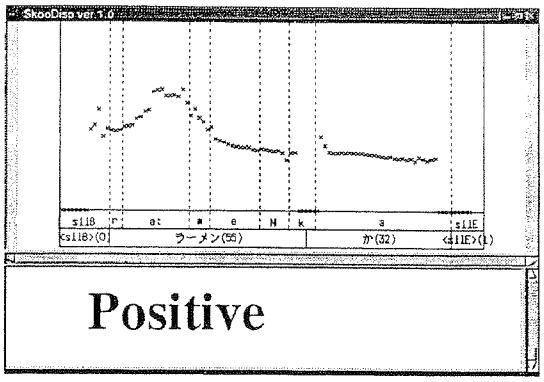


Figure 3: Prosody-based attitude recognizer.

4.3. Example

An example dialog is seen in Figure 4.

The system suggests another idea when the user's response is negative to the previous suggestion. It suggests the restaurant in the suggested category when the user's response is positive to the previous suggestion.

In Figure 4, the responses "Curry ka" and "Bento ne" which cannot be decided positive or negative with the only linguistic information, are appropriately interpreted using para-linguistic information.

5. Conclusion

In this paper, we applied the feature selection method to the prosody-based attitude recognition. Experimental results show that the stepwise method is the best from the point of view of the recognition rate. However, even if a feature combination marks high recognition rate, not all such combination marks high agreement value. We show a spoken dialog system with para-linguistic information, and these sorts of information influence the progress of a dialog effectively.

Evaluation of the spoken dialog system, particularly on the dialog quality, is the most important issue.

U: Ohiru gohan nandakedo, doko ka ii tokoro naikana
(I'd like going to lunch. Do you have some idea?)
R: Curry nante do
(What about curry?)
U: Curry ka **(negative)**
(Curry)
R: Sorejaa, hamburger nante doukana
(Well then, what about hamburger)
U: Ah, hamburger ne **(strongly positive)**
(Oh, hamburger)
R: Hamburger nara chikaku ni McDonald ga aruyo
(For hamburger, there's McDonald's nearby)

Figure 4: An example dialog
User's utterances and ROBISUKE's utterances are preceded by U: and R: respectively.

6. References

[1] Dellaert, F., Polzin, T. and Waibel, A., "Recognizing Emotion in Speech," in Proceedings of ICSLP '96, 1996.

[2] McGilloway, S., Cowie, R., Cowie, E. D., Gielen, S., Westerdijk, M. and Stroeve, S., "Approaching Automatic Recognition of Emotion From Voice: A Rough Benchmark," in Proceedings of ISCA workshop on Speech and Emotion, 2000.

[3] Lee, C. M., Narayan, S. and Pieraccini, R., "Recognition of Negative Emotions from the Speech Signal," in Proceedings of IEEE ASRU2001, 2001.

[4] Ang, J., Dhillon, R., Krupski, A., Shriberg, E. and Stolcke, A., "Prosody-based Automatic Detection Of Annoyance And Frustration in Human-Computer Dialog," in Proceedings of ICSLP2002, pp.2037 2040, 2002.

[5] http://www.speech.kth.se/snack/

[6] Cohen, J., "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37 46, 1960.

[7] Ejiri, Y., Nakajima, K., Fujie, S., Kobayashi, T., "Head Gesture Recognition for the Moving Conversation Robot," IEICE Tech. Rep., Vol. 103, No.453, PRMU2003 153, pp.13 18, Nov., 2003. (in Japanese)

[8] Fujie, S., Ejiri, Y., Matsusaka, Y., Kikuchi, H. and Kobayashi, T., "Recognition of Para-Linguistic Information and Its Application to Spoken Dialogue System," in Proceedings of IEEE ASRU2003, pp.231 236, 2003.

A Method to Classify Emotion in Text, Based on the Distribution of Emotional Words

Tosimitu Tukamoto, Futoshi Sugimoto and Masahide Yoneyama

Department of Information and Computer Sciences
Toyo University
2100 Kujirai Kawagoe, Saitama 335-8585 Japan
Email: tukamoto@r2.dion.ne.jp, f_sugi@eng.toyo.ac.jp, yoneyama@eng.toyo.ac.jp

Abstract

Representing emotional expressions on text-to-speech synthesis is a fascinating subject. Our aim is to add emotional expressions to synthetic speech. This system recognizes emotion based on the distribution of the words in the text and synthesizes emotional speech.

Keywords

Text-to-speech, Emotional expressions and Synthetic speech

INTRODUCTION

In recent years, it has become possible to have a computer read naturally expository texts with synthetic speech because of the technical progress of speech synthesis (Charpentier 1989). There are, however, still problems to be resolved in Text-To-Speech that treats texts which include a lot of emotion and wealthily expresses emotion in synthetic speech. These problems consist of the prosody control to express emotion and the classification of emotion in a text. In this paper, we propose a method to classify emotion in a text. In this method, we partition a text into discourse units in which it is easy to classify emotion by using a method used in the topic detection, and decide the emotion of each discourse unit based on the distribution of words having emotion.

OUTLINE OF TEXT TO SPEECH WITH EMOTION

The outline of the system in which we make a computer read text with an emotional voice, our ultimate goal, is shown in Figure 1. At first, we classify the emotion of an input text in perspective. In this part, the text is partitioned into discourse units and the emotion of each unit is classified based on the emotional words included in the unit. Then, the values of prosodic parameters, pitch, power and tempo, for each discourse unit are determined according to the emotion of the unit in order to give moderate emotion to the synthetic speech. Secondly, the local part in which the emotion is raised is found in every sentence or clause in the discourse unit, and the different values of prosodic parameters that can express stronger emotion are assigned to the part. As a result, the synthetic speech has nicely varied emotion. In this paper, we describe the part of the classification of emotion in perspective.

CLASSIFICATION OF EMOTION IN PERSPECTIVE

We partition a text into discourse units by using a method that is often used in the segmentation of text, the

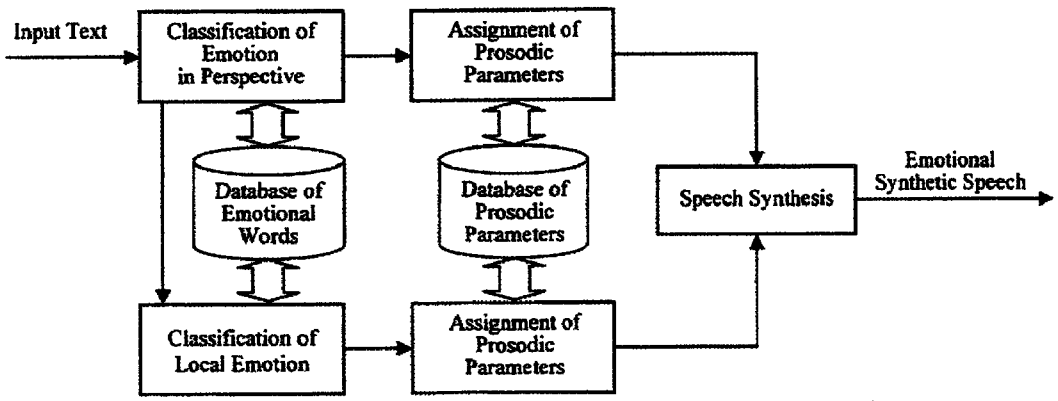


Figure 1. Outline of Speech Synthesis with Emotion

Hearst method (Hearst 1997), and determine the emotion of each discourse unit based on the distribution of emotional words included in the unit.

Optimum Parameters of Hearst Method in Emotional Text

Since the Hearst method is often used in the topic detection of expository texts but rarely in emotional texts, we have to find out the optimum values of the parameters of the Hearst method, step size and block size, in order to apply the method to our system. We determined the optimum solution through the following procedures. At first, 15 subjects were asked, respectively, to divide a text into discourse units, and the borders where more than four subjects agreed were adopted as the borders of subjects. Secondly, the values of parameters were heuristically determined so that the borders obtained by the Hearst method corresponded with the borders of the subjects most. The values are the optimum.

Emotional Words

In this paper, we limit the emotions that are objects of classification and words used. The emotions are joy, sorrow, anger, surprise, terror and aversion; these strongly characterize the speech and neutral emotion that is located at the center of these emotions. In addition, we use only words contained in the IPA Lexicon of the Japanese Language for computers (IPAL 1990).

Japanese sentences are organized into adjective sentences (S+V+C in English) and verb sentences (S+V and S+V+O). Since the emotion of adjective sentences is influenced by the emotion of an adjective or a noun assigned in the complement as in the following examples, the adjective and the noun become the emotional words (the emotional factors).

(1) この本はおもしろい。(Joy 1.0) = This book is interesting (Joy 1.0).

(2) 面倒な仕事はつらい。(Sorrow 0.8) = Troublesome (Anger 0.7) work is arduous (Sorrow 0.9).

Since the emotion of verb sentences is determined by a combination of a noun of the subject and a verb of the predicate as in the following examples, this pair becomes the emotional factor.

(1) 美しい婦人が嘆き悲しんでいる。(Sorrow 0.9) = A beautiful (Joy 0.7) lady weeps (Sorrow 0.9).

(2) 憎い敵が負ける。(Joy 0.7) = A hateful (Anger 0.9) enemy loses (Sorrow 0.8).

ASSIGNMENT OF PROSODIC PARAMETERS

An experiment to evaluate quantitatively the qualitative characteristics of the prosodic parameters, pitch, power and tempo was conducted using speech synthesis software. While listening to a Japanese sentence using the speech synthesis software, subjects adjusted the prosodic parameters interactively so that the sentence can express each emotion most. Subjects were asked to adjust prosodic parameters throughout the entire sentence uniformly, but they were allowed to adjust them partially when they were not satisfied. Ten subjects were engaged in this experiment. The average of all subjects in every prosodic parameter was calculated. We assign these values to the prosodic parameters for controlling the emotion of the discourse units.

CONCLUSION

We tried to segment a text into different discourse units by using the Hearst method in order to classify the emotion of a text in perspective. Here, we assume that the same emotion is distributed in each unit, but it is necessary to inspect it. Furthermore, it is comparatively easy to control the prosodic parameters equally in every scene, but it is necessary to study a method to control the parameters for a limited part in the future.

REFERENCES

Charpentier, F. and Moulines, E. (1989) Pitch Synchronous Wave form Processing Techniques for Text-to-Speech Synthesis using Diaphones, *Proceedings of Euro speech '89*.

Hearst, M. (1997) Text Tiling: Segmenting Text into Multi-paragraph Subtopic Passages, *Computational Linguistics, Vol. 23, No.1*.

Information-technology Promotion Agency (1990) IPA Lexicon of the Japanese Language for computers, URL http://www.ipa. go.jp/STC/NIHONGO/IPAL/ipal.html

2nd International Conference on Autonomous Robots and Agents
December 13-15, 2004 Palmerston North, New Zealand

A Conversation Robot with Back-channel Feedback Function based on Linguistic and Nonlinguistic Information

Shinya Fujie, Kenta Fukushima, Tetsunori Kobayashi
School of Science and Engineering,
Waseda University, Tokyo, Japan

Abstract

A conversation robot which can generate a back-channel feedback appropriately to the user is developed. The contents of these feedbacks depend on the contents of the dialogue partner's utterance, which are provided by the speech recognition. In order to determine the content of the feedback earlier than the end of the utterance, we used finite state transducer based speech recognizer which outputs the content of the feedback. And we used prosody information, especially the fundamental frequency(F0) and the power of the utterance, to extract the proper timing of the feedback. We implemented these modules and applied them to the spoken dialogue system on the humanoid robot ROBISUKE. Experimental results show the effectiveness of our methods.

Keywords : Conversation robot, back-channel feedback, spoken dialogue system, prosody, FST

1 Introduction

1.1 Back-channel Feedback in Conversation

In human conversation, we exchange the turn of talking one another. The natural exchange of turn realizes the natural conversation. Most of conventional spoken dialogue systems assume that the end point of the speech recognition of user's utterance is the transfer of the turn from the user to the system. And also they assume the end point of the system's speech synthesis is that from the system to the user. Those systems have to wait until the end of the user's utterance, once they finish their utterance and transfer the turn to the user. However, in human conversation, while the speaker speaks something, the listener feeds back his/her state in some ways. For example, the listener nods to show he/she listens to the speaker carefully, and the listener repeats the speaker's words to feed back how he/she understands what the speaker says. By receiving these feedbacks, the speaker feels easy to talk, because he/she is able to knows the listener listens to the utterance and how the listener understands. These feedbacks are generated by the listener unconsciously. It is called "back-channel feedback." If a spoken dialogue system can generate the back-channel feedback appropriately, the conversation becomes more effective and natural.

1.2 Studies about Back-channel Feedback Generation

To generate the back-channel feedback appropriately, the dialogue system must decide the content and the timing of the feedback according to the user's utterance. There are several studies about the back-channel feedback generation. They can be roughly classified into ones that use linguistic information and ones that use non-linguistic information.

Nakano et al. developed the spoken dialogue system which can do natural turn-taking by the incremental understanding using the context[1]. This system detects the speech recognition result earlier than the end of the utterance, and executes the language and semantic processing incrementally. The system decides what should be asked to the user and it synthesizes the content as the feedback. This kind of system can generate the feedback with the appropriate content, but it is afraid to make the user uncomfortable because it doesn't control the timing of the feedback.

Ward et al. developed the system that generates the back-channel feedback based on the rule that a feedback is generated after the utterance which contains the low pitch for a certain period of time[2]. Okato et al. detected the timing of the feedback using the prosodic templates obtained from the spoken dialogue corpus, and developed the feedback generation system with the same method[3]. Koiso et al. proposed the model that predicts the end/succession of the utterance by the analysis of correlation among the word, pitch pattern and energy pattern in the ending part of the utterances[4]. The system developed by Takeuchi et al. controlled the timing of the feedback generation, using the prosodic information and the part of speech of the spoken word[5]. This kind of system can decide the timing of the feedback generation appropriately. However, it cannot decide the contents of the feedback.

1.3 Proposed Method

In this paper, we aim at the system that can generate the back-channel feedback with the appropriate content at

the appropriate timing. As described above, the content of the feedback depends on linguistic information, i.e. the content of the utterance, and the timing depends on the style of the utterance.

In order to respond with the appropriate content in the middle of the user's utterance, the system must detect the results of the speech recognition before the end of the utterance. It seems to be difficult to realize the early detection using conventional speech recognizer, because of the complexity of the data structure and the implementation. In this study, we use the Finite State Transducer(FST) based speech recognizer, which becomes popular recently, to decide the content of the feedback. FST itself has an early detection function, and it can generates any symbols as output by composing several transducers.

The timing detection is realized by processing prosodic information for every frame. In this study, as prosodic information, we use the fundamental frequency (F0) and the logarithmic power of the user's utterance.

Using these two outputs, we implement the spoken dialogue system that can generate the back-channel feedback with appropriate content at the appropriate timing, on the humanoid robot.

2 Early Detection of Feedback Contents

2.1 Speech Recognizer Using FST

FST is an automaton with output symbols. A transducer with expected input/output symbols can be obtained by composing several transducers. The early decision of output symbols corresponding to the given input symbols can be realized by the optimizing the transducers with several operations such as minimization, determinization, and so forth.

In order to recognize user's speech with a large vocabulary continuous speech recognizer(decoder), we have to give it three databases, so called acoustic model, word dictionary and language model. Acoustic model is a probabilistic model that gives an acoustic feature output probability of each phoneme. Word dictionary is a list of phoneme sequences of all words. Language model is a probabilistic model that represents which word sequence is more plausible in a target language. Common decoders use Hidden Markov Model(HMM) for acoustic model and n-gram model (bi-gram or tri-gram) for language model.

HMM, word dictionary and n-gram language model can be represented by a network, and the recognition (decoding) is done by transition on the network. A single FST with early detection function can be obtained by constructing FSTs of these models individually, composing them into single FST and optimizing it. This kind of network is the most optimized network

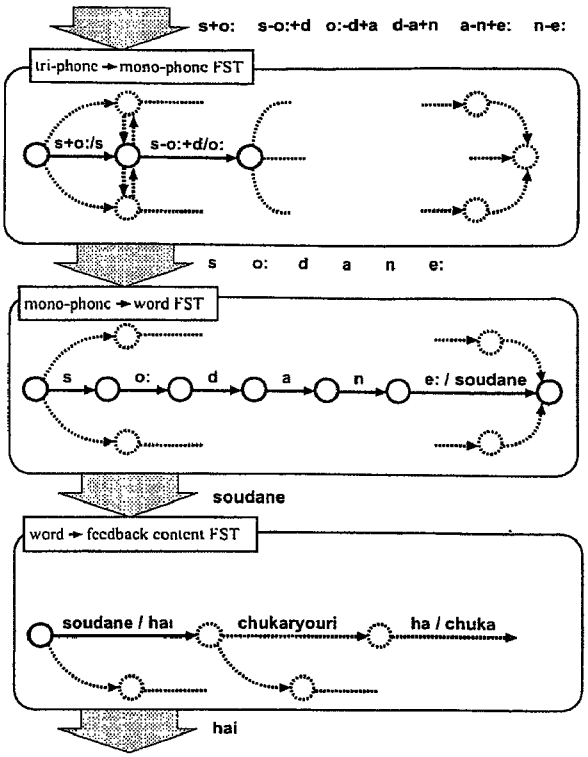


Figure 1: FST with the feedback contents as output symbols.

that can be obtained before the system works. And because the network is the only single one, we can use simple search algorithm, and the decoder implementation is very simple. Thus, recently, many researches focus on the application of FST to the speech recognition[6][7].

2.2 Early Detection of Back-channel Feedback Contents

Generally, the final output of speech recognition is a word sequence. In this study, the final output is the content of the back-channel feedback corresponding to the utterance of the user. We aim to detect it earlier than the end of the utterance using the early detection function of FST.

As shown in figure 1, a single network which accepts the tri-phone as input symbol and outputs the recognized words and the content of the back-channel feedback, is obtained by composing 3 networks. The first network accepts tri-phone as input and outputs mono-phone. The second one accepts mono-phone as input and outputs the recognized words. The third one accepts a word as input and outputs the given word and the contents of the back-channel feedback. By decoding with this network, the decoder outputs the contents of the back-channel feedback on the point where it should generate the feedback. And also, we give the special output symbol `delete` on the point where the feedback is not appropriate any more. Thus, the

feedback that must not be generated can be eliminated by decoding with this network.

We use tools provided by AT&T[8][9] in order to compose and optimize the FST.

3 Back-channel Feedback Timing Detection Using Prosodic Information

3.1 Back-channel Feedback and Prosody

The output of the decoder using FST, described in the previous section, is the appropriate content of the feedback when it should be generated. According to the early detection function of the FST decoder, the output can be obtained before the timing which the system should actually generate the feedback. But the decoder cannot determine the timing.

The decision whether the system should generate the feedback or not depends on the style of the user's utterance rather than the content of it. For example, in the case that the user's utterance is "karakute oishii mono ga tabetai na(I would like have something hot and tasty)." When the user says the part "karakute" with stretching the last phoneme 'e', the system should generate the feedback like "hai(yes)" or the repetition of the word "karai(hot)." On the other hand, when the user says "karakute oishii" at a breath, the feedback is not only unnatural in the conversation but also uncomfortable for the user.

In this section, we aim to detect the appropriate timing that the system should generate the feedback using the F0 and energy of the user's utterance.

3.2 Analysis of Dialogue Data

At first, in order to collect the partner's utterances right before the back-channel feedback, we recorded the simulated conversation between two people. 9 conversations, free topic, 5 minutes long each, were recorded with video camera. The all back-channel feedback parts over the recorded movie are annotated by hand with annotation tool Anvil[10].

We extracted and observed the partner's utterance right before the back-channel feedback. An example of the utterance "ah ichinichi me de (ah, on the first day)" is shown in figure 2. The graph shows the extracted F0 and logarithm power of the utterance. According to the observation of many samples, characteristic appears in the part around 100msec to 500msec before the feedback.

3.3 Feature and Model

In order to detect the proper timing of the feedback, we introduce the pattern recognition technique. The

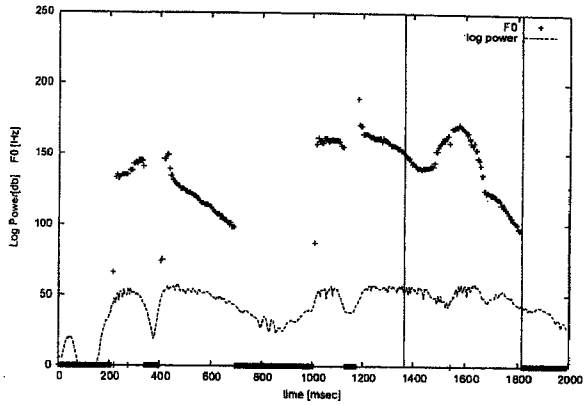


Figure 2: An example of F0 and power extraction of the utterance.

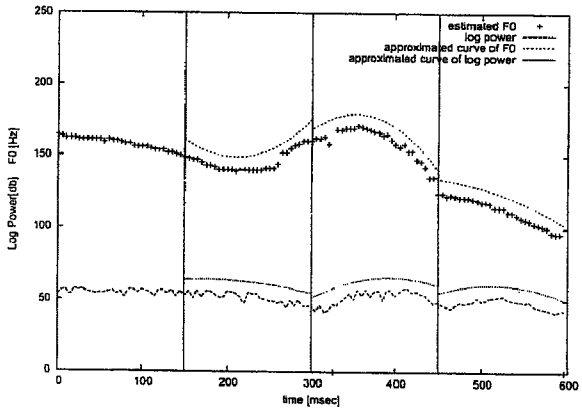


Figure 3: An example of approximated curve of extracted F0 and power.

feature vector consists of the coefficients of the approximated curve of the F0 and logarithm power over the divided 3 regions of the past 450msec region, and the average of the errors between curve and the extracted values. The approximated curve is calculated with least-square method.

$$y'(t) = at^2 + bt + c \tag{1}$$

We adopt the coefficients a and b as the feature. The average of the errors is calculated as follows.

$$\frac{1}{N}\sum_i |y_i \quad y'(t_i)| \tag{2}$$

Where, y_i represents extracted F0 or logarithm power, and N is the number of the extracted samples. In this study, the sampling rate is 16kHz and the frame shift is 80 samples, so the N is 30 for each region (the duration of one region is 150msec). Three kinds of features about F0 and power are calculated for each region, so the total dimension of a feature vector is 18. Figure 3 shows an example of approximated curve of F0 and power, calculated over the region shown in figure 2.

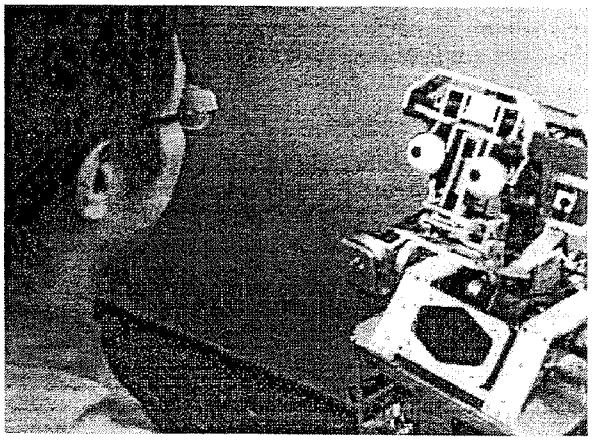


Figure 4: Conversation robot ROBISUKE

The feature vectors were calculated with 119 utterance samples which were obtained from the recorded database described in the previous section. And a Gaussian distribution was learned with these feature vectors. In the recognition stage, the feature vector is calculated, and the output likelihood of it from the learned distribution is evaluated for every frame. When the likelihood exceeds a threshold, it is decided to be the timing for the system to generate the back-channel feedback. The threshold is adjusted by hand before the system works.

4 Back-channel Feedback Function on a Conversation Robot

4.1 System Overview

We applied the proposed FST decoder and prosody process to the spoken dialogue system which we have developed. The spoken dialogue system has been implemented on the humanoid robot ROBISUKE[11] shown in figure 4. The system architecture is shown in figure 5.

The speech of the user's utterance is input into the prosody process module which calculates the feature vector and evaluates the back-channel feedback generation likelihood which we proposed in the previous section, as well as input into the MFCC extraction module which calculates the MFCC which is needed for the speech recognition. In the prosody process module, the sampling rate of the speech is 16kHz, the size of one frame is 1024 samples, and the frame shift is 80 samples(5msec). For the F0 extraction, we use the method proposed by Goto et al[12] which using instantaneous frequency and a comb filter. By more popular method proposed by Talkin[13] (the implementation provided in *ESPS/waves+* and *wavesurfer*) we are able to obtain more precise F0 extraction results. However, Talkin's method use the Dynamic Programming to generate precise results, so for our system, which needs the real-time response,

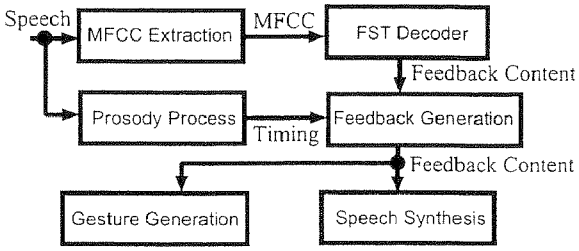


Figure 5: The system architecture

Goto's method is better. In the prosody process module, furthermore, the feature vector described in section 3.3 is calculated for each frame, and the likelihood of the timing for the back-channel feedback is evaluated. As soon as the likelihood exceeds the threshold, it sends a signal to back-channel feedback generation module.

FST decoder receives the MFCC feature and decodes with it. Because a decoding makes progress by spreading the hypotheses, the final output cannot be obtained until the end of the utterance, even if the FST has the early detection function. In our system, therefore, when the most plausible output doesn't change for several frames, it is decided to be the early detected output temporally. If a different output is more plausible than the previous output afterward, it outputs the special symbol `delete` which means the cancel of the previous one and outputs the newly detected one.

As the back-channel feedback generation module receives the content of the feedback from the FST decoder, it stores the content into the buffer. Afterward, as it receives a signal from the prosody process module, it sends the content in the buffer to the speech synthesis module and the gesture generation module.

The speech synthesis module synthesizes the received content. The gesture generation module makes ROBISUKE nod for a short time. This kind of head gesture often occurs along with the back-channel feedback.

5 Experiment

5.1 Target

In order to evaluate the developed system, we perform subjective evaluations. The target of this experiment is to confirm the effects of introducing the FST decoder and the prosody process module.

One of the expected effects of introducing the FST decoder is to enable the system to change the content of the feedback by what the user says. Moreover, because the decoder doesn't generate the output until the user says something for which the system should generate

Table 1: The result of pettern I

A is better	A is slightly better	even	B is slightly better	B is better
2	6	3	8	1

Table 2: The result of pettern II

C is better	C is slightly better	even	D is slightly better	D is better
3	4	0	15	10

a feedback, unnecessary feedbacks are expected to be reduced more than the case of using only the prosody process module.

The prosody process module is expected to enable the system to generate a feedback on the more appropriate timing than the case of using only the FST decoder.

5.2 Experimental Setup

Subjects talk with a pairs of the differently tuned dialogue systems, and compare these systems. The scenario of the conversation is previously prepared. And the evaluation is done by answering the question, "which system do you prefer?" with 5-point scale. We prepared 2 patterns (I and II) of pairs, a subject evaluated twice a pattern. There were 10 subjects for pattern I, and 16 subjects for pattern II.

Pattern I consists of system A which generates the feedback using only the prosody process module, and system B which generates the feedback using both the prosody process module and FST decoder. When the prosody process module sends a signal, system A generates the feedback regardless of the content of the user's utterance. In this experiment, we can see the effect of introducing the FST decoder.

Pattern II consists of system C which generates the feedback using only the FST decoder, and system D which generates using both modules like system B. In this experiment, we can see the effect of the prosody process module.

5.3 Results

The results of two experiments are shown in table 1 and 2. As we can see in table 1, there is few difference in pattern I. On the other hand, in table 2, there is clear difference where subjects prefer the system D which generates depending on both modules, in pattern II.

6 Discussion

The importance of the timing control based on the prosodic information is confirmed from the result of pattern II. On the other hand, the result of pattern I is separated to both sides, so the effects of using linguistic information cannot be seen clearly.

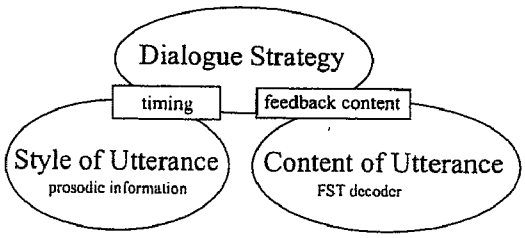


Figure 6: 3 factors to decide the back-channel generation.

Here, we discuss the factors that determine the timing and the content of the back-channel feedback.

The decision whether the back-channel feedback should be generated or not and the content of it seem to depend on the content and the style of the partner's utterance. In actual human conversation, however, they depend on the dialogue strategy of the participant that will generate the feedback.

For example, when the partner says "oishiii mono ga tabetai na (I would like to have something tasty)," the feedback plays a role quite differently depending on its content, "hai(yes)" or "oishii(tasty)." If it is "hai(yes)," it just represents the listener listens to the partner's utterance carefully, and it encourages the partner to continue the utterance. If it is "oishii," it represents the listener understands the partner's word "oishii" correctly in addition to listening carefully, but it doesn't encourage the partner to continue the utterance as strongly as "hai." And basically the feedback "hai" encourages the partner to continue the utterance, but too many feedbacks (even if its content is the same "hai") forces the partner to finish the utterance.

As shown in figure 6, the content and the timing of the back-channel feedback depend on not only the style and the content of the partner's utterance but also the dialogue strategy of the system itself. Therefore, in order to construct the spoken dialogue system that can generate appropriate back-channel feedbacks, we must consider the dialogue strategy of the system too.

7 Conclusion

In this paper, we proposed and implemented the FST decoder that generates the appropriate content of the back-channel feedback, and the method that detects the appropriate timing of the feedback using prosodic information. And we implemented the spoken dialogue system that can generate the back-channel feedback using proposed methods on the humanoid robot ROBISUKE.

Experimental results show the effectiveness of the FST decoder and the prosody process module. It shows that the timing is more important.

Future work includes the considering of the dialogue strategy of the system in order to make the content of the feedback more effective and more natural.

References

[1] M. Nakano, K. Dohsaka, N. Miyazaki, J. Hirasawa, M. Tamoto, M. Kawamori, A. Sugiyama, and T. Kawabata. Handling rich turn-taking in spoken dialogue systems. In *Proceedings of Eurospeech '99*, pages 1167–1170, 1999.

[2] N. Ward. Prosodic features which cue back-channel responses in English and Japanese. *Pragmatics*, 32:1177–1207, 2000.

[3] Y. Okato, K. Kato, M. Yamamoto, and S. Itahashi. Giving 'aizuchi' using prosodic information. *ISPJ Journal*, 40(3):469–478, 1998. (in Japanese).

[4] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogues. *Language and Speech*, 41(3-4):294–317, 1998.

[5] M. Takeuchi, N. Kitaoka, and S. Nakagawa. Timing detection for realtime dialog systems using prosodic and linguistic information. In *Proceedings of the International Conference Speech Prosody (SP2004)*, pages 529–532, 2004.

[6] Takaaki Hori, Chiori Hori, and Yasuhiro Minami. Speech summarization using weighted finite-state transducers. In *Proceedings of Eurospeech 2003*, pages 2817–2820, 2003.

[7] S. Kanthak, H. Ney, M. Riley, and M. Mohri. A comparison of two LVR search optimization techniques. In *Proceedings of ICSLP 2002*, pages 1309–1312, 2000.

[8] M. Mohri, F. Pereira, and M. Riley. Weighted finite-state transducers in speech recognition. In *Proceedings of ISCA ITRW, ASR 2000*, pages 97–106, 2000.

[9] AT&T Finite State Machine Library. Home page. http://www.research.att.com/sw/tools/fsm/.

[10] M. Mipp. Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of Eurospeech 2001*, pages 1367–1370, 2001.

[11] Shinya Fujie, Yasushi Ejiri, Hideaki Kikuchi, and Tetsunori Kobayashi. Recognition of paralinguistic information and its application to spoken dialogue system. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2003)*, pages 231–236, December 2003.

[12] Masataka Goto, Katunobu Itou, and Satoru Hayamizu. A real-time filled pause detection system for spontaneous speech recognition. In *Proceedings of Eurospeech '99*, pages 227–230, September 1999.

[13] D. Talkin. *Speech Coding and Synthesis*, chapter 14, pages 495–510. Elsevier Science, 1995.

Proceedings of the Ninth IASTED International Conference
INTERNET AND MULTIMEDIA SYSTEMS AND APPLICATIONS
February 21-23, 2005, Grindelwald, Switzerland

CONSTRUCTION OF THE WALKING MOTION MODEL FOR ANIMATED CHARACTERS' MOTION

Takehiro Morino, Daiki Osafune, Takayuki Itano, Katsuhiko Shirai
Department of Computer Sciences
Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555
Tokyo, Japan
email: takehiro-morino@suou.waseda.jp
Makoto Murakami
Department of Information and Computer Sciences
Toyo University
2100 Kujirai, Kawagoe-shi, Saitama 350-8585
Saitama, Japan

ABSTRACT
To represent and control the motion of animated characters for computer animations easily, it is necessary to construct a model using which we can control the motions with an intuitive and easy-to-use interface. To construct the model, we analyze the human motion data captured with a motion capture system. Initially, we clarify that the human walking motion is composed of a few hierarchical sub-motions that are independent of each other. Then, we extract some motion control parameters from the captured human walking motions with physical alteration. Finally, we demonstrate that it is possible to control the motions appropriately.

KEY WORDS
Computer Animation, Human-Computer Interfaces, Human Walking Motion, Motion Generation.

1 Introduction

In recent years, the processing power of computers has increased, and we have been able to create three-dimensional computer animations on our personal computers. In most computer animations, animated characters are present and they play an important role. Therefore, it is necessary to generate and control their motions appropriately. However, motion capture systems, which are used to capture human motions and to apply this information to animated characters, are not available to all because of both the equipment size and the expense involved. We can also use keyframe animation which is a technique by calculating the continuous presentation from some keyframes at regular intervals, and manually created animation by assigning a coordinate to every joint in all frames. However, in these methods, sophisticated skills are essential to appropriately generate the motions of animated characters.

After all, the method should generate and control the motions with an intuitive and easy-to-use interface. One of the ideal methods is to generate the motions through the interaction between a user and the character generated in a system.

Several researchers have proposed models based on the motion database [1], pattern generator [2], space-time constraints [3], labanotation [4], and musculoskeleton [5, 6]. They do not regard every joint motion as a motion that correlates closely with the others. We analyze the human motion captured with an optical motion capture system to construct a model in which all the joint motions are classified hierarchically and for which we have to control only a few groups of joint motions to control the animated characters' motion. Moreover, we extract the features such as "speed," "posture," and "amplitude" from the captured data and use them as the human motion control parameters. Further, we apply them to standard motions, such that the various motions are generated intuitively.

In our study, we use the walking motion, which is one of the most standard human motions, and formulate a variation. Additionally, we construct the human walking motion model to control the animated characters' walking motions.

2 Definition of the Walking Motion Model

Our human walking motion model consists of two elements: the walking motion representation model and the walking motion control model.

2.1 Walking Motion Representation Model

Initially, in the motion space in which we assume all possible human motions to be represented, we construct the subspace of the human walking motion by analyzing the captured data. Namely, the subspace implies the restriction and freedom of the human walking motion. Moreover, the human walking motion is generated by the interaction between each joint. Therefore, we construct the representation model by analyzing the correlation between each joint

motion, which is one of the most important restrictions.

2.2 Walking Motion Control Model

Then, we clarify the mechanism that controls various walking motions by extracting features from the captured human walking motions with some physical alterations. Using the weighted sum of the extracted features, we can control the various walking motions of the animated characters.

3 Construction of the Walking Motion Model

3.1 Walking Motion Data

3.1.1 Captured Human Walking Motion Data

The various walking motions seem to be caused by a combination of features such as "speed," "amplitude," and "posture." Therefore, we use the following eight types of human walks to construct the walking motion model.

- Normal walk
- Fast walk
- Slow walk
- Walk with head held high
- Walk with head held low
- Walk with swinging arm movement
- Walk with vigorous leg movement
- Walk with short steps

We define "Normal walk," which has no features, as the standard walking motion. "Slow walk" and "Fast walk" are the walking motions in which the feature of "speed" is added to the standard walk. "Walk with swinging arm movement," "Walk with vigorous leg movement," and "Walk with short steps" are the walking motions in which the feature of "amplitude" is added to the standard walking motion. "Walk with head held high" and "Walk with head held low" are the walking motions in which the feature of "posture" is added to the standard walking motion. We use these motions as the walking motions with physical alterations.

We captured walking motion data of 20 male college students with an optical motion capture system, and converted it to BioVision Hierarchical data format (BVH data format) originally developed by Biovision, a motion capture services company, in which the hierarchy and initial pose of the human body model is defined and the joint angle data is described, as shown in Figures 1 and 2, respectively. In our BVH data human body model with 19 joints, the angle data for every joint and the location of root are described. The total number of parameters in every frame is $60 = 19 \times 3 + 3$.

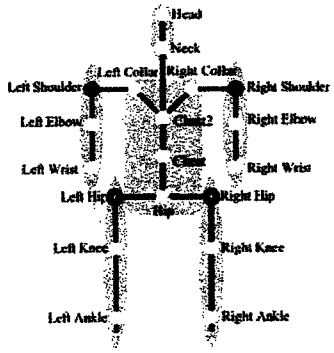


Figure 1. Human body model

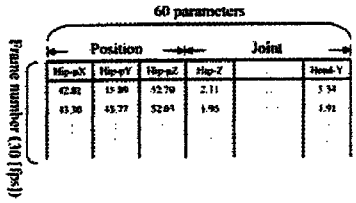


Figure 2. BVH format

3.1.2 Transformation into a Spectrum

Every joint motion in a real human walking motion is periodic; hence, it can be regarded as a periodic continuous-time signal. Therefore, the signal $\Theta_m(t)$ of the m-th joint angle is expanded into a Fourier series

$$\Theta_m(t) = \sum_{k=-\infty}^{\infty} C_{mk} e^{i\frac{2\pi k}{T}t}, \quad (1)$$

$$C_{mk} = \frac{1}{T}\int_0^T \Theta_m(t) e^{i\frac{-2\pi k}{T}t} dt, \quad (2)$$

where C_{mn} is the amplitude of the n-th harmonic component, and T is the period. It should be noted that the captured walking motion data is a discrete-time and approximate periodic signal; therefore, the amplitude is calculated by a discrete Fourier transform

$$C_{mk} = \sum_{n=0}^{N-1} \Theta_{mn} e^{-i\frac{2\pi nk}{N}}, \quad (3)$$

and the captured walking motion data is reproduced by an inverse discrete Fourier transform

$$\Theta_{mn} = \frac{1}{N}\sum_{k=0}^{N-1} C_{mk} e^{i\frac{2\pi nk}{N}}. \quad (4)$$

Yukawa et al. shows that all the joint motions can be sufficiently represented as a spectrum up to the third harmonic component [7]. Figure 3 shows the wave pattern of the captured back-and-forth motion of the right knee and the pattern approximated by up to the third harmonic. When a discrete walkimg motion data is periodic with a period N, we must control N parameters in order to control every joint motion. However, the approximation allows us to control only three parameters for every joint motion. Therefore, we construct the model with the spectrum up to the third harmonic.

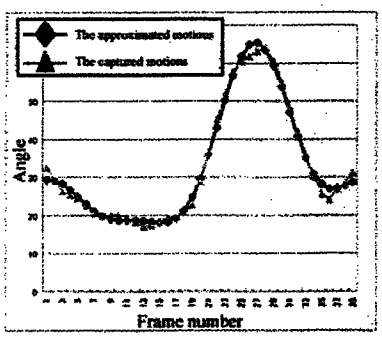


Figure 3. Approximated motion by up to the third harmonic

3.2 Construction of the Walking Motion Representation Model

3.2.1 Rule of the Joint Motions Correlation

To clarify the correlation of each joint motion in the human walking motion, we formulate a rule for it using multiple linear regression analysis. Initially, we define the key joint motions of walking as the joint motions that appear to correlate with several other joint motions. Then, as illustrated in Figure 4, we use the key joint motions spectrum as explanation variables and the other joint motions spectrum as purpose variables to calculate a group of functions by which the other motions are generated from the key motions using multiple linear regression analysis. We assume the group of functions to be a rule of each joint motion correlation.

3.2.2 Classification of the Joint Motions

We define four types of joint motions: "left hip," "right hip," "left shoulder," and "right shoulder," as key joint motions I, and formulate a rule for each joint motion correla-

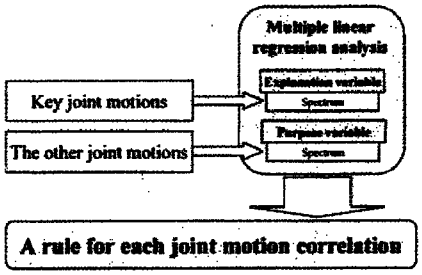


Figure 4. Formulate a correlation rule for each joint motion

tion in the human walking motion. Using the rule, we then generate the walking motions from the key joint motions I in the other captured human walking motions. We calculate the correlation coefficient between the generated walking motions and the original captured walking motions. We use 15 walking motions and calculate the average of the correlation coefficients for every joint motion.

Table 1 shows the averages of the correlation coefficients of the joint motions in the right leg. It is clear that certain joint motions can be generated from the key motions, while others cannot be sufficiently generated.

Table 1. The averages of the correlation coefficients of the right foot

Joint motions	Correlation coefficients
Right Knee-Z	0.8553
Right Knee-X	0.9410
Right Knee-Y	0.7048
Right Ankle-Z	0.6801
Right Ankle-X	0.8541
Right Ankle-Y	0.3531

We define the dependent joint motion I as the motion for which the correlation coefficient is over 0.5 on an average. Therefore, the dependent joint motions I can correlate closely with the key joint motions I. In other words, the dependent joint motions I can change regularly depending on the change in the key joint motions I.

On the other hand, we define the independent joint motion I as the motion for which the correlation coefficient is under 0.5 on an average. Therefore, the independent joint motions I cannot be generated using the rule. In other words, the independent joint motions I can change, the changes being independent of the key joint motions I.

Moreover, we define the key joint motions II as the motions which are a subclass of the independent joint motions I and seem to correlate with several other independent joint motions I. Further, we classify the independent joint motions I into the dependent joint motions II and do the same for the independent joint motions II.

Figure 5 shows the classified joint motions. Finally, we classify all the 57 joint motions in the human walking motion into 40 joint motions that correlate with each other: the key joint motions I, II, the dependent joint motions I, II, and 17 joint motions that change independently: the independent joint motions II.

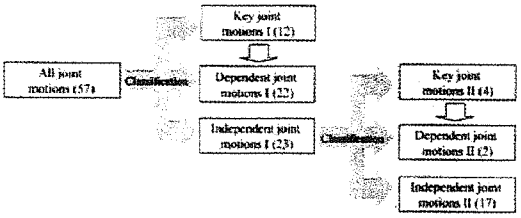


Figure 5. Classification of the joint motions

3.2.3 Structure of the Joint Motions

Further, we clarify the structure of all the joint motions in the human walking motion as shown in Figure 6. It follows that the joint motions that correlate closely with each other can be grouped into the upper body and lower body joint motions. Conversely, most of the independent joint motions are motions that occur at the end of the body.

3.3 Construction of the Walking Motion Control Model

3.3.1 Extraction of the Control Features

In this section, we extract the features such as "speed," "amplitude," and "posture" from the captured human walking motions. In section 3.2., we classified the joint motions into certain groups wherein the motions correlate with each other. Therefore, we have to control only some features in the joint motions of every group to generate and control the various walking motions of the animated characters. It should be noted that the independent joint motions II are not controlled because they do not correlate with the other joint motions. Moreover, as shown in Figure 6, the joint motions that correlate closely with each other can be grouped into the upper body and lower body joint motions. Therefore, we use two types of "amplitude": "amplitude of arms" and "amplitude of legs." In other words, we extract four features such as "speed," "posture," "amplitude of arms," and "amplitude of legs," to control the walking motions of the animated characters. Tables 2 and 3 show the human walking motion data used and the group of joint motions that correspond to every feature, respectively.

We explain the feature extraction method. For this extraction method, we can choose Unuma's method [7] or Kraus's metod [8], which provide possibilities of application of extracted emotional part of the motion. But this

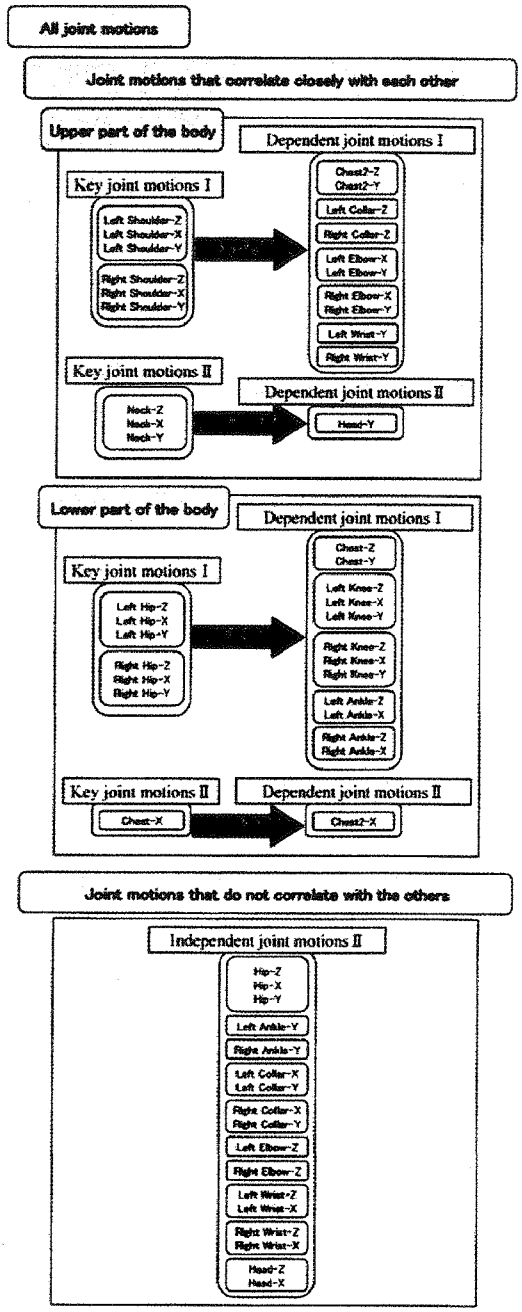


Figure 6. Structure of the joint motions

method can not generate the variable motions, so we proposed original method which can generate the variable motions based on these methods.

Initially, we define the standard walking motion spectrum to be the average of 20 spectrums of the captured "Normal walk" motions. It is represented as

$$\mu_{mk}^{(S)} \equiv E\left[C_{mk}^{(S)}\right]. \tag{5}$$

Then, we calculate the expected value of 20 spectrums of the captured walking motions with every physical alteration

$$\mu_{mkl}^{(P)} \equiv E\left[C_{mkl}^{(P)}\right], \tag{6}$$

and the variance

$$\left\{\sigma_{mkl}^{(P)}\right\}^2 \equiv E\left[\left\{C_{mkl}^{(P)} - \mu_{mkl}^{(P)}\right\}^2\right], \tag{7}$$

where l is the index of the types of features. The feature corresponding to every physical alteration is given by

$$\delta_{mkl} \equiv \mu_{mkl}^{(P)} - \mu_{mk}^{(S)}. \tag{8}$$

If the captured walking motion data with physical alterations is sufficient, we assume that the variation can be represented by a normal distribution $N\left(\mu_{mkl}^{(P)}, \sigma_{mkl}^{(P)}\right)$. Therefore, we add variation to the controlled walking motions by adding certain other features, using the normal distribution of the physical alterations in the walking motions.

Table 2. Walking motion data for every feature

Feature	Walking motions
Speed	Fast walk Slow walk
Posture	Walk with head held high Walk with head held low
Amplitude of arms	Walk with swinging arm movement Walk with short steps
Amplitude of legs	Walk with vigorous leg movement Walk with short steps

Table 3. Group of joint motions for every feature

Feature	Group of joint motions
Speed	Upper part of the body + lower part of the body
Posture	Upper part of the body + lower part of the body
Amplitude of arm	Upper part of the body
Amplitude of leg	Lower part of the body

3.3.2 Controlling the Walking Motions

We generate the various walking motions by adding certain features and variations to the "Normal walk." Therefore, the generated walking motion is computed by

$$\Theta_{mn} = \sum_{k=0}^{2}\left[\mu_{mk}^{(S)} + \sum_{l=0}^{3}\left\{w_l\delta_{mkl} + \lambda_{mkl}\right\}\right] e^{i\frac{2\pi nk}{N}}, \tag{9}$$

where, w_l is the weight or control parameter for every feature, and λ_{mkl} is the variation that is calculated using the normal distribution $N\left(\mu_{mkl}^{(P)}, \sigma_{mkl}^{(P)}\right)$. Figure 7 depicts the method of controlling the walking motions.

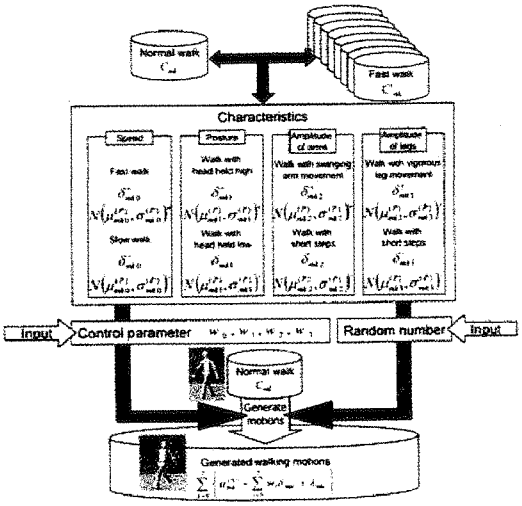


Figure 7. The method of controlling the walking motions

4 The Evaluation Experiment of the Model

In this chapter, we use an experiment to evaluate the model proposed in Chapter 3.

4.1 The Evaluation Experiment Method

We use the captured walking motions of 11 college students with four positive or negative physical alterations and three emotions as the target walking motions:

- Normal walk
- Fast walk
- Slow walk
- Walk with head held high
- Walk with head held low
- Walk with swinging arm movement
- Walk with vigorous leg movement
- Walk with short steps
- Walking happily
- Walking angrily
- Walking sadly

Then, we generate the walking motions using our model for the target walking motions, to clarify the motion generation ability of our model. In this experiment, every control parameter w_l in our model is altered in 61 steps; therefore, 61^4 walking motions can be generated. Further, we generate all of these walking motions to evaluate our model based on their similarity to the target motions. Conversely, the variation term λ_{mkl} in our model is changed depending on the normal distribution $N\left(\mu_{mkl}^{(P)}, \sigma_{mkl}^{(P)}\right)$ every time the control parameter w_l is changed. Therefore, it should be noted that all the possible walking motions using our model are not generated in this experiment.

Further, we estimate the parameter that minimizes the mean squared error between the generated motions and the target motions. Finally, we calculate the correlation coefficient between the target motions and the estimated motions, and use its average as an evaluation value.

4.2 The Result of the Evaluation Experiment

Table 4 shows the Reproduction degree as the average of the correlation coefficients of 40 joint motions, except for the 17 independent joint motions that do not have much effect on the other joint motions. So the Reoroduction degree(RD) is given by

$$RD = \sum_{m=0}^{39} r_m^{(TE)}, \tag{10}$$

wehre $r_m^{(TE)}$ is the m-th joint correlation coefficients between the target motions ($\Theta_{mn}^{(T)}$) and the estimated motions ($\Theta_{mn}^{(E)}$). It is represented as

$$r_m^{(TE)} = \frac{\sum_{n=0}^{N-1}(\Theta_{mn}^{(T)} - \overline{\Theta_m^{(T)}})(\Theta_{mn}^{(E)} - \overline{\Theta_m^{(E)}})}{\sqrt{\sum_{n=0}^{N-1}(\Theta_{mn}^{(T)} - \overline{\Theta_m^{(T)}})^2}\sqrt{\sum_{n=0}^{N-1}(\Theta_{mn}^{(E)} - \overline{\Theta_m^{(E)}})^2}}.$$

Table 4. The result of the evaluation experiment

Walking motions	Reproduction degree
Fast walk	0.8770
Slow walk	0.9155
Walk with head held high	0.9345
Walk with head held low	0.8961
Walk with swinging arm movement	0.9407
Walk with vigorous leg movement	0.9173
Walk with short steps	0.7675
Walking happily	0.5511
Walking angrily	0.6436
Walking sadly	0.5039

It is clear that for every target walking motion with physical alteration, our model generates the appropriate motion; however, for the walking motions with emotions, it does not generate the appropriate motions adequately. This is caused by the inadequacy of the motion control features used: "speed," "posture," "amplitude of arms," and "amplitude of legs," to generate the walking motions with emotions. Therefore, we need to use additional features such as "walking with bow-legs," "walk putting up things," and so on.

5 Conclusions

We have constructed a walking motion model that can adequately control various walking motions. Initially, we calculated the spectrum up to the third harmonic using a discrete Fourier transform for the captured human walking motions. Then, we formulated a rule for each joint motion correlation by multiple linear regression analysis, to clarify the hierarchical structure of joint motions in the human walking motion. Based on this result, we extracted features from the captured human walking motions with physical alterations such as "speed," "posture," "amplitude of arms," and "amplitude of legs," and constructed a model that could generate various walking motions based on the weight of the features. As a result of the evaluation experiment of the model, our model generates the appropriate motion for the target walking motions with physical alteration; however, for the walking motions with emotions it does not generate the appropriate motions adequately. We endeavored to construct a model that could generate and control a greater variety of walking motions using additional features.

References

[1] K. KAWACHI and H. SUZUKI, Realtime Motion Synthesis with Motion Database, *TVRSJ*, 7(4), 2002, 453-458.

[2] Y. Kurihara, S. Kuriyama, and T. Kaneko, Real-time controls of human gait with pattern generators, *IPSJ Graphics and CAD*, 100(13), 2000, 73-78.

[3] C. Rose, B. Guenter, B. Bodenheimer, M. F. Cohen, Efficient generation of motion transitions using spacetime constraints, *SIGGRAPH 96*, New Orleans, Louisiana, USA, 1996, 147-154.

[4] Y. Yoshida, Y. Matsuoka, and K. Hachimura, Processing of Human Body Movement based on Labanotation - LabanReader and LabanEditor-, *IPSJ/SIG Computers and Humanities*, 38(6), 1998, 61-68.

[5] T Komura, Y Shinagawa, and T L Kunii, Muscle-based feed-forward controller of human body, *Computer Graphics Forum*, 16(3), 1997, 165-176.

[6] T Komura, Y Shinagawa, and T L Kunii, Creating and retargetting motion by the musculoskeletal human body model, *The Visual Computer*, 5, 2000, 254-270.

[7] M. Unuma, R. Takeuchi, Generation of Human Walking Motion with Emotion for Computer Animation, *IEICE Journal*, J76-D-II(8), 1993, 1822-1831.

[8] M. Kraus, Human motion and emotion parameterization, *CESCG 2004*, Budmerice castle, Slovakia, 2004.

対面対話の収録と解析のための信号処理とパターン認識

松坂　要佐　　小林　哲則
Yosuke MATSUSAKA　　Tetsunori KOBAYASHI
早稲田大学理工学部
School of Science and Engineering, Waseda University
{yosuke,koba}@tk.elec.waseda.ac.jp

ABSTRACT

対面対話の収録と解析のための信号処理技術とパターン認識技術の開発を行っている。その中で、対話音声の調波構造に着目した混合音分離法は、防音壁のない状態で収録された対話音声を比較的簡易な処理で話者ごとの音声に分離する技術であり、対話研究のツールとして有用なものであると考える。本稿では、その原理と性能について紹介する。また、対話における話者の頭部動作の認識などのパターン認識技術についても紹介する。

This paper describes our current effort to realize the recording and the analysis of the face-to-face conversation data. We developed a signal processing technique to realize the separation of speech signal, recorded without the sound barrier. In this paper, we introduce the principle and the performance of the developed technique. Also, we introduce the pattern recognition techniques to realize the analysis of the head movement of the participant of the conversation.

1. はじめに

近年対話の研究、中でも対面対話の研究の持つ重要性が非常に高まってきている。擬人化エージェント・ロボットなどへの応用、人間同士の活動の補助を行う CSCW システムなどへの応用、掲示板やメールなど非同期コミュニケーションの円滑性向上のための基礎的なデータを得る手段として、広い分野の人間が人間同士の対話の研究に興味を持ち始めている。

近年の信号処理技術やパターン認識技術の進歩により、情報通信やヒューマンインターフェース分野において、音声・画像などのマルチモーダルな情報を扱うシステムが良好な性能をあげることができるようになってきた。それら背景の中で、工学的方法論を用いた対話研究へのアプローチが可能となった。例として、アイトラック技術を用いた Vertegaal による解析[2]、モーションキャプチャ技術を用いた綿貫らによる解析[3]などがある。我々は、数年前から対面対話、特に多人数の話者がかかわる対面対話に興味をもち、対話におけるマルチモーダル対話情報の収録と解析を行うシステムの作成に取り組んできた[4]。

対面対話の解析は、ふるまい等の煩雑で記号化の難しい情報を多く扱わなければならないので、工学的な方法論が特に役に立つ分野であると考えている。本稿においては、その中で特に信号処理技術とパターン認識技術に焦点を当てる。

2. 対面対話の収録と解析における問題

2.1 収録における問題

話者と話者が近距離で対峙して行う対面対話においては、話者の音声の収録における困難が存在する。例え話者ごとにマイクを設置したとしても、ある話者が発生した音声は、その近傍に設置されたすべてのマイクに収録されてしまう。音声の回り込みの問題は、対面対話の収録における大きな問題点のひとつである。

対面した状態でそのまま収録された音声を利用した場合、自然な対話の収録を行うことができる反面、音声が混合して収録されているため、その後の解析の段階において困難が生じていた。クリーンな音声を収録するために、話者と話者の間に透明の防音壁を設置することが一般に行われるが、防音壁の設置には下記のようなデメリットがある。

収録装置の巨大化：対話の収録のために専用の部屋を用意する必要があり、手軽に収録を行うことができない。また、収録機器を手軽に対話の行われている場所に移動してセットアップすることが難しい。

臨場感・定位感の喪失：話者同士がワークスペースを共有して作業するような状況における対話を収録するのが難しい。また、話者がどこから発話を行っているのかといった音像の定位感を伝えることが難しい。定位感の喪失の問題は、特に、多人数の対話を収録する場合においては、致命的であると考える。

被験者への心理的影響：対話の収録のために特別に作られた部屋や巨大な防音壁は、被験者へ圧迫感を与えることが少なくない。

これらの問題に対して、本研究では、対話の収録に防音壁は設置せずに、対面対話を音声が混合されることを前提として収録し、後処理によって満足のいく品質に分離することを試みる。これらを実現する信号処理手法を3章において議論する。

2.2 解析における問題

対面対話は非常に豊かな対話である。言語による情報伝達だけでなく、視線や表情などさまざまなチャネルを通して表現のやり取りが行われる。したがって、対面対話を理解するためには、言語レベルでの解析だけでなく、振る舞いレベルでの解析を行う必要がある。しかしながら、振る舞いレベルのデータは非常に煩雑で膨大であり、従来のような人手に頼った解析で十分な量のデータを得ることは難しい。また、振る舞いレベルの解析を行うためには、従来の言語レベルでの解析で用いられていたような記号とは異質の記号を用いる必要がある。日常的に訓練された記号である言語でのラベル付け作業に比較して、定義の難しい振る舞いのような現象を記号化し、揺らぎなくラベルを振ることは、人間にとって非常に難しい作業である。

我々は、パターン認識技術は、これら解析における問題を補助する上で有力な手段であると考えている。また、パターン認識技術の開発後に、ブートストラップ的に解析データを増やすことができるメリットも大きい。4章において、我々が開発中の頭部動作のスポッタの概要について紹介する。

3. 混合された対話音声の分離

3.1 周波数軸上での音声の分離手法

自然な対話の収録を実現するために、一般的な対話の収録で行われているような、話者間への防音壁の設置を行っていない。したがって、各話者の胸元に設置されたマイクで収録した音声には、他の話者が発生した音声が混合されて収録されている。対話の解析へと進む前に、この混合された音声を分離するための前処理を検討する必要がある。

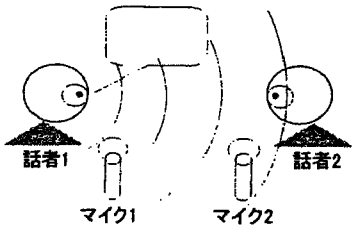


図1：対話音声収録時における音声の回り込みの問題。話者から発話された音声は、すべてのマイクに収録されてしまう。

混合された音声を分離する手法にはいくつかの方法が提案されているが、そのなかで、周波数領域上でサブバンドごとに音声を分離する手法[5][6]を採用した。これは、比較的制約の少ない状況で収録された音声に適用可能であり、話者ごとの平均基本周波数の差や、対話音声における文頭・文末の周波数の差異から、音声の重畳区間においてもある程度の分離性能が得られることが見込まれるからである。以下、分離手法の原理と性能について説明する。

3.2 音声スペクトルの基本構造

(1) 音声の生成プロセス

人の音声の生成プロセスには大きく二つの種類に分けられる。ひとつは、声帯の振動と調音器官の形によって生成されるプロセスであり、そのプロセスを経て発生された音声は、有声音と呼ばれる。もうひとつは、声道の狭めや閉鎖と急激な開放に伴って発生されるプロセスであり、そのプロセスによって生成された音声を無声音と呼ぶ。

有声音の生成プロセスにおいては、声帯の振動によって、まず特定の周波数(基本周波数と呼ばれる)とパワーを持った信号が生成される。生成された信号は、顎、舌、唇などの器官(まとめて声道と呼ばれる)を通ることで、さまざまに変化し、口または鼻から放射される。

無声音の生成プロセスにおいては、空気が狭められた声道を通過する際の摩擦音や、声道の閉鎖と急

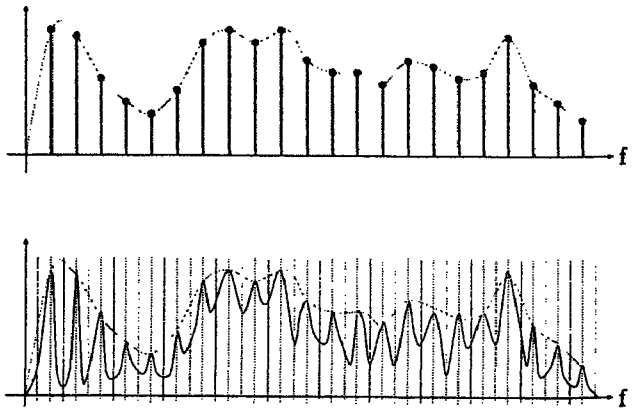


図2：音声の持つ周波数軸上での構造
上が理想的に考えた場合の調波構造。下は実際のスペクトルの構造。

峻な開放により発生する破裂音が、口や鼻から放射される。

(2) 基本周波数と調波構造

有声音における音声は、単一の周波数を持った元信号に対して、声道における共振などのフィルタ処理を加えたものと考えることができる。フィルタ処理によって原音の持つ基本周波数以外の周波数成分が作り出され、それが音色の違いとして知覚される。原音とする信号が単一の周波数であるため、発生された音声の周波数成分も元信号の周波数に強いピークを持つ。フィルタによって作り出されたその他の周波数成分も、元信号に対する受動的な処理により生成された信号であるため、元信号の特性を保存する。具体的には、元信号の持つ周波数を n 倍した周波数にそれぞれのピークが現れる。

有声音を周波数領域で表現すると図 2 のようになる。それぞれの周波数成分は、基本周波数とその n 倍の周波数に強いピークを持った櫛状の形状になる、これを調波構造という。音の周波数成分は、理想的には、それぞれの周波数に鋭いピークを持った構造を持つはずであるが、現実には、短時間スペクトルを求めるために用いる窓関数の影響で、ピークを中心としてある程度の広さの裾野を持った形状となる(図 2)。

音色の決定要因となるスペクトルの概形を表わす言葉としてフォルマントという表現が使われることがあるが、それに対して調波構造を指し示す場合、スペクトルの微細構造と呼ぶことが多い。

3.3 回り込み音声分離の原理

(1) 回り込み音声のスペクトル形状

回り込み音声を考えた場合、その音は話者の最も近くで収録された音に比較して、空気中を伝播することで、パワーが弱まった状態で収録される。

パワーは弱まるが、波形の時間軸上での伸縮は起こらないため、周波数成分の構造が変化することはない。それぞれの場所で収録された音声のスペクトルは、ほぼ同一の調波構造を持ちながら、パワーの差を持った形状となる。

(2) 音声の重畳区間におけるスペクトル形状

複数の話者が重畳して発話している場合のスペクトル形状を考えてみる。

ここでは、2 話者の発声した音声を 2 つのマイクで収録した場合に単純化して考える。それぞれの話者を話者 1・話者 2、それぞれの話者の近傍に設置されたマイクをマイク 1・マイク 2 とすると、話者 1 の発声した音声は、マイク 1 とマイク 2 に対して同じ調波構造を持って収録される。しかしながら、パワーは、マイク 1>マイク 2 となっているはずである。同様にして、話者 2 の発声した音声は、調波構造においては、マイク 1=マイク 2 であり、パワーにおいては、マイク 1<マイク 2 となっている。各々のマイクに収録される音声のスペクトルの形状は結果として、それらそれぞれ入力の持つスペクトル成分を重ね合わせた形状を持つ。この形状を模式的に図 4 に示す。

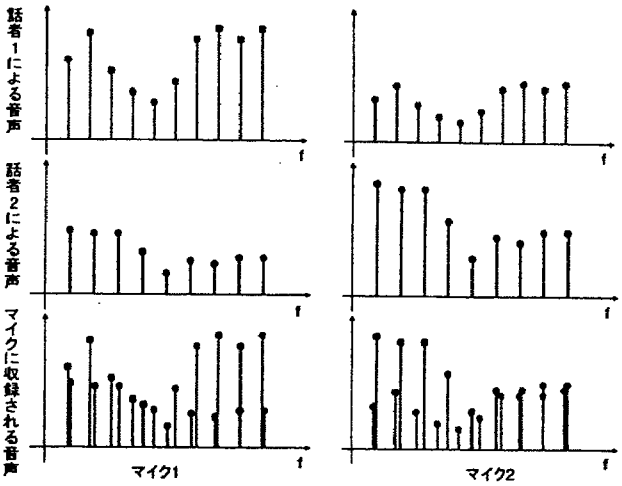


図 4：収録された音声のスペクトル形状
上段が話者 1 からの発話のスペクトル成分、中段が話者 2 からの発話のスペクトル成分、下段が最終的にマイクに収録されるスペクトル形状(左列がマイク 1、右列がマイク 2)。

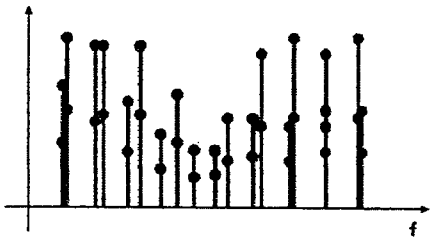


図 5：2 チャンネルのピークの重ね合わせ
両チャンネルのスペクトルのピークは、同一の箇所に出現し、パワーの差などの情報からどちらの話者から発話されたかを判別できる。

(3) 分離の原理

収録音声の分離の原理を以下に示す。

収録されたそれぞれのチャンネルのスペクトルは、複数のピークを持った形状をしている。これらのピークは発声されたそれぞれの音声に帰属するものであり、音声の分離の問題はこれらのピークがそれぞれどの話者からの発声に帰属するかを判別する問題となる。

(2)において示したように、これらのピークは、収録されたそれぞれのチャンネルのスペクトル上の同じ周波数帯に出現する。したがって、複数のチャンネルの同一の周波数帯に着目し、(2)で示した各チャンネルにおけるパワーの差などを手がかりとして、そのピークがどの話者からの発声に帰属するかを判別することができる。

判別・分離した各々のピークを音声波形として再合成することで、話者ごとに分離された音声が復元できる。

以上の説明は、収録された各チャネルにおけるスペクトルを重ね合わせてみると理解がしやすい(図 5)。

3.4 分離を実現するために必要な前提条件

上記した原理に基づく分離を実現するために必要な前提条件を以下に示す。

1. 対象とする音が調波構造を持った音である
2. 音声の各周波数成分の持つパワーが、話者毎に設置したマイクに対して、対象話者が最も強く、それ以外の話者が弱く収録されている
3. 音声の重畳区間においては、2 話者の発声する音声に、基本周波数のずれがある

上記した前提条件が、実際の収録の状況において、どれだけ妥当であるか検討する。

条件 1 については、日本語音声の多くの部分は、有声音であることが知られており、発話区間の多くの部分が条件に当てはまることが期待できる。しかしながら、運悪く音声の重畳区間に無声音が発生した場合、ある話者が発声した無声音区間にもう一人の話者の有声音区間が重なることで、無声音であったはずの音声が弱い調波構造を持った形で分離されたり、無声音同士が互いの成分を打ち消しあったりする可能性は否定できない。

条件 2 については、収録時において、マイクの配置に留意をする必要がある。無指向性のマイクを用いて収録する場合、一般に音声はどの周波数であっても音源から収録機器までの距離が離れるほど減衰するので、対象話者の収録マイクは、常にほかの話者のマイクよりも、対象話者の近くの位置にある必要がある。指向性を持ったマイクで収録する場合、そのマイクの持つ指向特性に配慮してマイクの配置と角度を設定する必要がある。人の口からの発声には、弱い指向性しかないので、マイクの取り付け角度にはある程度の自由度がある。多くの場合、あえて口元に設置するなどしなくても、胸元につける程度で十分に条件を満たすことができるであろう。これら収録におけるマイクの距離・角度の条件は、多くの対話の収録の状況において容易に満たすことができると考える。

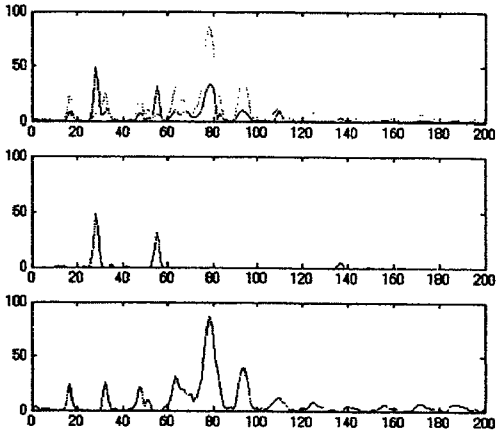


図 6：混合された対話音声の分離処理
上段は、2チャンネルの入力音声のスペクトルパターンを重ねたもの。中段、下段は、それぞれ分離された 2チャンネル音声のスペクトルパターン。

条件 3 については、平均基本周波数が一般に話者ごとで異なることや、音声における基本周波数の変動が複雑なパターンを持つことから、たとえ同時に同じ内容の発話を行ったとしても、2 話者の発話した音声の調波構造が完全に重なることは稀である。また、実際にはこれらの処理は、前フレームと重複した時間窓を少しずつシフトしていく形で行われるため、たとえある区間の中で数フレームの欠損があったとしても、再合成された音声の可聴性に与える影響は少ない。

3.5 分離処理

音声分離処理の概要は以下の通りである。

1. 各チャンネルの音声の周波数変換
2. 周波数軸上で、サブバンドごとの比較・フィルタ処理
3. 各チャンネルの逆周波数変換

上記処理をフレームごとに行っていく。

周波数軸上でのフィルタ処理には、もっとも単純には各々の周波数上でパワーの大小比較をすればよい。公開ツール(3.7 節参照)においては、現状では上記した単純な大小比較による実装となっているが、音声の調波構造やフレーム間の連続性、各周波数における位相情報などを評価することで、収録状況の悪い音声におけるチャンネル間のパワーの逆転を防ぐことや、複数話者の音声が複雑に重畳する区間において分離性能を向上できる可能性が残されている。

3.6 性能と限界

図 7 にピークの判別結果を着色したスペクトログラムを示す。複数話者の発話が重畳する区間においても、美しい調波構造が確認できることから、妥当な分離が行われていることが見て取れる。

処理の結果、単一の話者が発話している区間のみならず、複数の話者が重畳して発話している区間においても良好な分離性能を得ることができた。しかしながら、周波数構造が複雑に重畳する一部の区間において、ワブノイズの湧き出しが認められた。ノイズへの対策として、収録時のサンプリング周波数を上げて周波数領域における時間・周波数解像度を向上させることや、音声の持つ特性を手がかりとしてフィルタの性能を上げること(3.5 節参照)などが考えられる。

分離性能は、比較的制約の少ない状況で収録した音声にもかかわらず、多くの部分で非常に良好であり、ノイズが発生している区間においても発話内容がわからなくなるほどの可聴性の低下は確認されなかった。また、本手法は、音声スペクトルの基本構造を利用した分離手法であるため、分離後も発話内容のみならず、発話のパワーや基本周波数などの韻律的情報も保存されている。当初の目標のような対話研究のツールとしての役割は十分に果たすものと考える。

3.7 公開ツール

上述した信号処理による対話音声の分離ツールをホームページ[9]にて公開している。ツールの使用方法や使用条件、最新の情報の取得などについては、同ホームページを参照されたい。

4. パターン認識による自動ラベル

対話収録データの画像中に現れる典型的な頭部ジェスチャを図 8 に示す。対話に現れる頭部ジェスチャには、例えば、うなずき・かしげには、対話のリズムの調整や理解のフィードバックとしての役割が、視線の動きやアイコンタクトには、興味(アウェアネス)の伝達や、ターンテイクの調整など役割があると考えられる。これらのパターンの出現を分析するために、パターン認識技術を用いて収録データに自動的にラベルを振るシステムの構築を試みている。

4.1 静的パターン認識によるアイコンタクトの判別

ICA による特徴抽出に基づく静的パターン認識を用いて、対話中におけるアイコンタクトの判別を行うシステムを作成した。テストセットとして収録したデータに対して、90%以上の識別精度を持ち、実環境においても高い性能を達成している[7]。

4.2 動画像処理による頭部動作の認識

頭部ジェスチャの動き特徴量として、オプティカルフローによる特徴抽出と HMM による時系列データをモデル化を用いて、うなずき、かしげ、首振りのスポッティングを行うシステムを作成している。自発話、相手発話の韻律的特徴に基づく頭部ジェスチャ生起の事前確率と統合することで、オペレータにより振られたタグと高い一致率を示すスポッタを作成することに成功している[8]。

5. 本稿で使用した対話収録システムの概要

本稿において使用した対話収録システムの概要を図 9 示す。システムの詳細については、文献[4]を参照されたい。

6. まとめ

対面対話の収録と解析のための信号処理技術とパターン認識技術の開発を行った。その中で、対話音声の調波構造に着目した混合音分離法は、防音壁のない状態で収録された対話音声を比較的簡易な処理

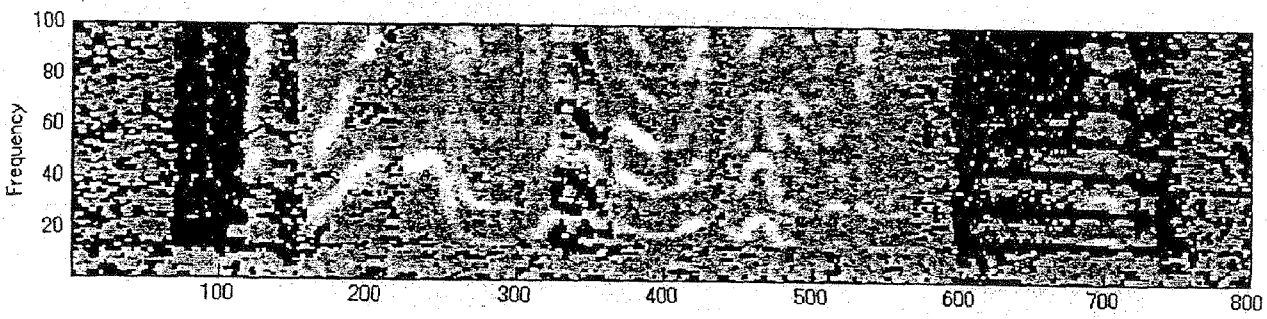


図 7：対話音声の重複区間における色つきスペクトログラム
白く着色されたのが話者 1、黒く着色されたのが話者 2 のスペクトログラム。時間軸上では混合しているように見える音声も、周波数軸上ではほとんどが重ならない構造を持つ。

図 8：対話に現れる典型的な頭部ジェスチャ
アイコンタクト(左上)、笑い(右上)、傾げ(左下)、
想起(右下)

で話者ごとの音声に分離する技術であり、対話研究のツールとして有用なものであると考える。その原理と性能について紹介し、ツールとして公開した。また、対話における話者の頭部動作の認識などのパターン認識技術についても紹介した。

参考文献

[1] 松坂要佐，東條剛史，小林哲則，“グループ会話に参与する対話ロボットの構築，” 電子情報通信学会論文誌 Vol.J84-D-II No.6, pp.898-908, 2001.

[2] R.Vertegaal, “Look who's talking to whom? Mediating Joint Attention in Multiparty Communication and Collaboration,” Cognitive Ergonomics Department University of Twente, 1998.

[3] 綿貫啓子，関進，三吉秀夫，“発話時の人間の振舞い -マルチモーダル対話データの解析-,” 人工知能学会研究会資料 SIG-SLUD-9902, pp49-54, 1999.

[4] 松坂要佐，小田勇一郎，小林哲則，“グループ会話におけるマルチモーダル会話データの収録・解析システムの構築，” 第 15 回人工知能学会全国大会 3B2-02, 2001.

[5] 田中英智，小林哲則，“音声認識のためのマイクロフォンアレーによる周波数構造を考慮した音源分離，” 日本音響学会 2000 年秋季講演論文集, 2-7-10, pp.429-430, 2000.

[6] 青木真理子，古家賢一，吉岡理，山口義和，“チャネル間情報を利用した音源分離手法による混合音声の認識，” 日本音響学会 2001 年秋季講演論文集, 1-1-23, pp.45-46, 2001.

[7] 久保田千太郎，松坂要佐，小林哲則，“グループ会話ロボットにおける顔画像処理システム，” 電子情報通信学会技術報告 NLC99-85, PRMU99-268, pp.49-56, March, 2000.

[8] 江尻康，松坂要佐，小林哲則，“対話中における頭部ジェスチャの認識，” 電子情報通信学会 PRMU 研究会 (2002 年 7 月 19 日発表予定)

[9] convclean のホームページ (http://www.tk.elec.waseda.ac.jp/convclean/)

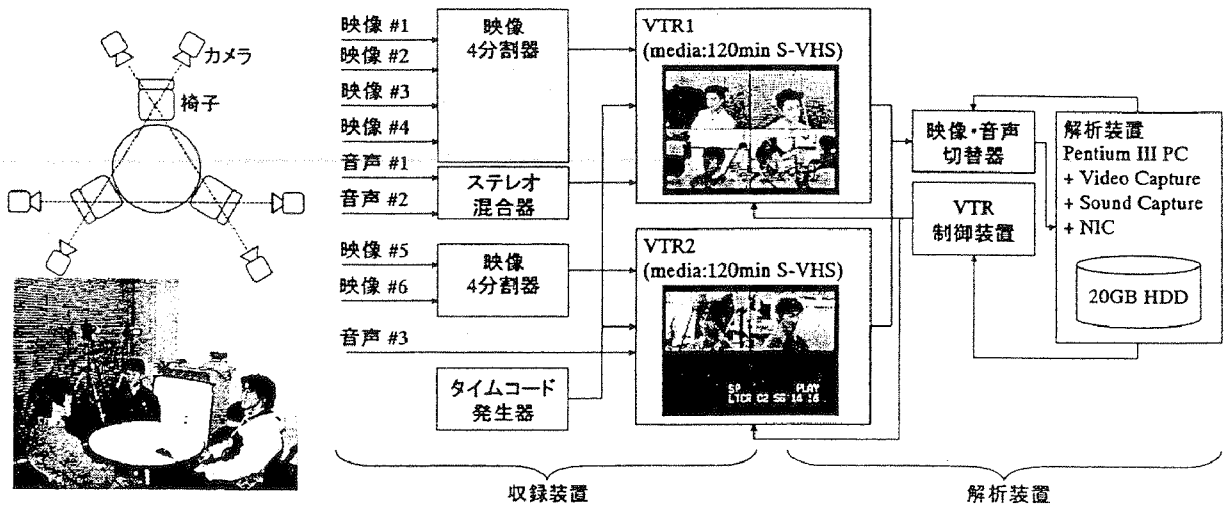


図 9：対話収録システムの概要と機器構成

社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
SP2002-40 (2002-06)

部分隠れマルコフモデルによる連続音声認識

小川 哲司† 小林 哲則†

† 早稲田大学 理工学部 〒169-8555 東京都 新宿区 大久保 3-4-1
E-mail: †{ogawa,koba}@tk.elec.waseda.ac.jp

あらまし 部分隠れマルコフモデル (Partly-Hidden Markov Model;PHMM) における状態と出力の依存関係を一般化し，連続音声認識に適用した．従来の PHMM においては，出力確率と状態遷移確率を決定するための隠れ状態と観測可能な状態の組は共通なものを用いていた．ここでは，出力確率と状態遷移確率を決定するにあたり，隠れ状態に関しては共通のものを使い，観測可能な状態に関しては，出力の決定と，状態遷移の決定に異なるものを利用することを考える．このような簡単な改良により大きな自由度が与えられ，より精度の高い確率過程のモデルを実現できる．連続音声認識実験の結果，PHMM はトライフォン HMM に比べ 19% の誤認識率を改善できることが示された．

キーワード HMM, PHMM, SPHMM, 半音節モデル, 連続音声認識

Continuous speech recognition by Partly-Hidden Markov Model

Tetsuji OGAWA† and Tetsunori KOBAYASHI†

† School of Science and Engineering, Waseda University 3-4-1 Okubo, Shinjyuku-ku, Tokyo, 169-8555 Japan
E-mail: †{ogawa,koba}@tk.elec.waseda.ac.jp

Abstract Generalization of state-observation dependencies in Partly-Hidden Markov Model (PHMM) is performed and it is successfully applied to the continuous speech recognition. PHMM, which was proposed in our previous paper, is the novel stochastic model, in which the pairs of the hidden states (H-state) and the observable states (O-state) determine the stochastic phenomena of the current observation and the next state transition. In the previous formulation of PHMM, we used common pair of H-state and O-state to determine both of these phenomena. In the formulation of modified PHMM proposed here, we use common H-state but different O-states for the current observation and for the next state separately. This slight modification brought the big flexibility in the modeling of phenomena. Experimental results showed the effectiveness of PHMM (without delta parameters): it reduced the word error by 19% compared to triphone HMM (with delta parameters), respectively.

Key words HMM, PHMM, SPHMM, Demi-syllable model, Continuous speech recognition

1. はじめに

現在，ほとんどの音声認識システムが音響モデルとして left-to-right 型の HMM を採用しているが，このモデルには様々な問題が存在している．例えば，このモデルでは各状態内で出力確率分布が一意に与えられるため，特徴量の局所的な動的変化情報がモデル内部に反映されず，区分定常な確率過程しか扱うことができない．一方，実際の音声スペクトルは，調音上の理由などにより渡りの部分で複雑な時間変化を生じる．

この問題を解決するために，より複雑な特徴量の時間変化を音響モデル内部で扱うことを目指して，様々な確率モデルが提案されている．坪香らや Deng らは，セグメント内での特徴ベクトルの軌跡を多項式近似するセグメント HMM を提案している [1] [2]．Wellekens は，HMM における出力確率を前フレームの出力の条件付き確率で表現する条件付き HMM を提案している [3]．Ostendorf は Wellekens のモデルを過去の複数フレームの条件付き確率で表現するよう拡張している [4]．その他にも，様々な拡張が試みられている [5] ～ [9]．

同様の目的で，我々は，二つの状態系列を有する確率過程のモデルを考え，一方を隠れ状態に，他方を観測可能な状態におくことで，単純な HMM よりも過渡部の表現能力に優れる新たな確率モデルを提案してきた．出力確率および状態遷移確率の条件部の一部を隠れ状態においたことから，このモデルを部分隠れマルコフモデル（Partly-Hidden Markov Model;PHMM）と呼んでいる．PHMM は，Wellekens や Ostendorf のモデルを特殊な形として含む，より一般性の高いモデルである．先行

研究が，出力確率のみを過去の出力による条件付き確率で表現しているのに対し，出力確率，状態遷移確率の双方を過去の出力による条件付き確率で表現している点が大きな特徴である．状態遷移確率が過去の出力と独立に決定する場合，一つの状態に停留する限り一定の状態遷移確率を与えるため，多くの場合どの時刻で状態遷移を行っても遷移確率の累積値には大差がない．つまり，この場合状態遷移確率は状態の切り替わり時刻に対して影響力を持たないと言える．このため，従来のモデルが抱えていた，異なるカテゴリのデータに対しても，出力確率の高い状態だけを選択的に使うことによって不当に高い尤度を与えてしまうという問題を，PHMM では解決している．

これまでに，PHMM はジェスチャ認識 [10] や孤立単語音声認識 [11] に適用が試みられ，特に過渡状態が多いジェスチャ認識において良好な結果が得られている．

従来の PHMM においては，出力確率と状態遷移確率を決定するための隠れ状態と観測可能な状態の組は共通なものを用いてきた．しかし，このような制約は必ずしも本質的なものではない．そこで今回，このような拘束条件を除くことにより，確率過程のモデル化において，より大きな自由度を与えることを試みる．具体的には，出力確率と状態遷移確率を決定するにあたり，隠れ状態に関しては共通するものを用い，観測可能な状態に対しては，出力を決定するものと，状態遷移を決定するものとで，異なるものを用いることを考える．この枠組みにより，PHMM は従来よりも離れたフレーム間における出力の依存関係を扱えるように一般化される．

以下，2. で本稿の基礎となる部分隠れマルコフモデル，および平滑化部分隠れマルコフモデルについて概観した後，3. において，PHMM における状態・出力依存関係の一般化について述べる．そして 4. で，連続音声を対象とした提案モデルの性能評価実験について述べる．この際，音素の切り替わり部分における性能を向上するためのモデリングの単位ついても検討を行なう．

2. 部分隠れマルコフモデル

2.1 部分隠れマルコフモデルの概要

提案するモデルでは，出力確率および状態遷移確率は次の 2 つの状態 s_t^h，s_t^o により，以下のように条件付けられるとする．

$$\begin{aligned} &\text{出力確率} \ : \ P_r(x_t|s_t^h, s_t^o) \\ &\text{遷移確率} \ : \ P_r(s_{t+1}^h|s_t^h, s_t^o),\ P_r(s_{t+1}^o|s_t^h, s_t^o) \end{aligned}$$

いま，状態 s_t^h を H 状態 (Hidden State) と呼び，隠れ状態におく．また，状態 s_t^o を O 状態 (Observable State) と呼び，観測可能な状態におく．出力確率および状態遷移確率の条件部の一部を隠れ状態においたことから，このモデルを部分隠れマルコフモデル (Partly-Hidden Markov Model;PHMM) と呼ぶことにする．

ここで，s_t^h と s_t^o が共に過去の出力から一意に決るとすれば，このモデルは単純マルコフモデルと等価であり，過去の出力から確率的に決るとすれば，隠れマルコフモデル (HMM) と等価である．

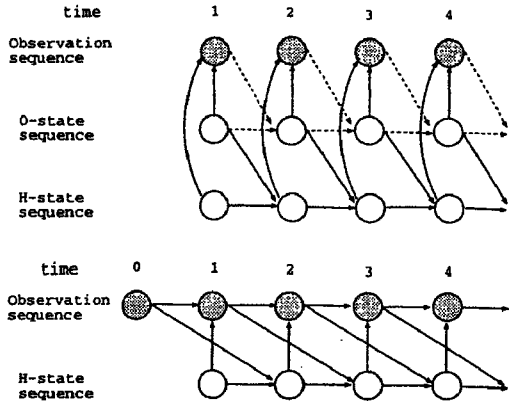


図 1 PHMM における状態・出力依存関係図 (上段：O 状態を表示したもの，下段：O 状態を $s_t^o = x_{t-1}$ の関係を用いて非表示にしたもの，実線：確率的な依存関係，点線：確定的な依存関係)

提案モデルにおいては，O 状態は過去の全ての出力から観測可能であるが，ここでは簡単のため，O 状態を過去 1 フレーム前の出力のみから観測可能であるとして議論を進める．つまり，$s_t^o = x_{t-1}$ と定める．この関係を用いたときの，PHMM における状態と出力の依存関係を図 1 に示す．

HMM が出力と状態遷移の両方において，状態依存の区分定常過程しか扱えないのに対して，PHMM では，隠れ状態と観測可能な状態の組により，出力のみならず状態遷移においても確率的な現象の記述が可能となる．O 状態を導入することにより，出力と状態遷移共に前出力依存性が実現され，区分定常以上の複雑な過程を扱うことが可能となる．

このように，通常の HMM やセグメント HMM，Δ パラメータを含むモデルが数フレームの出力列，あるいはそれと等価な情報の同時出力確率を扱うのに対して，PHMM は過去の出力に関する条件付きの出力確率を扱う点，更に出力確率のみならず，状態遷移も過去の出力に依存している点が大きく異なる．

2.2 確率計算とパラメータ

PHMM において，出力列 $x_1, x_2, \cdots, x_T$ が観測される確率 $Pr(x_1, x_2, \cdots, x_T)$ を求める．

出力列 $x_1, x_2, \cdots, x_T$ が，H 状態系列 $s_1^h, s_2^h, \cdots, s_T^h$，O 状態系列 $s_1^o, s_2^o, \cdots, s_T^o$ から生起する確率を P_s とおけば，その値は，前述までの議論により次式で表される．

$$\begin{aligned} Ps &= Pr(x_1, x_2, \cdots, x_T, s_1^h, s_2^h, \cdots, s_T^h, s_1^o, s_2^o, \cdots, s_T^o) \\ &= Pr(s_1^h, s_1^o) Pr(x_1|s_1^h, s_1^o) \\ &\quad \cdot \prod_{t=1}^{T-1} Pr(s_{t+1}^h|s_t^h, s_t^o) Pr(x_{t+1}|s_{t+1}^h, s_{t+1}^o) \\ &= Pr(s_1^h, x_0) Pr(x_1|s_1^h, x_0) \\ &\quad \cdot \prod_{t=1}^{T-1} Pr(s_{t+1}^h|s_t^h, x_{t-1}) Pr(x_{t+1}|s_{t+1}^h, x_t) \\ &= Pr(s_1^h) Pr(x_0, x_1|s_1^h) \\ &\quad \cdot \prod_{t=1}^{T-1} \frac{Pr(s_{t+1}^h|s_t^h) Pr(x_{t-1}|s_t^h, s_{t+1}^h)}{Pr(x_{t-1}|s_t^h)} \end{aligned}$$

$$\cdot \frac{Pr(x_t, x_{t+1}|s^h_{t+1})}{Pr(x_t|s^h_{t+1})} \quad (1)$$

求める $Pr(x_1, x_2, \cdots, x_T)$ は，可能性のある全ての H 状態遷移 $s^h_1, s^h_2, \cdots, s^h_T$ の組み合わせに対し，式 (1) を加えればよい．

以上の議論から，PHMM は，次の 5 種類のパラメータで定義できることがわかる．

- $\pi_i = P_r(s^h_1 = S^h_i)$:
  初期時刻に H 状態が S^h_i にある確率
- $a_{ij} = P_r(s^h_{t+1} = S^h_j | s^h_t = S^h_i)$:
  時刻 t に H 状態が S^h_i にあり，次時刻に S^h_j に遷移する確率
- $b_i(x_{t-1}) = P_r(x_{t-1}|s^h_t = S^h_i)$:
  時刻 t に H 状態が S^h_i で，O 状態つまり前時刻出力が x_{t-1} である確率
- $c_{ij}(x_{t-1}) = P_r(x_{t-1}|s^h_t = S^h_i, s^h_{t+1} = S^h_j)$:
  H 状態が時刻 t に S^h_i から S^h_j に遷移する時に，O 状態つまり前時刻の出力が x_{t-1} である確率
- $d_j(x_{t-1}, x_t) = P_r(x_{t-1}, x_t|s^h_t = S^h_j)$:
  時刻 t に H 状態が S^h_j であるときに O 状態つまり前時刻の出力が x_{t-1} で，現時刻の出力が x_t である確率

PHMM のパラメタは，HMM と同様にして EM アルゴリズムあるいはセグメンタル K-means アルゴリズムにより推定され，また，viterbi アルゴリズムやフォワードアルゴリズムにより尤度を計算することができる．

2.3 平滑化部分隠れマルコフモデル

一般に，識別器に用いる確率モデルが与える正解カテゴリに対する尤度は，誤認識が生じている場合でも十分高いことが多い．このため，誤り傾向の異なる二つのモデルを統合することで，誤りを改善することができる．これを拠として，PHMM の出力確率・状態遷移確率と，HMM の出力確率・状態遷移確率を統合した，平滑化部分隠れマルコフモデル (Smoothed PHMM;SPHMM) を提案している [12]．PHMM の状態遷移確率・出力確率を，HMM の状態遷移確率・出力確率との相乗平均で置き換えると，確率計算式 (1) は以下のようになる．

$$\begin{aligned} Ps &= Pr(s^h_1)Pr(x_0, x_1|s^h_1) \\ &\times \prod_{t=1}^{T-1} \left\{ \left(\frac{P_r(s^h_{t+1}|s^h_t)P_r(x_{t-1}|s^h_t, s^h_{t+1})}{P_r(x_{t-1}|s^h_t)} \right)^{w_t} \right. \\ &\qquad \left. \cdot Pr(s^h_{t+1}|s^h_t)^{(1-w_t)} \right\} \\ &\cdot \left\{ \left(\frac{P_r(x_t, x_{t+1}|s^h_{t+1})}{P_r(x_t|s^h_{t+1})} \right)^{w_o} \cdot Pr(x_{t+1}|s^h_{t+1})^{(1-w_o)} \right\} \quad (2) \end{aligned}$$

ここで，w_t は遷移確率の平滑化の重みを表し，w_o は出力確率の平滑化の重みを表している．重みを変化させると表 1 に示すように，様々なモデルを表現することができ，SPHMM は自然な形で種々のモデルを含んだ枠組となっていることがわかる．

表 1　平滑化重みとモデルの関係

w_t	w_o	Model
0	0	HMM
0	1	conditional HMM
1	1	PHMM
$0 \leqq w_t \leqq 1$	$0 \leqq w_o \leqq 1$	SPHMM

文献 [11] において，SPHMM を用いたとき出力確率を平滑化することにより大幅な効果が見られるが，遷移確率については，PHMM が与える確率は十分有効であり，平滑化の効果は見られないことが示されている．

3. 状態・出力依存関係の一般化

PHMM の従来の定式化における隠れ状態と観測可能な状態は，出力確率を条件付けるものに対しても，状態遷移確率を条件付けるものに対しても，共通な組を用いるように定めている．本研究では，この拘束条件を除き，出力確率と状態遷移確率を条件付けるための隠れ状態と観測可能な状態に関して，隠れ状態に関しては共通のものを用い，観測可能な状態については出力を条件付けるものと，状態遷移を条件付けるものとで，異なるものを導入する．

この枠組みを実現するために，前述した従来の枠組みに対し s^{oo}_t，s^{os}_t という 2 つの観測可能な状態を導入する．ここで，s^{oo}_t は出力確率を決定するために用いられ，s^{os}_t は状態遷移確率を決定するために用いられる．この枠組みでの状態・出力依存関係を以下の図 2 に示す．図 2 は，s^{oo}_t が過去 N フレーム前の出力から観測可能，s^{os}_t が過去 1 フレーム前の出力から観測可能である場合の依存関係を示してあるが，この枠組みにおいては，s^{oo}_t と s^{os}_t は過去の全ての出力から観測可能である．PHMM

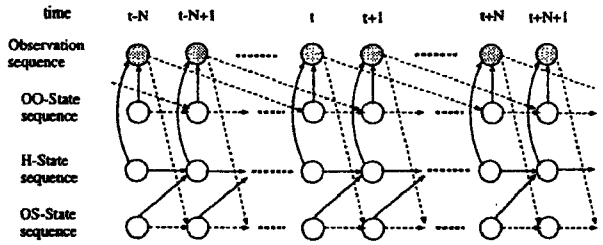


図 2　PHMM における状態・出力依存関係図 (実線：確率的な依存関係を表現，点線：確定的な依存関係を表現)

の従来の定式化は，この枠組みにおいて s^{oo}_t と s^{os}_t が共に 1 フレーム前の出力のみから観測可能である場合に相当する．つまり，出力確率は 1 フレーム前の出力により条件付けられる構造になっている．しかし，音声認識において使用する特徴量の隣接フレーム間における変化は小さく，さらにこの枠組みでは，Δパラメタを付加しない HMM にΔパラメタを付加することと等価であるともみなせる．したがって，この枠組みでは冗長な情報を扱っている可能性がある．

そこで本研究では，特徴パラメータのフレーム間相関として，より長いフレーム間隔を扱うために，s^{oo}_t を N フレーム前の出力である x_{t-N} から観測可能であるとし，s^{os}_t を 1 フレーム

前の出力である x_{t-1} から観測可能であるとする枠組みを採用する．このような，$s_t^{oo} = x_{t-N}$，$s_t^{os} = x_{t-1}$ という関係を用いて図 2 の枠組みを簡略化した場合の PHMM の状態・出力依存関係を図 3 に示す．

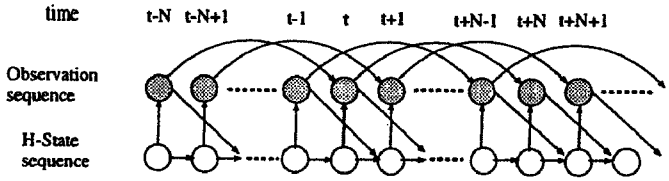


図 3　PHMM の状態・出力依存関係図（$s_t^{oo} = x_{t-N}$, $s_t^{os} = x_{t-1}$ の関係を用いて簡略化したもの）

このとき式 (1) は，

$$P_s \cong P_r(s_1^h)P_r(x_{-N+1}, x_1|s_1^h) \cdot \prod_{t=1}^{T-1} \frac{P_r(s_{t+1}^h|s_t^h)P_r(x_{t-1}|s_t^h, s_{t+1}^h)}{P_r(x_{t-1}|s_t^h)} \cdot \frac{P_r(x_{t-N+1}, x_{t+1}|s_{t+1}^h)}{P_r(x_{t-N+1}|s_{t+1}^h)} \quad (3)$$

と書き直すことができ，従来の PHMM のパラメータセットに，出力確率を規定する次の 2 種類のパラメータが追加されることになる．ここで，$N = 1$ とすると従来の枠組みと等価である．

- $e_i(x_{t-N}) = P_r(x_{t-N}|s_t^h = S_i^h)$
  時刻 t に H 状態が S_t^h で，N 時刻前の出力が x_{t-N} である確率
- $f_j(x_{t-N}, x_t) = P_r(x_{t-N}, x_t|s_t^h = S_j^h)$
  時刻 t に H 状態が S_j^h であるときに N 時刻前の出力が x_{t-N} で，現時刻の出力が x_t である確率

HMM と同様にして，PHMM に対しても以下の手順により Viterbi アルゴリズムを適用できる．与えられた PHMM に対し，時刻 t に H 状態が S_i^h に至る状態遷移のうち，$x_1, x_2, \cdots, x_t$ を出力する確率の最大値を与えるものを考え，そのときの確率値を $\delta(i,t)$，1 時刻前の H 状態へのポインタを $\psi(i,t)$ とする．n_s を H 状態の数とし，N は PHMM のフレーム間相関に用いる時間差とする．また，観測系列の最終フレーム T における H 状態を $s_T^h = S_J^h$ と表す．$\hat{S}$ は，与えられたモデルパラメータによる最適なセグメンテーションの結果を表し，最適状態系列が与えられたもとでの出力確率を P_v とする．

1. 初期化　$(1 \leq i \leq n_s, N = 1, 2, \cdots)$

$$\delta(i,1) = \pi_i f_i(x_{-N+1}, x_1)$$
$$\psi(i,1) = 0$$

2. 繰り返し　$(2 \leq t \leq T, 1 \leq j \leq n_s, N = 1, 2, \cdots)$

$$\delta(j,t) = \max_{1 \leq i \leq n_s} [\delta(i,t-1)\frac{a_{ij}c_{ij}(x_{t-2})}{b_i(x_{t-2})}] \cdot \frac{f_j(x_{t-N}, x_t)}{e_j(x_{t-N})}$$

$$\psi(j,t) = \arg\max_{1 \leq i \leq n_s} [\delta(i,t-1)\frac{a_{ij}c_{ij}(x_{t-2})}{b_i(x_{t-2})}] \cdot \frac{f_j(x_{t-N}, x_t)}{e_j(x_{t-N})}$$

3. 終了

$$P_v = \delta(J,T)$$
$$\hat{s}_T^h = \psi(J,T)$$

4. パス（状態遷移）バックトラック
   $(t = T-1, T-2, \cdots, 1)$

$$\hat{s}_t^h = \psi(\hat{s}_{t+1}^h, t+1),$$

4. 連続音声認識実験

PHMM の性能，および状態と出力の依存関係を一般化することの効果を評価するために，連続音声認識実験を行った．ここでは，まず，音素の切り替わり部分における高精度化を目的として導入した半音節モデルについて述べ，その後，認識性能と出力系列のフレーム間相関に用いる時間差の関係について評価した結果について述べる．

4.1 半音節モデル

PHMM はその構造的な特徴から，ある目標値に様々な初期値から漸近的に収束する現象において，その過渡部を表現するのに優れている [13]．そこで，前音素からの過渡的な遷移を含み，定常状態までをモデリングの単位とする半音節モデル (demi-syllable model) を音素単位として採用する．

図 4 に，音響特徴の時間変化を表したサウンドスペクトログラムと，モノフォン，バイフォン，トライフォン，および半音節の各々の音素系列のラベル境界との対応関係を表す概念図を示す．モノフォンおよびその他のコンテキスト依存モデルの状態系列のラベルも同時に示してある．図には，音素の状態数が 3 である場合を示した．ここで，バイフォンやトライフォンのようなコンテキスト依存モデルは，調音によるスペクトル特性の違いに対処したモデリングであるが，音素ラベルの境界は基本的にモノフォンと同一とみなせる．したがって，モノフォンならびに他のコンテキスト依存モデルは，図 4 に示したような，ラベル情報と音響特徴変化の関係に関する仮定の基では，前音素からの過渡的な現象を正確に記述しきれない可能性がある．半音節モデルは，このような音素の切り替わりにおける過渡的な現象に対する性能を向上することを目的としている．

半音節モデルの初期モデルを学習するために用いる音声データは，モノフォンにおける音素の過渡的な部分を包含するようにラベル付けされるべきである．そこで，まずモノフォンによって状態ごとに強制アライメントを実行する．モノフォンや他のコンテキスト依存モデルにおいて音素の切り替わりに関する過渡的な現象は，図 4 のように音素の最終状態から始まると仮定し，前音素の最終状態からを半音節モデルの単位として，強制的にラベル付けを行った．

4.2 実験条件

学習および認識に用いる音声特徴量を，以下にまとめる．本

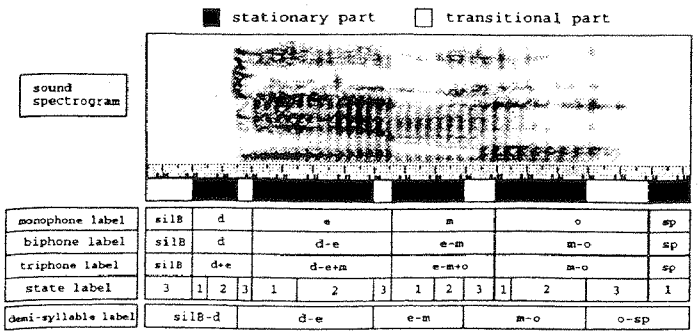


図 4　音響的特徴と半音節モデルのラベル境界の関係

実験では，12 次元の MFCC とパワーを特徴量とした場合 (差分なし特徴；13D) と，それに加えて各々の Δ パラメータを用いた場合 (差分あり特徴；26D) について検討した．ただし，以下で述べる **Tri-HMM-25D** に関しては，26D からパワーを除いたものを用いている．各々フレーム長は 25[ms]，フレーム周期は 10[ms] である．

13D : MFCC + power

25D : MFCC + ΔMFCC + Δpower

26D : 13D + ΔMFCC + Δpower

音響モデルは，ASJ-JNAS，ASJ-PB の男性話者 100 人分程度 (計 20406 文) から学習し，状態数 5，ループ数 3 の left-to-right 型とした．本実験の評価項目となる音響モデルを表 2 に示す．**DS-PHMM-13D** および **DS-PHMM-26D** が各々差分無し特徴 13 次元，差分あり特徴 26 次元を用いて学習した半音節 PHMM であり，**DS-HMM-13D** および **DS-HMM-26D** が各々差分無し特徴 13 次元，差分有り特徴 26 次元を用いて学習した半音節 HMM である．以上のモデルは，学習データおよびモデルの共有等に関して，全く同じの条件で学習を行なったものである．また、**Tri-HMM-25D** は差分あり 25 次元特徴を用いて学習したトライフォンであり，IPA の日本語ディクテーション基本ソフトウェア [14] 提供による音響モデルである．ただし，この **Tri-HMM-25D** は前述の 3 種類のモデルと学習データは同じである．

表 2　評価対象となる音響モデル

モデル	特徴量	混合数	状態数	分散行列
DS-PHMM-13D	13D	1	1200	全共分散
DS-PHMM-26D	26D	1	1200	全共分散
DS-HMM-13D	13D	1	1200	全共分散
DS-HMM-26D	26D	1	1200	全共分散
Tri-HMM-25D	25D	16	2000	対角共分散

ASJ 音声データベースは音素ラベルが付与されていないため，HMM の音素系列を用いて音素ごとに強制アライメントを取り，これにより得られたラベル情報に基づいて，各音素カテゴリごとに学習データを切り出す．そして，切り出された複数観測系列を状態毎にセグメンテーションし，状態内で観測系列の平均化処理を行なうことにより，初期パラメータを荒らく求める．その後，EM アルゴリズムによる再推定を繰り返し行なうことにより，最終的なパラメータを得る．

言語モデルは，語彙数 20000 の辞書を用いて作成した．語彙は，毎日新聞の 1991 年から 1994 年の 4 年分の新聞記事から構成されている．

評価データは，ASJ-JNAS から学習外話者 20 人分 (計 100 文) を用いた．本実験で用いた認識エンジンは、トライグラムと音素を単位とした木構造辞書を用いた，ワンパスフレーム同期ビームサーチを行う．また，木構造辞書の分岐ではファクタリングを行う．

4.3 実験結果

出力系列のフレーム間相関に用いる時間差と認識精度の関係を図 5 に示す．従来の PHMM はフレーム間相関に用いる時間差が 1 の場合に相当する．

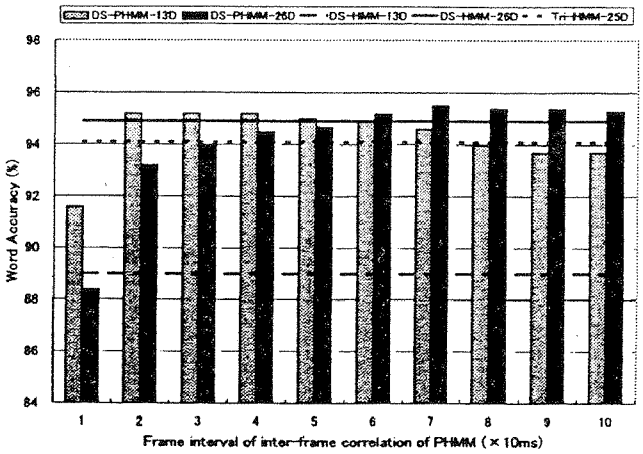


図 5　PHMM におけるフレーム間相関に用いる時間差と単語認識精度の関係

図 5 より，PHMM における状態と出力の依存関係を一般化することにより，認識性能は大幅に向上する．つまり，PHMM において状態遷移を決定するために用いられる出力と，出力を決定するために用いられる出力の組み合わせを同一にすることは，必ずしも適切ではないと言える．

半音節 PHMM は，差分なし特徴を用いた場合 (**DS-PHMM-13D**)，フレーム間相関に用いる時間差を 2 フレームから 6 フレームとしたときに認識精度が良好な値を取り，最大で 95.2% となった．また，差分あり特徴を用いた場合 (**DS-PHMM-26D**) は，フレーム間相関に用いられる時間差を 6 フレームから 10 フレームとしたときに良好な精度が得られ，最大で 95.5% となった．これは，従来の枠組みである，フレーム間相関に用いる時間差が 1 フレームのときの PHMM と比較して，差分なし特徴の場合で 43%，差分あり特徴の場合で 61% の誤認識率を改善したことにあたる．また，**DS-PHMM-13D** は，同じ差分なし特徴を用いた半音節 HMM である **DS-HMM-13D** (89.0%) と比較して 56%，より高次の差分あり特徴を用いた **DS-HMM-26D** (94.9%) に対しても 6% の誤認識率を改善した．さらに，最良な音響モデルとして一般的に用いられているトライフォン HMM (**Tri-HMM-25D**)(94.1%) と比較しても，19% の誤認識率を改善した．これより，半音節 PHMM は，トライフォン HMM と比較してより少ない特徴量，混合数，コンテキストで構成されているにも関わらず，より精度の良いモデル

と言える．同様に，**DS-PHMM-26D** は，**DS-HMM-26D** の誤認識率を 12% 改善し，**Tri-HMM-25D** の誤認識率を 24% 改善した．

次に，2.3 で述べた PHMM における出力確率の平滑化 (SPHMM) を，状態と出力の依存関係を一般化した PHMM において適用することを試みる．

出力のフレーム間相関に用いる時間差を前述の実験で最良の性能を与えたときの値に固定し，平滑化の重み (w_o) を 0 から 1 まで 0.2 刻みで変化させて認識実験を行った．ここでは，フレーム間相関に用いる時間差を，**DS-PHMM-13D** の場合は 3 フレームに，**DS-PHMM-26D** の場合は 7 フレームに固定して評価を行った．結果を図 6 に示す．ここで，重み w_o が 0 のときは出力確率は HMM と等価であり，状態遷移確率が前出力により条件付けられていることのみ HMM と異なる．また，重み w_o が 1 のときは通常の PHMM と等価である．

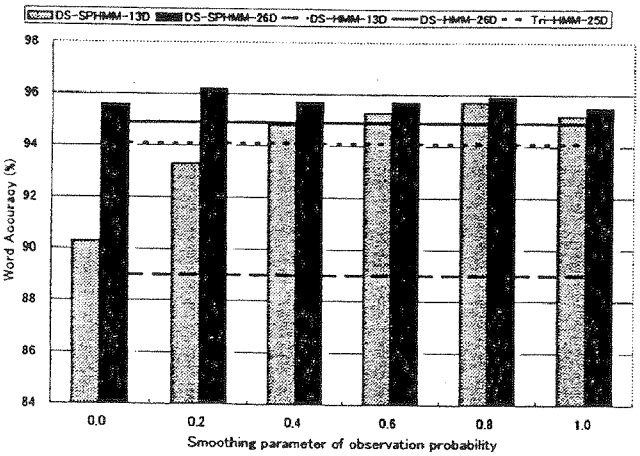


図 6　PHMM における出力確率平滑化の効果

図 6 より，SPHMM は差分なし特徴を用いたとき (**DS-SPHMM-13D**) 最高で 95.7%($w_o = 0.8$)，差分あり特徴を用いたとき (**DS-SPHMM-26D**) 最高で 96.2%($w_o = 0.2$) となり，各々PHMM($w_o = 1.0$) を上回る性能が得られている．これは同じ特徴量を用いた場合の半音節 HMM の結果と比較すると，差分なし特徴を用いた場合で 61%，差分あり特徴を用いた場合で 25% の誤認識率改善にあたる．また，差分あり特徴を用いた場合のトライフォン HMM と比較すると，最大で 25% の誤認識率を改善した．

重み w_o を 0 とした SPHMM，つまり，遷移確率が前出力により条件付けられていることのみ HMM と異なるモデルの認識精度と，純粋な HMM の認識精度を比較すると，差分なし特徴，差分あり特徴によらず，$w_o = 0$ とした SPHMM の方が良好な性能を与えている．これより，遷移確率を過去の出力の条件付き確率とすることの効果がうかがえる．

5. まとめ

部分隠れマルコフモデルにおける状態と出力の依存関係について検討し，連続音声認識実験によって評価を行った．また，音素の渡り部分の揺らぎを吸収することを目的に，半音節モデルを導入した．この半音節 PHMM において出力のフレーム間相関の扱いを一般化することにより，従来の半音節 PHMM に対しては差分なし特徴を用いた場合で 43%，差分あり特徴を用いた場合で 61% の誤認識率を改善し，その有効性が示された．また，同じ次元数の特徴量を用いた半音節 HMM に対しては，差分なし特徴を用いた場合で 56%，差分あり特徴を用いた場合で 12% の誤認識率を改善し，最良な音響モデルとして現在一般的に用いられているトライフォン HMM をも上回る程度にまで性能を改善することができ，最大で 24% の誤認識率を改善するに至った．それに加え，出力確率の平滑化を併用することにより，さらに認識性能を改善することができた．

この結果を受け，今後は自然発話音声に対する PHMM の性能を評価することを考えている．これは，自然発話音声のような複雑な現象に対する PHMM の表現能力を調べることを目的としたものである．発話速度が読み上げ音声よりも速い自然発話に対処するには，出力が時間的に過去の出力のみから影響を受けるという枠組みから，時間的に先の出力からの依存関係も同時に扱えるような枠組みとすることが望ましいと考えている．

文　献

[1] 坪香 英一, 中橋 順一, "音声スペクトルの動的特徴を組み込んだ HMM, " 信学論 (A), vol.J77-A, no.2, pp.162-172, Feb. 1994.

[2] L. Deng, M. Aksmanovic, "Speaker-Independent phonetic classification using Hidden Markov Models with mixture of trend functions, " IEEE trans on Speech and Audio Processing, vol. 5, pp.319-324, JULY. 1997.

[3] C.J. Wellekens, "Explicit correlation in Hidden Markov Model with optimal inter-frame dependence, " Proc. ICASSP87, pp.383-386, 1987.

[4] Mari Ostendorf, "From HMMS to segment models : stochastic modeling for CSR, " Automatic speech and speaker recognition, Kluwer Academic Publishers, pp.185-210, 1996.

[5] Y. Ariki, "Mixture density HMMs with two-level transition, " Journal of Acoustic Society Japan(E), vol.14, no.4, pp.279-280, Sep. 1993.

[6] 高橋敏, 松岡達雄, 南泰浩, 鹿野清宏, "フレーム間相関を利用した音韻 HMM による音声認識, " 信学論 (A), vol.J77-A, no.2, pp.153-161, Feb. 1994.

[7] 中川 聖一, 山本 一公, "セグメント統計量を用いた隠れマルコフモデルによる音声認識, " 信学論 (D-II), vol.J79-D-II, no.12, pp.2032-2038, Dec. 1996.

[8] J-F. Mari, J-P. Haton, "Automatic word recognition based on second-oder hidden Moarkov models, " IEEE Trans. on Speech and Audio Process, vol.5, n.1, Jan. 1997.

[9] J.A.Bilmes, "Buried Markov Models for speech recognition," ICASSP99, vol.2, pp713-716, Mar. 1999.

[10] 益満健, 小林哲則, "部分隠れマルコフモデルとそのジェスチャ認識への応用, " 情処論, Vol.41, No.11, pp.3060-3069, 2000.

[11] 古山純子, 小林哲則, "部分隠れマルコフモデルによる単語音声認識," 信学論 (D-II), **J83-D-II**, No.11, pp.2379-2387, Nov. 2000.

[12] 古山純子, 田中英智, 益満健, 小林哲則, "部分隠れマルコフモデルによる音声認識," 日本音響学会秋期講演論文集, pp.93-94, 1998.

[13] T.Kobayashi, J.Furuyama, K.Masumitsu., "Partly Hidden Markov Model and its Application to Speech Recognition," Proc. ICASSP99, vol.1,pp.121-124, Mar. 1999.

[14] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄, "IT TEXT 音声認識システム," オーム社, 2001.

社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE
PRMU2002-118, HIP2002-25(2002-11)

合意形成型対話における声・顔の機能的表情の分析と認識

松坂 要佐† 江尻 康† 小林 哲則†

† 早稲田大学 理工学部，東京都
E-mail: †{yosuke,ejiri,koba}@tk.elec.waseda.ac.jp

あらまし 人間同士の合意形成型の対話を対象として、頭部に現れる動作パターンと声に現れるF0パターンの分析と認識を行った。分析の結果、談話の半数は、頭部の動作と共起していた。頭部うなずきの動作パタンにおいては、動作の速度・強さと、それと共起する発話の信念の程度との間に相関が見られた。音声のF0パターンにおいては、言明・応答の発話において、ピッチの強い高下が見られた。それぞれの特徴を用いて、談話の分類を推定する認識器を作成し、対話における表情に基づくパターン認識器の可能性について検討した。

キーワード 合意形成型対話，表情，対面対話，マルチモーダルシステム

Analysis and Recognition of the Functional Expression of Voice and Face in Agreement Formation Dialog

Yosuke MATSUSAKA†, Yasushi EJIRI†, and Tetsunori KOBAYASHI†

† School of Science and Engineering, Waseda University, Okubo 3–4–1, Shinjuku-ku, Tokyo, 169–8555 Japan
E-mail: †{yosuke,ejiri,koba}@tk.elec.waseda.ac.jp

Abstract This paper describes an analysis and recognition of the head movement, and F0 pattern for human's agreement formation dialog. The half of a discourse coincided with a head movement as a result of analysis. In the pattern of a head nodding motion, correlation was seen between grade of the belief and the speed and strength of the motion. In the F0 pattern, strong fluctuations of pitch were seen in the utterance of declaration and response. The recognition system which presumes the classification of a discourse was created using each feature, and the possibility of a pattern recognition system for the expression in dialog was examined.

Key words Agreement formation dialog, Expression, Face-to-face conversation, Multi-modal system

1. はじめに

従来の音声対話システムにおいて扱われていたのは、主にタスク指向型の対話であった。人間の話し相手となって、蓄えた情報を元に議論に参加し意思決定に貢献するシステムを考えた場合、議論を伴う合意形成型の対話に参与する能力が必要となる。我々は、合意形成型の対話を扱うことのできる対話システムを作ることを目指している。

タスク指向型の対話システムは、予約システムや情報検索システムなどの分野で大きな成功を収めているが、意思決定の場面で利用されることはまれである。これは、タスク指向型の対話システムが、人間の合意形成時における諸現象を扱うように設計されておらず、それに対処することができないことが大きな原因である。

人間同士の合意形成プロセスをモデル化し、それをコンピュータシステムによって支援する試みも行われているが、それらシステムにおいては、支援を受けるためには、人間がコンピュータの提供するインターフェースを通して対話を行う必要がある。人間同士の自然な対話を妨げないで、それらの支援をすることを考えた場合、声・顔の表現によって行われるその対話を音声・画像処理によって認識する技術が不可欠となる。

本稿においては、合意形成型の対話システム構築へ向けて、合意形成型対話をそれに現れる声・顔の表情の面から特徴分析し、人間同士の対面対話の分析に基づいた認識システムを構築する。

2. 合意形成型対話の特徴分析

合意形成型対話には、以下のような特徴がある

2.1 信念と表情

各々の話者は、その時の話題に対して、その時々の信念で対

処する。信念は、その個人のそれまでの経験により形成される。各個人によって持つ経験が異なることから、信念は各個人に帰属されるものとなる。合意形成型対話とは、同一の話題に対して各個人の持つ信念を、相互に意見を投げかけあうことですり合わせるための作業ということができる。

各個人の持つ経験を限界のある発話量によって相手に伝達することには、一般に困難が伴う。足りない発話によって相手の信念を動かすためには、自らの強い信念を相手に提示することが時に必要とされる。人間同士の対面対話においては、話者の信念の強さの程度が、声・顔の表情によって相手に提示される。相互の信念を提示しあう必要のある合意形成型の対話は、時として非常に表情豊かなものとなる。

2.2 対話に表れる表情の機能的な役割

内部にある信念と現在行われている議論との間には、多くの場合ギャップが存在する。そのギャップが感情となる。表情とは、そのギャップの外部への表出として説明される。

表情を、人の内部に持つ感情が社会的な要因によって抑制された形で露出したものとして見る場合が多いが、本研究では、それをコントロールされ、戦略的に用いられるものとして捉える。社会的能力を獲得した成人であるほど、内部に持つ感情に依存しない形で表情をコントロールでき、感情を抑制した形としてだけでなく、能動的に強調して表出する場合の両方について行うことができると仮定する。

上記仮定に立つと、表情は、相互の信念を伝達し、それを調整するための重要な機能を持つものとなる。本研究では、表情を従来的な解釈における感情の露出としてだけではなく、信念のすりあわせを行うために用いられる機能的な役割を持つものとして扱う。

2.3 まとめと本稿におけるアプローチ

タスク指向型対話においては、タスクの目標は最初から具体化されており、誤解の解消や不足データの聞きだしが問題となる。それに対して、合意形成型の対話の場合は、誤解の解消や不足データの聞き出しを行うとともに、発せられた発話を自らの信念に照らし合わせた上で、その妥当性の判断や、自らの新たな信念の形成が行われる。

タスク指向型対話における誤解は、相互にやり取りされた情報の差異に起因するものであり、多くの場合、数回のやり取りによって円満に解決される。それに対して、合意形成型対話における誤解は、信念の差異に起因する場合がある。信念の差異に起因する誤解は、解消に合意形成の過程を必要とするため、解消がすぐには進まず、長期にわたってすり合わせのためのやり取りが行われる場合が多い。

合意形成の過程において、対話参加者は相互に信念の強さを提示しあう必要があり、その提示は表情によって行われる。表情は、相互の信念を調整する役割を持つ。本研究においては、それが発話者のコントロールの下で機能的に用いられているとの立場に立つ。

合意形成型対話に表れる表情を認識・理解することは、合意形成の過程を理解する上で重要である。本稿では、合意形成型の対話において使われる機能的表情の理解を行うために、

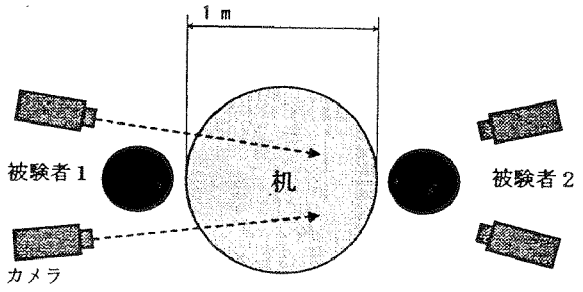


図 1 収録環境
Fig. 1 Data recording setup

（1） 人間同士の合意形成型対話の収録
（2） 機能に基づいた発話の分類
（3） 発話の特徴の抽出と分析
（4） 機能的表情の分類を行う認識器の実現と評価

を行う。

3. 分析方法

3.1 合意形成型対話の収録

人間同士の合意形成型対話を収録するために、名前付け課題を用意した。

名前付け課題においては、出題者が提示したおもちゃに対して、それに適した名前を対話参加者が決定する。対話参加者は、出題者からの要求に適していると考えられる名前を発案、議論し、限られた時間内に参加者間での合意を形成する必要がある。今回は、出題者からの要求は「売り上げの高くなりそうな名前」であり、議論の時間は 20 分として収録を行った。

2 組のカメラ・マイクを用いて、対話を行っている話者の音声・画像を同期して収録する [6]。2 人の被験者に対し，頭部をクローズアップしたカメラと上半身を撮るカメラを図 1 のように設置した．計 4 台のカメラの映像を 4 分割器を通し S-VHS ビデオテープに録画した（図 2）。二人の音声はピンマイクを用いてそれぞれの音声を左右に分けて収録した。対面状況で収録した 2 話者の音声は、相互のマイクに回り込んで収録されるため、収録後に信号処理によって分離した [7]。今回の収録では、このようにして 20 歳代前半の 9 人分（男：7 人，女：2 人）、計 170 分間のデータを得た。

今回は、収録した中から活発な議論の行われている 24 分のデータを分析対象とした。議論の内容の転記の一部を図 3 に示す。

3.2 談話タグの検討

対話の分析を行うためには、まず、そのときの発話がどのような意味を持つ発話であるかを分類する必要がある。発話を分類する方法は多数考えられるが、各々の発話の持つ機能に着目して分類する方法が一般的である。

日本語のタスク指向型の対話を分類する体系として、標準談話タグが提案されている [1]。しかしながら、このタグ体系を議論を伴う対話に適用すると、現時点では、特定の機能タグに偏ってタグが振られるなどの問題があることは、文献 [2] の指摘するとおりである。

図 2　録画像の例
Fig. 2　Example of recorded image

A:そうだね、じゃあ、車、車だよねー、カーをつける？
B:いや、それはないでしょ？
A:ないか、むずかしっすねー、なんかネーミングでうるってのはね、
B:そうだねー、
A:これ車だよね、だけど
B:うん
A:なんだろう、いや、車ってのはね、あんまりこうないほうがいいかもね
B:ああ
A:うん、逆に流行ったのミニ四駆ってのがあるじゃん、それでー、ハイパーヨーヨーでもさ、
B:うん
A:わかるじゃん、遊び方が
B:うん
A:ようよう、ミニ四駆も四駆だから車、だからその名前でどう遊ぶか、あ、でもポケモンは違うのかでも、遊戯王も違うか。いやー、そおか
B:うん

図 3　収録された対話の転記 (一部)
Fig. 3　Transcription of the collected dialog(part).

議論における発言を機能的側面から分類しようという試みに、トゥールミンモデル [3] などがあるが、これらの分類は、よく整った (フォーマルな) 対話においては十分な説明を与えることができるものの、インフォーマルな形で進められる対話においては適用が難しかった。インフォーマルな形で行われた議論を伴う対話にも適用可能なタグ体系として DAMSL タグ [4] が検討されている。DAMSL タグにおいては、タグ付けの体系を話題遷移の層、談話のやり取りの層、対話の調整の層などに階層化する方法を用いている。

本稿においては、インフォーマルな対話を扱うための階層化を行いつつも、日本語標準化タグ体系においてタグ付けの難しかった分類項目のいくつかを再検討し、下記のようなタグ体系を用いて発話の分類を行った。

話題、態度、談話、調整の階層に階層化したタグ体系を用いる。話題の階層においては、そのときに議論されている話題を

1[独り言] そうだね、
2[-]{D じゃあ }、
3[その他の働き掛け] 車、
4[確認] 車だよねー、
5[提案] カーをつける？
6[否定 5] いや、
7[強い主張] それはないでしょ？
8[合意 7] ないか、
9[弱い主張] むずかしっすねー、
10[-]{F なんか }
11[弱い主張] ネーミングでうるってのはね、
12[独り言] そうだねー、
13[確認] これ車だよね、
14[-]{C だけど }
15[合意 13] うん
16[独り言] なんだろう
17[-]{F いや }、
18[弱い主張] 車ってのはね、あんまりこうないほうがいいかもね
19[部分合意 18] ああ

図 4　談話タグの例
Fig. 4　Example of the discourse tags.

記述する。態度の階層においては、そのときに議論されている話題に対する、各々の参加者の興味の状態を記述する。話題に対する興味の度合いを 5 段階の覚醒度として、また、快・不快の度合いを 5 段階の評定であらわす。

談話の階層においては、標準談話タグに基づく発話の分類を記述する。議論を伴う対話を扱うために、標準談話タグに加えて、信念の表明を行う主張・言明タグ、表明された信念に対する応答・意見のタグを新たに定義した。主張・言明タグには、発話の中で、論拠に客観性の乏しい発話をこれに分類する。信念の程度に応じて、強い主張 (Strongly Assert)、弱い主張 (Weakly Assert)、独り言 (Self Talk) の 3 種類を用意した。応答・意見タグには、発話の中で、論拠に客観性の乏しい応答発話をこれに分類する。信念の程度に応じて、否定 (Reject)、部分否定 (Reject Part)、部分合意 (Accept Part)、合意 (Accept) の 4 種類を用意した。

調整の階層においては、発話権の調整を行うための発話を扱う。相手発話の (表層的な) 受理と更なる発話の継続を許可する働きを持つ相槌、自己の発話権の維持に貢献するフィラーがこれに該当する。

タグ付けは、訓練を受けたオペレータによって行われた。その一部を図 4 に示す。

4. 表情の分析

対話中に人間は相互に様々な表情の提示を行う。対面状況においては、それら表情は主として、頭部の動作と発話の韻律の中に現れる。それら表情の中から、いくつかの特徴を選び出して分析・認識対象とした。

4.1 頭部に表れる表情の分析

対話中において人間は、主に眉間や口元の変形、頭部全体の

表 1　談話タグと頭部動作の共起関係

Table 1　Co-occorrence between discourse tags and head motion events.

	うなずき	かしげ	首振り	合計	タグの合計
依頼				0	1
勧誘	5			5	8
提案	12			12	25
確認	10			10	35
真偽情報要求				0	1
意見の要求	6			6	17
情報要求	2			2	3
その他の働き掛け	2			2	7
希望				0	2
情報伝達	1			1	10
主張	21	6		27	69
独り言	1	1		2	33
その他の言明	1			1	9
合意	15			15	33
否定		3	3	6	8
了解	3			3	5
未知情報応答			1	1	7
その他の応答	3			3	16

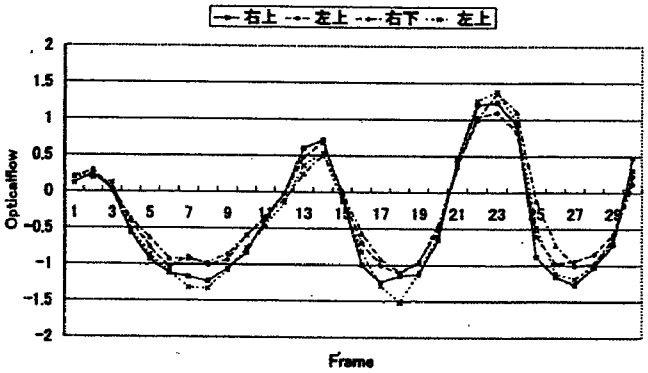


図 5　典型的なうなづきのオプティカルフロー

Fig. 5　Optical flow feature of the nodding motion

動作として表情を表出する。本稿においては、頭部に表れる表情の中から頭部全体の動作を特徴として注目した。頭部動作をうなずき、かしげ、首振りの 3 種類に分類した。

各々の談話の分類における頭部動作の共起関係を表 1 に示す。特に相手に意見を求める発話や応答を行う発話において、生起する動作の種類の特徴が見られるが、多くの発話はその発話の中にうなづきを伴っていることがわかる。

頭部動作は、頭部領域を 4 分割し、それぞれの領域においてオプティカルフローによる特徴抽出を行った。うなづきにおけるオプティカルフローの変動を図 5 に示す。

同一の談話の分類の中で、信念の程度の差が存在する場合におけるオプティカルフロー特徴量の比較を行った。オプティカルフローのゼロクロス数と振幅を特徴として、合意と部分合意と共起したうなづきを分析した結果を図 6 に示す。信念の程度によって、同一分類の中でも頭部動作の強度や頻度の差が存在する。

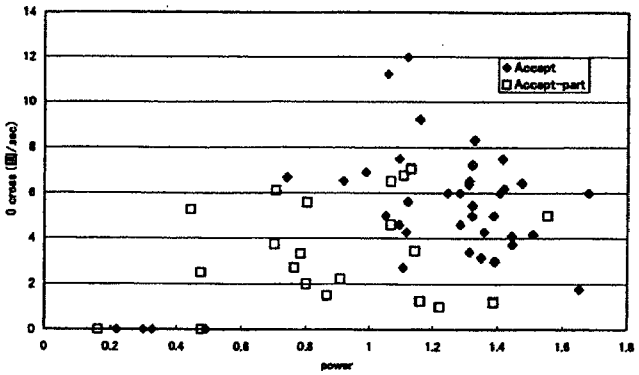


図 6　うなづきの速さ・強度と談話における信念の程度の関係

Fig. 6　Relation between nodding motion and a belief in discourse

4.2　発話の韻律に表れる表情の分析

発話の韻律的特長としては、音の大きさ・話速・F0 があるが、本稿では発話における F0 の変化に注目して分析・認識を行う。

発話区間の切り出し方法としては、談話単位で切り出した。F0 パターンによる分析を考えた場合、アクセント句によって区切る方法も有効であると考えられるが、収録したデータの多くの発話において、アクセント句が不鮮明であるため、この方法をとらなかった。

発話の F0 パターンの抽出には、Snack2.2 を用いた。発話中の各部位 (今回は 5 箇所とした) における F0 の高さ、発話の平均 F0、発話の長さ、一発話内の F0 の分散を発話の持つ特徴として注目した。

談話分類に対する、各部位における発話の F0 の平均と分散を図 7 に示す。今回新設した言明・応答の分類に該当する発話を、従来の談話の分類に該当する発話と比較すると、新設した分類の発話は、全体として F0 の平均が高く、分散が大きい傾向がある。特に否定を表す談話標識において、特に急峻な F0 の上昇が見られる。

5.　表情の認識

5.1　頭部に表れる表情の認識

前章で述べた特徴量を正規分布によってモデル化する。特徴量としては、

（1）　動作区間内でのゼロクロス数

（2）　オプティカルフローの振幅

（3）　頭部動作の発話区間内での位置

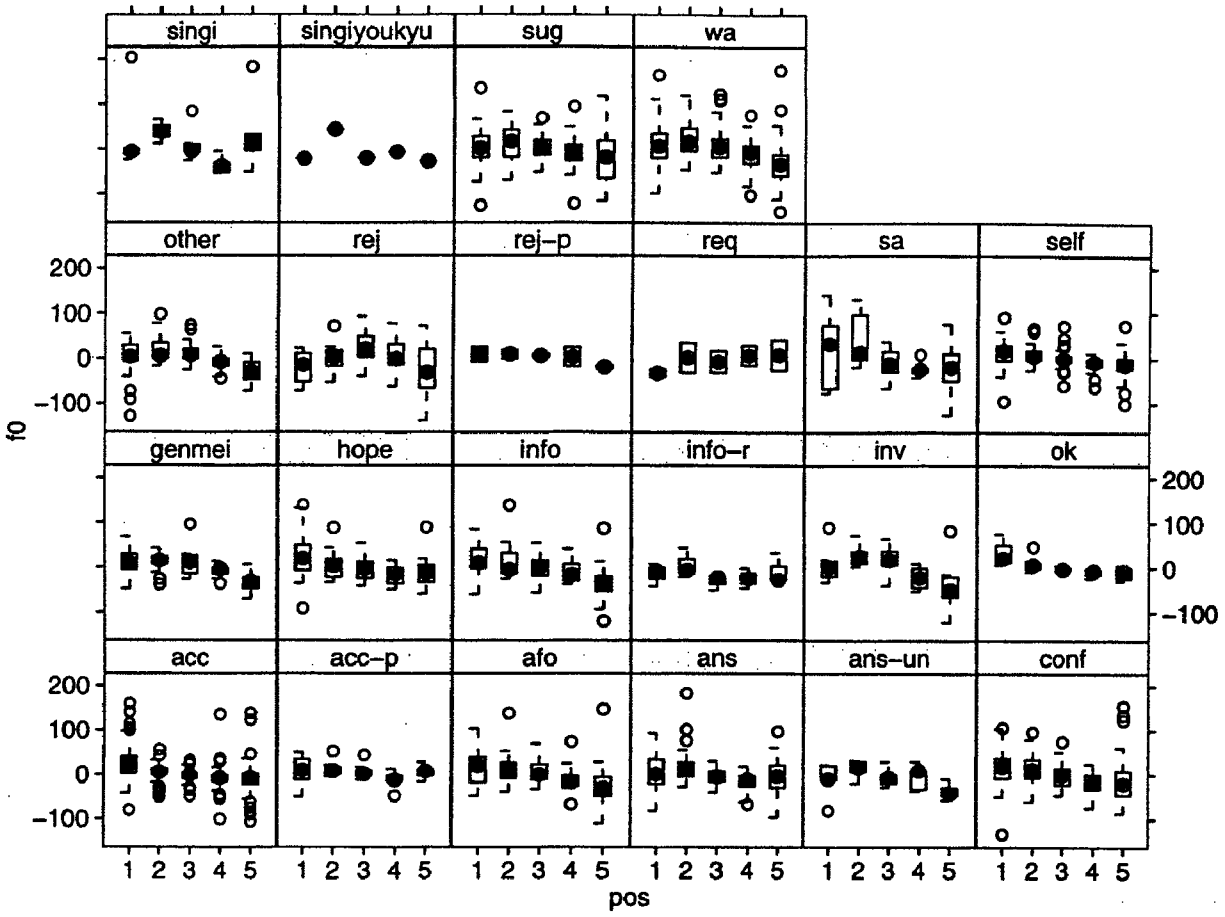
（4）　動作の継続時間

を用いた。

各頭部動作の検出と分類は、HMM を用いたスポッタによって行われる [8]。この検出結果に基づいて上記の特徴量を抽出し、頭部動作内での特徴をと談話の分類との共起関係の事前確率に基づいて、談話の分類を推定した。今回は、談話の種類を対話の中に頻出する 4 種類を選択して認識実験を行った。その結果を表 2 に示す。

5.2　発話の韻律に表れる表情の認識

前章で述べた特徴量を正規分布によってモデル化する。この



図中の省略は以下のとおり。依頼 (req)、勧誘 (inv)、提案 (sug)、確認 (conf)、真偽情報要求 (singiyoukyu)、意見の要求 (afo)、情報要求 (info-r)、その他の働き掛け (other)、希望 (hope)、情報伝達 (info)、真偽情報応答 (singi)、強い主張 (sa)、弱い主張 (wa)、独り言 (self)、その他の言明 (genmei)、主張の合意 (acc)、主張の部分合意 (acc-p)、主張の否定 (rej)、主張の部分否定 (rej-p)、未知情報応答 (ans-un)、了解 (ok)、その他の応答 (ans)

図 7　談話の種類とピッチパターンの分布

Fig. 7　Relation between pitch pattern and a discourse

表 2　頭部に現れる表情を用いた談話の認識結果

Table 2　Result of classification of discourse using facial expression

	認識結果					
	主張	合意	否定	確認	Total	%result
主張	5	1	0	7	13	38.4
合意	0	23	0	0	23	100.0
否定	0	0	1	0	1	100.0
確認	0	0	0	2	2	100.0

表 3　声に現れる表情を用いた談話の認識結果

Table 3　Result of classification of discourse using expression in prosody

	認識結果				
	主張	合意	確認	Total	%result
主張	78	25	36	139	56.1
合意	10	87	14	111	78.4
確認	20	19	12	51	23.5

モデルを用いて、出現頻度の高い 3 種類の談話を分類した結果を表 3 に示す。

6. 議　　論

本稿では、対話の過程において現れる、頭部の動作と発話の F0 に現れるパターンを分析し、それに基づく認識器を構成した。

綿貫らは、うなずきの頻度や発話の平均 F0 が対話の盛り上がりによって、高くなる傾向を持つことを示した [5]。本稿では、局所的なインタラクションにおける表情を分析・認識対象としたが、対話全体における長期の変化も視野に入れる必要がある。

また、今回注目しなかったことに発話の内容がある。発話内容を手がかりとして信念の程度を推定することも考えられる。しかしながら、対話における多くの発話は小さなフラグメントであり、信念の程度を表す語 (「絶対」「多分」「きっと」「間違いなく」など) は、発話によって出現する場合としない場合がある。言語のみに頼った分類を考えた場合、その情報の少なさから、高い性能を得られるとは考えにくい。しかしながら、それらを補助的な情報として用いることで性能改善に貢献する可

能性はある。その点については、今後の検討課題としたい。

7. まとめ

合意形成型の対話を対象として、頭部に現れる動作パターンと声に現れる F0 パターンの分析と認識を行った。

分析の結果、談話の半数は、頭部の動作と共起していた。頭部うなずきの動作パタンにおいては、動作の速度・強さと応答発話の信念の程度との間に相関が見られた。音声の F0 パターンにおいては、言明・応答の発話において、ピッチの強い高下が見られた。

それぞれの特徴を用いて、談話の分類を推定する認識器を作成し、対話における表情に基づくパターン認識器の可能性について検討した。

文　献

[1] 荒木雅弘, 伊藤敏彦, 熊谷智子, 石崎雅人, "発話単位タグ標準化案の作成," 人工知能学会誌, Vol.14 No.2, pp.251-260, 1999.

[2] 矢野博之, 伊藤昭, "同意・不同意表現のための対話タグに関する一考察," 人工知能学会誌, Vol.14 No.2, pp.290-295, 1999.

[3] 足立幸男, 議論の論理, 木鐸社, 1997.

[4] James Allen, Mark Core, "Draft of DAMSL: Dialog Act Markup in Several Layers, " http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/ RevisedManual.html

[5] 綿貫啓子, 外川文雄, "対話における感情の変化の解析," 情報処理学会研究報告 SLP 7-13, pp.79-84, 1995.

[6] 松坂要佐, 小田勇一郎, 小林哲則, "グループ会話におけるマルチモーダル会話データの収録・解析システムの構築," 第 15 回人工知能学会全国大会 3B2-02, 2001.

[7] 松坂要佐, 小林哲則, "チュートリアル:対面対話の収録と解析のための信号処理とパターン認識," 人工知能学会研究会資料 SIG-SLUD-A201, pp.27-32, June, 2002

[8] 江尻康, 松坂要佐, 小林哲則, "対話中における頭部ジェスチャの認識," 電子情報通信学会技術研究報告 PRMU2002-61, pp31-36, 2002.

人物動作における感情表出モデル構築に向けたディジタルアクタ学習システムの設計

Construction of Digital Actor Learning System for Building Emotive Model of Human Behavior

大隈 健史† 長船 大毅† 森野 岳宏† 村上 真‡ 白井 克彦†
Takeshi OKUMA†, Daiki OSAFUNE†, Takehiro MORINO†
Makoto MURAKAMI‡, Katsuhiko SHIRAI†

†早稲田大学理工学部 ‡東洋大学工学部
†Waseda University, ‡Toyo University
E-mail: okuma@shirai.info.waseda.ac.jp

Abstract

人間が感情を動作に表出させる感情表出モデルを構築することを目的として，ディジタルアクタシステムに対する演技指導を通してモデル化を行なう手法を提案する．本システムにおいてはディジタルアクタに対する演技指導を通して人間の中にある感情表出のイメージと，アクタの動作を近づけようとする．この演技指導の過程に用いられた感性語などの情報と動作の特徴との対応を表現する感情空間は，人間の動作による感情表出プロセスをモデル化したものだと言える．本研究では対象領域を歩行動作に限定し，感情表出モデルを構築する．本稿では，ディジタルアクタシステムの概要と，そのために行なった演技指導の分析について報告する．

1 はじめに

近年，コンピュータの処理能力の向上により，民生用機材での高品質な3DCGキャラクターアニメーション制作の可能性が開けてきた．しかし，従来から人間の動作の記述に利用されてきたモーションキャプチャは，機材の設置場所の確保やコストの面から個人で利用できるものではない．また，個人で利用できる動作生成の手法として，キーフレームアニメーション，数値入力やプログラミングによるアニメーションなどが考えられる．しかし，それらの手法は，利用にあたり熟練した技能が必要であることから，誰しもが利用することは困難である．

そこで，理想的な動作生成の手法として，インタラクションを通じたキャラクターアニメーション制作 [1][6] が考えられている．この手法では，CGキャラクター（ディジタルアクタ）の動作に対し，制作者が自分のイメージと違う部分をアクタに演技指導を行ない，アクタはその演技指導を理解して自らの動作を修正し，ユーザに再度提示する．このやり取りを繰り返すことで，制作者の頭にあるイメージに近い動作が生成される．つまり，制作者がアクタと対話をすることで，アクタが制作者の考えていることを推測し，推測した内容を反映した表現を行うことにより，アクタ自身が自律的に動作を行うのである．このことを可能にするためには，演技指導で用いられる感性語などの情報と動作の特徴との対応を表す感情表出モデルを構築する必要がある．

また，人物動作を認識することができれば，感情表出モデルを用いることで感情を読み取ることが可能となり，将来的にはロボットに代表される機械と人間との豊かなコミュニケーションに寄与できるものと考える．本稿ではディジタルアクタシステムの概要と，そのために現時点までで行なった分析について報告する．

2 ディジタルアクタシステムの概要

基本動作制御モデル，感情表出モデルという2つのモデルを通して，アクタが豊かな感情表出を伴った動作を生成するシステムを構築する．それぞれの役割と定義は以下の通りである．

人間型の身体を持つアクタを用いて感情表現をする際には全く自由に動いても良いというわけではなく，物理的，運動学的に妥当な動作を生成するための拘束条件なども考慮に入れなければならない．そこで本システムにおいては，ある特定の動作における拘束条件を記述した基本動作制御モデルと，そこに感情を付与す

る感情表出モデルの2つのモデルを用いて動作による感情表出の実現を目指す．つまり，人間のある特定の動作（例えば「歩く」「走る」）を生成する基本動作制御モデルとその動作に対して演技を加える（「悲しそうに…」「うれしそうに…」）感情表出モデルを別々に構築することになる．

2.1 基本動作制御モデル

基本動作制御モデルとは，感情を反映させる元となる動作を定義したモデルであり，動作における関節の連動性のルールセットにより成り立つ．全ての人物動作というのは，個々の関節を連動させて制御される．故に，個々の関節の連動性を動作ごとにルール化し，そのルールを積み重ねて構築されたルールセットによって，動作の制御が可能となる（図1参照）．

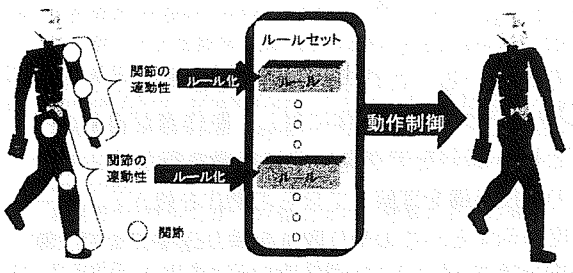


図1　基本動作制御モデル概念図

本研究で構築する基本動作制御モデルは，個性や感情を反映する幅を持ったルールセットでなくてはならない．そこで，モーションキャプチャによって収録した多人数の動作データを用いて動作をモデル化することで，動作者の個性を吸収した汎用的なモデルを構築する．

2.2 感情表出モデル

本システムにおいて感情表出モデルは，人間とのインタラクションを通してある特定の動作から，感情を持つ動作を作成することにより構築される．つまり，感情表出モデルは初期状態では，感情を持つ動作を作り出すためのパラメータを持っておらず，学習によってのみ取得することになる．何故なら感情を伴う動作を予め分類しておくことは困難だからである．これにより，感情を持った動作を恣意的に分類するのではなく，連続的なものとして獲得することができる．

3 演技指導の分析

本システムにおいては，演技指導の内容と出力動作の対応を蓄積し，モデルを構築していく．よって，感情語を用いた明確な演技指導を引き出すために，効果的なインタフェースを構築することを目標とし，演技指導の分析を行なった．

3.1 主観的な動作と客観的な動作の比較

ディジタルアクタは感情表出に関する知識が充分ではない状態だと，実際には近い感情を探したり既知の情報を組み合わせることで動作を生成する．こうして得た動作を学習することで，モデルを拡張していくことが可能となる．その際に感情語から想像される動作のイメージと，そこから動作に表出させている特徴，そして感情表出モデルの修正方法などを規定するためには，人間が実際に行なっているイメージの動作への表出法や修正法を明らかにする必要がある．そこで，図2のように主観的な動作と客観的な動作を比較する．

ここで主観的な動作とは，実際にモーションキャプチャで収録した動作であり，頭の中にあるイメージを元に生成された動作である．しかし，この時動作者は自分の動作を客観的に捉えることができないため，イメージと一致していない可能性がある．一方，客観的な動作とは，本システムのプロトタイプにより演技指導を通して獲得した動作である．この両者を比較することで，人間の思考プロセス及び意思の反映の仕方を捉えることが可能ではないかと考える．

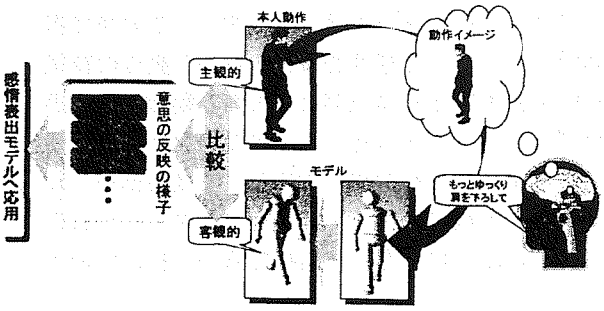


図2　主観的な動作と客観的な動作の比較

3.2 主観的動作の収録

赤外線式モーションキャプチャシステム ReActor を用いて，大学生男子35人の歩行動作について収録を行なう．収録の際，以下に示すように，速さ，大きさ，姿勢，感情を示す言葉を用いて指示を出す．

- 普通に歩く
- ゆっくり歩く
- 速く歩く
- 大きく手を振って歩く
- 大股で歩く
- 細かく歩く
- 胸を張って歩く
- うつむいて歩く

- 楽しそうに歩く
- 悲しそうに歩く
- 怒っているように歩く

この収録によって得た動作を1周期，つまり2歩分を切り出して，それぞれループ化を行なう．こうして得られた主観的動作と以下に示すような手法で合成した客観的動作を比較する．

3.3 客観的動作の作成

主観的動作として収録した「普通に歩く」という動作から，図3に示す3通りの方法で客観的動作を作成する．

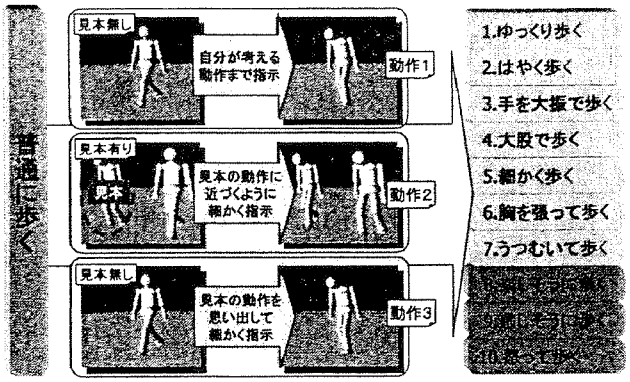


図3　動作合成

動作1は，頭の中にある動作のイメージとモデルの動作が一致するように作成する．動作1を主観的動作と比較することで，イメージとして客観的に持っている特徴と主観的に動作へ表出させている特徴との差異を求めることができる．動作2は，演技指導を行なう人間の主観的動作を見本動作として提示し，作成する．動作2を主観的動作と比較することで，動作を視覚的に提示されたときに認識できる特徴，またそこから指示として表現できる特徴がわかる．動作3は，動作2を作成することで修正された頭の中のイメージを元に作成する．動作3を動作1および主観的動作と比較することで，修正されたイメージと実際の動作との差異を求めることができ，人間の頭の中でのイメージの修正のされ方を確認することができる．

実験は主観的動作を収録した35人に対してWizard Of OZ方式[8][9]で行なった．被験者には図4に示すシステムを用い，指示をもとに動作を変化させることで，上述の3通りの方法で図3に示す10種類の動作を作成してもらった．

モデルの動作及び指示内容は表示用のアプリケーション「IVA Logger」(図5)により記録される．記録された作成動作はBVH形式のデータで与えられる．そして，専用分析アプリケーションを用いて作成動作の分析を行い，分析結果を統計ソフト(JMP)で集計した．

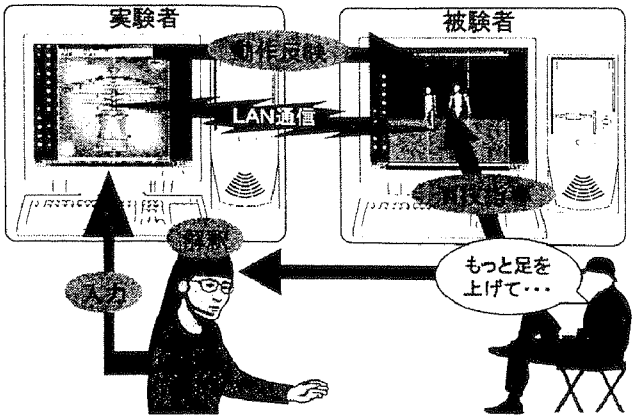


図4　実験風景

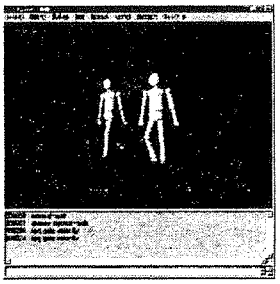


図5　IVA Logger

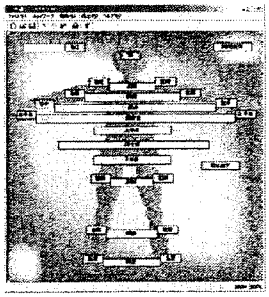


図6　IVA Remote

動作を行わせるモデルは，BVHに対応した(人間の体の関節に対応した)アクションユニット(19個)を持っている．各アクションユニットに関節の角度データを時系列データとして与えることで動作を記述することができる[10]．動作へ変化を加えるためには各アクションユニットに与える角度データを操作する必要がある．しかし，個々のアクションユニットを操作するのは煩雑であり，直感的な操作が困難である．そこで，ある程度直感的に動作を作成できるようにコントロールユニットを定義した．コントロールユニットに必要なコマンドを送ることにより，関節をまとまった単位で操作することが可能になり，動作の作成が容易になる．IVA LoggerはLAN経由で，コントロールユニットの操作をGUI化した「IVA Remote」(図6)により操作することが可能である．

3.4 特徴量の評価

主観的動作と客観的動作から，以下に示す情報を動作の特徴を表す変量として求める．

関節の回転周期の平均　：パーツの回転速度の変化
関節の回転中心の平均　：姿勢の変化
関節の運動量の平均　：パーツの動きの激しさの変化

次に，客観的動作については動作作成の前後それぞれの変量から変化の度合いを求め，動作作成に関する

特徴量とする．また，同様に主観的動作の「普通の動作」から「特別な意思の働いた動作」への変量を求め，主観的に動作へ特別な意思を働かせる際の変化に関する特徴量とする．

客観的動作とそれに対応する主観的動作との比較には，式 (1) で定義される類似度を用いる．式 (3) は変化量の行列を表しており，例えば，「腿の X 軸方向」のみの類似度を求める場合であれば「腿の X 軸方向」の変化量が用いられる．

$$S = E(M_P, M_S) : -1 \leq S \leq 1 \quad (1)$$

$$(M_P : \text{本人変化量}, M_S : \text{合成動作変化量})$$

$$E = \frac{Cov(M_1, M_2)}{Stdev(M_1) \cdot Stdev(M_2)} \quad (2)$$

$$M = \begin{bmatrix} \text{姿勢の変化率} \left(= \frac{\text{変化後の姿勢}}{\text{通常の姿勢}}\right) \\ \text{運動量の変化率} \left(= \frac{\text{変化後の運動量}}{\text{通常の運動量}}\right) \\ \text{周期の変化率} \left(= \frac{\text{変化後の周期}}{\text{通常の周期}}\right) \\ \vdots \end{bmatrix} \quad (3)$$

求めた変化類似度は以下に示す方針で分析を行う．

- 被験者集団での類似度のヒストグラムを作成
- コマンド入力表[1]を用いて入力内容の反映具合を確認
- 人間の動作形成過程を把握できる有意情報を探す

4 分析結果

作成したヒストグラムの例を図 7 に示す．ヒストグラムは「ゆっくり歩く」における腿の X 軸 (脚の前後運動) の類似度の変化である．動作 1 より動作 2 の方が類似度が高い方に偏りが現れる．しかし，動作 3 になるとその偏りが崩れることが確認できる．これは，動作 2 を作成する際に認識できた動作の違いが記憶できなかったということになる．その他の体の部位や軸に関してもヒストグラムを作成すると，類似度が高い部位には概して同様の特徴が得られた．ここから考えられることとしては，芝居をすることに，もしくはさせることに慣れていない人間が指示をする際には，自分の生成するであろう動作に関わらず，自分が確立したイメージに近づけるように指示を行なっていくということである．また，一度学習しても，手本を視覚的に提示され続けない限り，確立されたイメージは変化しないと予想される．

図 8 から図 10 は大きく手を振って歩かせたときの類似度をまとめたものである．これらを見ると，3 つの動作に共通して「腕」「肘」の類似度が高くなっていることがわかる．同様に「大股で歩く」やという場合には「腿」「膝」の類似度が高くなっている．ここから，パーツが具体的に示される個所のみ類似度が高くなる傾向

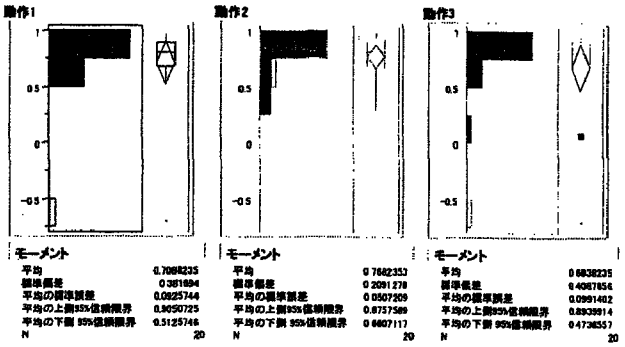


図 7 「ゆっくり歩く」:腿の X 軸の類似度の変化

があるということを確認できた．「細かく歩く」といった場合にも「腿」「膝」の類似度が高くなるという傾向が見られたが，これは歩行動作において「細かく」という指示を「歩幅を狭めよ」と同義に捉えた人間が多かったため，脚部の類似度が高くなったのではないかと予想される．

図 11 から図 13 は楽しそうに歩かせたときの類似度をまとめたものである．これらを見ると，動作 1 では「腕」「肘」の類似度が低いが，動作 2 では「腕」「肘」の類似度が改善されている．これらの特徴は，「悲しそうに歩く」「楽しそうに歩く」「怒っているように歩く」といった感情を表現する動作では共通して見られる．ここから「腕」「肘」は「感情を表現する」歩行動作において「普通に歩く」という動作と比べて実際には主観的動作を変化させているうえ，提示されれば認識できるが，指示を行なう際には意識されにくい部位だと言える．こうした変化については，変化のさせ方を考える必要があると思われる．

5 まとめ

本稿では，感情表出モデル構築に向けたディジタルアクタ学習システムを提案し，モーションキャプチャを用いて収録した主観的動作と，その本人がモデルに対して指示を行い作成した客観的動作を比較することで人間の思考プロセスを確認する手法について分析を行った．

今後は基本動作制御モデルを構築することで，そのモデルを用いた実験を行うことが望ましい．しかし，今回得られた結果の多くは，合理的な説明を行いやすいものが多く，従って，感情表出モデルを構築する際には人間の思考プロセスを確認する為に，本稿で提案した手法を用いることが可能であると考えられる．今後はより複雑な動作及び演技背景を用いて同様の実験を行い，それぞれの傾向を分析することで，より多様な表現が可能となる．

また，俳優のような芝居の訓練を受けた人間の感情

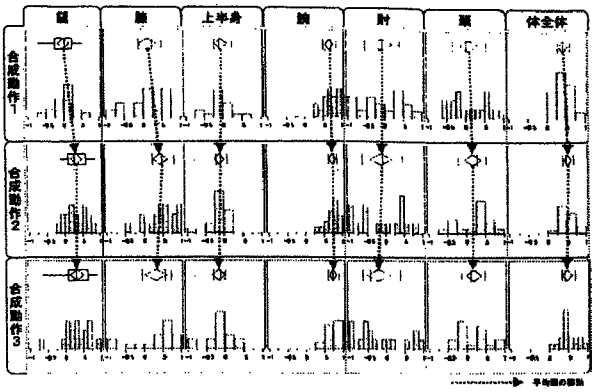


図 8　大きく手を振って歩く-X

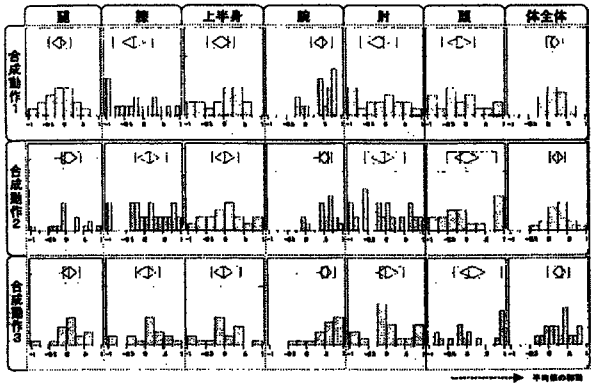


図 9　大きく手を振って歩く-Y

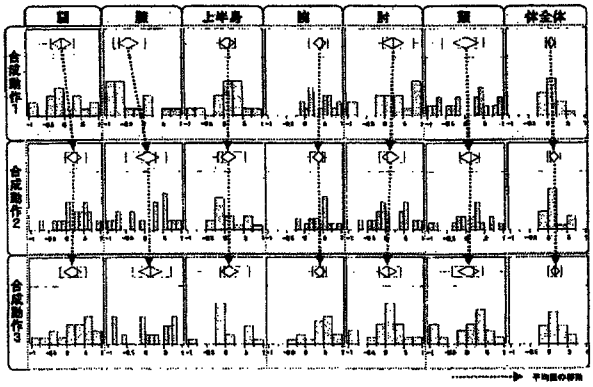


図 10　大きく手を振って歩く-Z

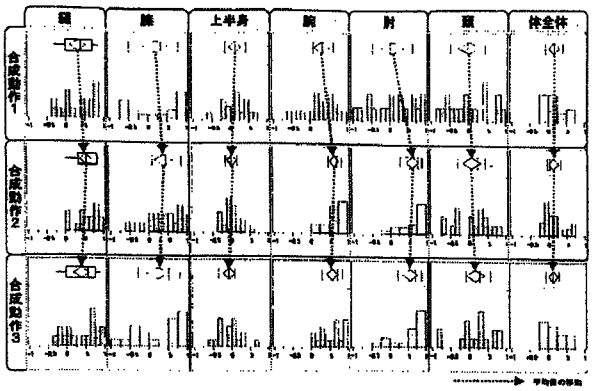


図 11　楽しそうに歩く-X

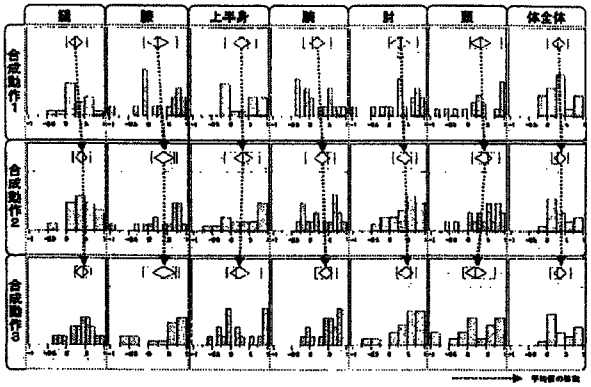


図 12　楽しそうに歩く-Y

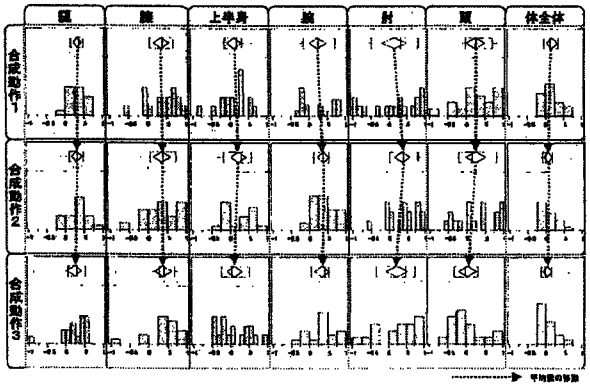


図 13　楽しそうに歩く-Z

表出プロセスと，そうでない人間の感情表出プロセスは異なることが容易に想像できることから，どういった性質の感情表出プロセスをモデル化するのかを検討する必要がある．そのためには劇や芝居，ダンスなど意図的に動作中に感情表出を行なっているものについて分析を行なう必要があると考えられる．

今後は，以上のような分析や効果的なインタフェースの検討などを経て，ディジタルアクタシステムとしての統合を行なう．

参考文献

[1] Gael Sinnier, Selim Balcisoy, Nadia Magnenat Thalmann, Daniel Thalmann, "An Interactive Interface for Directing Virtual Humans", ISCIS 1998,1998

[2] Seilm Balcisoy, Remy Torre, Michal Ponder, Pascal Fua,Daniel Thalmann, "Augment Reality for Real and Virtual Humans", 2000

[3] Gael Sinnier, Selim Balcisoy, Nadia Magnenat Thalmann, Daniel Thalmann, "VHD: a system for directing real-time virtual actors", 1999

[4] 鵜沼宗利, 武内良三, "コンピュータアニメーションにおける感情を伴った人間の歩行動作の生成方法", 電子電子情報通信学会論文誌,D- Vol.J76-D pp.1822-1831, 1998-08

[5] 中川智史, 藤本昌宏, 馬場口登, 北橋忠宏, "言語を用いた指示による人体動作アニメーションの作成", HIR-2002-57, 2002

[6] 米元聡, 谷口倫一郎, "視覚認識による 3 次元人体動作の解析・生成インタラクション", HIR2002-58, 2002

[7] 和氣早苗, 渡邉亮志, 河野泉, 上窪真一, 福住伸一, "感情表現としての非言語音に関する実験報告", 情報処理学会論文誌ヒューマンインターフェース,89-9, 2002-07

[8] N.M.Fraser and G.N.Gilbert, "Simulating Speech Systems", Computer Speech and Language,Vol..5,No.1,pp.81-99, 1991

[9] J.M.Carroll and A.P.Aaronson, "Learning by Doing with Simulated Intelligent Help", Communications of the ACM,Vol.31,No.9,pp.1064-1079, 1988

[10] 西田論司, 杉浦彰彦, "アクションユニットを用いた人間の下半身動作の合成", HIR2002-57, 2002

2003－SLP－48 (3)
2003／10／17

パラ言語の理解能力を有する対話ロボット

藤江 真也† 江尻 康† 菊池 英明‡ 小林 哲則†

†早稲田大学 理工学部 ‡早稲田大学 人間科学部

あらまし:

音声対話における人間同士のやり取りは，発話に含まれる言語情報だけでなく，それを補助する別の情報も活用して行われていると考えられる．この発話に付随して生成され言語情報の円滑な伝達を補助する情報をパラ言語情報と呼ぶ．本論文では，パラ言語情報として，韻律情報を用いた態度認識と画像情報を用いた頭部ジェスチャの認識手法を示すとともに，それを用いた対話システムを構築する．前者は，発話者の態度が肯定的か否定的かを，F0 パターンと音素アライメントを用いて識別する．後者は，頷き，傾げ，横振りの 3 ジェスチャを，オプティカルフローを特徴量，left to right HMM を確率モデルとして用いることによって認識する．実験結果からこれらの手法が，パラ言語情報としてユーザの態度を表すのに十分な性能を持っていることを示す．

Dialogue Robot with an Ability to Understand Para-Linguistic Information

Shinya FUJIE†, Yasushi EJIRI†, Hideaki KIKUCHI‡, Tetsunori KOBAYASHI†

†School of Science and Engineering, Waseda University
‡School of Human Science, Waseda University

Abstract:

The human-human interactions in a spoken dialogue seem to use not only linguistic information in the utterances but also some sorts of additional information supporting linguistic information. We call these sorts of additional information "para-linguistic information". In this paper, we present a recognition method of attitudes by prosodic information, and a recognition method of head gestures. In the former method, in order to recognize two attitudes, such as "positive" and "negative", F0 pattern and phoneme alignment are introduced as features. In the latter method, in order to recognize three gestures, such as "nod", "tilt" and "shake", left-to-right HMM is introduced as the probabilistic model as well as optical flow is introduced as features. Experimental results show that these methods are sufficient to recognize user's attitude as para-linguistic information. Finally, we show a proto-type spoken dialogue system using para-linguistic information and how these sorts of information contribute the efficient conversation.

連絡先：〒 169 東京都新宿区大久保 3-4-1 早稲田大学理工学部 コンピュータネットワーク工学科 小林 哲則
TEL: (03)5286-3379 FAX: (03)-3205-9381 E-mail: koba@tk.elec.waseda.ac.jp

1 はじめに

音声対話における人間同士のやり取りは，発話に含まれる言語情報だけでなく，それを補助する別の情報も活用して行われていると考えられる．この言語情報を補助する情報をパラ言語情報と呼ぶ．

従来の音声対話システムの多くは，ユーザとシステムが交互に発話を繰り返し，またそこでやり取りされる情報は発話文の表層的な言語情報のみを頼りにしている．しかしながら，人間同士の対話ではパラ言語情報を用いて効率的に，また豊かな表情を持って対話を進めている．より自然に対話の出来る音声対話システムを構築するにあたって，パラ言語情報を考慮するのは必要不可欠なことだと考えられる．

本論文では，ユーザが発するパラ言語情報として，発話中の韻律情報に含まれる態度と，頭部ジェスチャーの認識手法を提案，実装した．これらの実験結果を示す．また，これらのパラ言語情報を統合，利用する音声対話システムを実装し，効率的に対話を進行可能なことを示す．

2 音声対話中のパラ言語情報

2.1 パラ言語情報

パラ言語情報は主に聴覚的なものと視覚的なものに分類される．聴覚的なものとしては発話に含まれる基本周波数 (F0)，話速等の韻律的な情報や，フィラーや言い直し等の言語的には無意味，いわゆる不要語と呼ばれる類いのものなどがある．視覚的なものとしては，発話者の表情や視線の方向，頷き等の頭部ジェスチャなどがある．

韻律情報を用いた対話の研究については，韻律情報を用いて構文解析の速度や精度を向上させるもの [1]，韻律情報を用いて談話タグ (Dialogue Act) を推定するもの [2][3] などが見られる．頭部ジェスチャに関しては，個々の認識のアルゴリズムに関する研究 [4][5] や，ノンバーバルなコミュニケーションの手段として用いる研究 [6] が見られる．また，音声言語と頭部の動きの関連性に関してのデータベースが収集された例もある [7]．

2.2 対話のタスク

本研究で扱う音声対話は，ある事柄に関して決断を迷っているユーザに対して，ユーザの要求を聞き出し，提案を行う対話を想定している．本研究では，このような対話を相談型対話と呼ぶ．

今回は，昼食に何を食べるか，またどの店へ行くかを迷っているユーザの相談に応じるタスクを対象としている．

3 韻律情報を用いた発話態度認識

3.1 目的

本節では，ユーザの発話態度を韻律情報から認識する手法について述べる．

例えば，システムがユーザに対して「ハンバーガーはどうですか」と提案した状況を考えてみる．ユーザが提案に対する評価を明示的に発言に含めることは稀である．ただ単に提案の内容を「ハンバーガーかー」等と繰り返すことが多い．この発言内容自体はユーザの評価 (肯定的であるか，否定的であるか) を含んでいない．しかし，その評価はその顔の表情や動き，また発話の韻律などにパラ言語情報として現れることが多い．その評価を認識することが出来れば，効果的で円滑な対話が実現できると思われる．

本節の目的は，韻律情報を用いて，提案に対する発話に含まれるユーザの評価を認識することである．

3.2 データ

システムの提案に対するユーザ応答の音声データの収録を行った．今回は，肯定的な態度，否定的な態度のデータを数多く収集するため，システム側の提案に対応する発話を音声合成装置を用いて合成し，その応答を肯定的，否定的な態度で発話するという形で収録を行った．収録した発話の種類を表 1 に示す．この表の言い回しの項目で，「○○」はシステムの提案に含まれるカテゴリや店の復唱を表す．収録は研究室の男子学生 20 人分 (計 2000 発話) について行った．

例えば，「ラーメン」「○○かー」「否定的」なデータを収録する場合，まず合成音で「ラーメンなんてどうかな」という音声が流れる．その後ユーザが

表 1: 収録音声の種類

カテゴリ・店 (計 10 種)		言い回し (5 種)	態度 (2 種)
ハンバーガー	マクドナルド	○○かー	肯定的
ラーメン	味源	○○ねー	否定的
弁当	夢民	○○	
カレー	ホカ弁	いいんじゃない	
学食	そばの実	そうだね	

「ラーメンかー」という内容を否定的な気持ちを込めて発話するという形で行った.

3.3 認識手法

収録した発話に関して, 基本周波数 (F0) の抽出, 音素アライメントを取り観察を行った. 検討の結果, 図 1 に示すように, 以下の 3 次元の特徴 $x = (x_1, x_2, x_3)$ を識別に用いることとした.

x_1 : 第 1 モーラの母音部分の F0 の傾き

x_2 : 発話全体の F0 レンジ

x_3 : 最終モーラの継続長

ここで, 第 1 モーラの F0 の傾きは最小自乗法で求めた. また, 発話の最初に, 「あー」や「うん」等の適当なフィラーを言うことを許容したため, 特徴量の抽出はその部分を除いて行った.

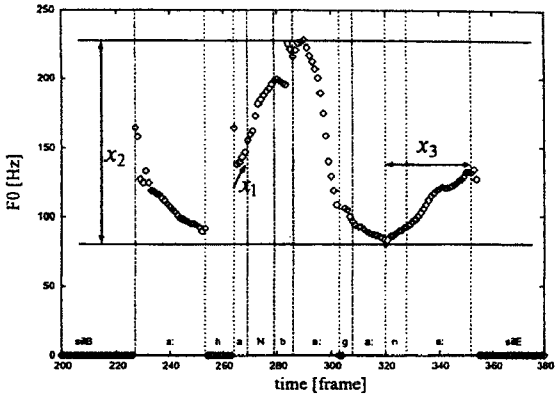
図.1 を見ると, おおよそ次のような傾向が見られる.

- x_1 は肯定的な場合に正になり, 否定的な場合は負になる.
- x_2 は肯定的な場合の方が, 否定的な場合よりも大きい値になる.
- x_3 は肯定的な場合の方が, 否定的な場合よりも小さい値になる.

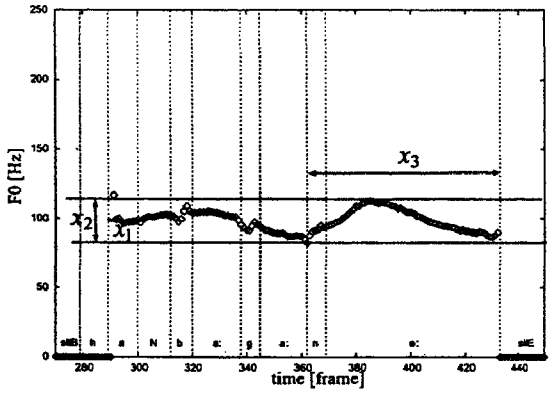
抽出した特徴量を用いて, 肯定的, 否定的の 2 カテゴリについて単一正規分布を最尤推定により学習した. 識別はこの分布を用いたベイズ識別によって行う.

3.4 実験と結果

収録した 20 人分計 2000 発話のデータを, 4 セット (1 セットあたり 5 人分 500 発話) に分けた. 3 セットで学習を行い, 残る 1 セットの識別を行う. これを各セットにつき計 4 回行い, 各セットの正解率の平均を取った. 結果を図 2 に示す.



(a) 肯定的



(b) 否定的

図 1: 特徴量抽出の例

言い回しによる違いを見るため, 最終モーラの継続長が常に短くなる「いいんじゃない」を除いた実験, また傾向が似ている「○○かー」「○○ねー」のみを用いた実験, 「○○かー」のみを用いた実験をそれぞれ行った.

図 3 に人毎の認識結果を示す.

結果を見ると, 言い回しを限定することで, ある程度の認識率の上昇が得られるが, 逆に限定をしすぎると認識率が落ちる. これは学習データ不足のためだと考えられる. また, 図 3 を見ると, 人によって認識率に違いが出ているのが分かる. 数人については肯定の方が認識し易く, 別の数人については否定の方が認識し易い. これは, 人によって認識結果が肯定的か否定的に傾く傾向があるということを表している. この結果を対話制御に応用する際に非常に重要な情報である.

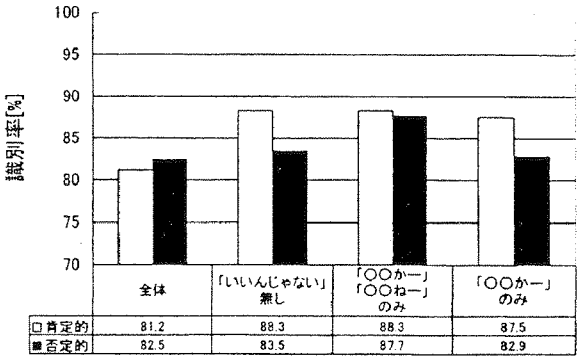


図 2: 韻律による発話態度認識の実験結果

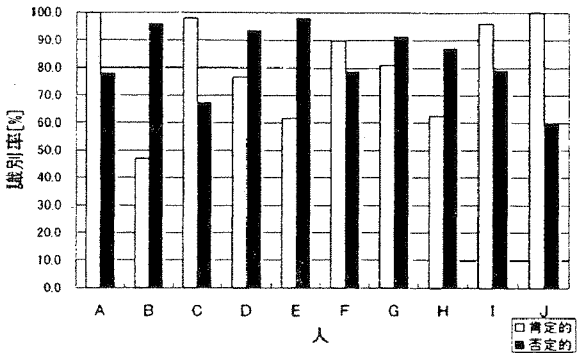


図 3: 認識率の人による違い

3.5 人による認識との比較

前節での実験は, 発話者が収録時に想定した態度を正解としている. ここでは, 人間の持つ識別能力との比較のため, 収録した発話を人が識別した結果と, 前節の手法で識別した結果との一致率を見る.

収録したデータのうち, 態度が肯定的, 否定的のものからそれぞれ 20 発話, 計 40 発話をランダムに選択した. この 40 発話を 5 人の作業者 (A〜E) に順不同に聞かせて, それぞれ否定的か肯定的かを決めさせた.

一致率の計算には, 対話データのタグ付けの評価等に用いられる Cohen の κ[8] を用いた. κ は 2 人の作業者毎に求まる値なので, 前節で述べた識別結果 (M) を含む 6 組の結果について, それぞれの組合わせで κ を求めた. その結果を人毎に集計し, 最小値, 最大値, 平均値を計算した. 結果を表 2 に示す.

平均値を見ると, 今回の識別結果は 6 人中 5 位であるが, 最小と最大の一致率は人間のものと近い. また, Cohen の κ は, $0.60 < \kappa < 0.75$ の範囲にあるときに良好な結果とされており, その点からも十分人間に近い識別能力を持つと言える.

表 2: Cohen's κ の計算結果

	最小	最大	平均
M	**0.52**	**0.80**	**0.66**
A	0.52	0.85	0.72
B	0.52	0.79	0.62
C	0.65	0.80	0.72
D	0.65	0.85	0.76
E	0.61	0.79	0.72

4 頭部ジェスチャ認識

4.1 目的

うなずき, かしげ, 首振りなどの頭部ジェスチャは, 対話中での応答発話と共起することが多い. これらの頭部ジェスチャは, 前節で述べた韻律情報と共に, 提案に対する評価を表すパラ言語情報の伝達手段の一つと言える.

本節の目的は, これら頭部ジェスチャを動画像から認識することである.

4.2 認識対象

まず, 自然な対話中にどのような頭部ジェスチャが生起するかを観察するため, 映像付きの人同士の対面対話を収録した. 相手の発話に対するフィードバックを表す頭部ジェスチャとして以下の三種類のものが顕著に現れた.

うなずき 相手の質問に対する肯定的応答, 相手の発言に対する同調や理解, 自分の発言に対する相手に同調や協調の促し等を表す.

かしげ 相手の発言に対する否定的応答, 自分の発言に対する自信の無さ, 考え込んでいる様子などを表す.

首振り 相手の発言に対する否定的応答を表す.

本研究での認識対象は上記の 3 カテゴリである.

4.3 認識手法

本手法では，頭部の動きを特徴量として用いる．頭部領域の抽出は，入力画像 (図 4(a)) から肌と髪の色情報を用いて，頭部領域を抽出する (図 4(b))．さらに，頭部の縦横の比を用いて首領域の除去を行った (図 4(c))．抽出した頭部領域の全画素に対しグラディエント法によって求めたオプティカルフローを，頭部の動きとして用いた．

(a) 入力画像

(b) 色モデルによる抽出　　(c) 首領域の除去

図 4: 頭部の抽出

図 5 のように頭部領域を 4 分割し各領域毎のフローに平均ベクトルを求め，計 8 次元の特徴量とした．このようにすることで，図 6 のように各ジェスチャの特徴を捕らえることが可能となることが期待できる．

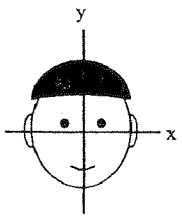
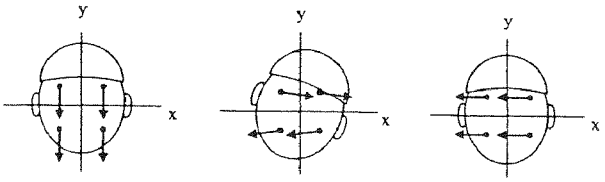


図 5: 頭部領域の分割

認識には，音声認識でよく使われる left to right の HMM(Hidden Markov Model) を用いた．本研究では，対話中に現れる頭部ジェスチャを認識するためスポッティング認識を行う必要がある．スポッティング認識を行うために，今回設定した三つのジェスチャ以外に，静止モデルとガーベッジモデルの二つのモデルを用意した．静止モデルは頭部が動かないものを表すモデル，ガーベッジモデルは設定した三つのジェスチャ以外の動きを表すモデルである．



(a) うなづき　　(b) かしげ　　(c) 首振り

図 6: 各ジェスチャの特徴的なフロー

4.4 実験と結果

実験のために集めたデータは，4.2 節で述べた自然対話中のデータ (15 セット各 4 分) と，新たに収録した 25 セット計 144 分の，計 40 セット 204 分のデータである．新たに収録した 25 セットは，ユーザにジェスチャを指定して実際に行って貰ったものを収録したものである．指示中の動作も含まれるため，一部自然なジェスチャも含まれる．

このデータのジェスチャ部分に，人手でタグ付けを行なった．その結果，データから計 2148 個のサンプルが得られた (静止，ガーベッジは除く)．

これら 40 セットのデータを，学習データ 35 セット，認識データ 5 セットに分けて認識実験を，各認識セット毎に行った．認識結果をコンフュージョンマトリックスで表したものを表 3 に示す．

脱落誤りは正解ジェスチャが静止もしくはガーベッジとして認識された回数，挿入誤りは静止もしくはガーベッジが正解ジェスチャのどれかとして認識された回数を示す．

この実験に用いたモデルのパラメータは，HMM の状態数が 11，正解ジェスチャモデルの混合数が 4，静止モデルの混合数 1，ガーベッジモデルの混合数 16 である．様々なパラメータで実験を行った結果，認識率的に最良のものである．

表 3: 頭部ジェスチャ認識の実験結果 (認識率最良)

	認識結果			
	うなづき	かしげ	首振り	脱落誤り
うなづき	**1144**	6	0	233
かしげ	9	**322**	4	189
くびふり	2	2	**290**	18
挿入誤り	347	157	42	

実際の対話に応用する際には, 挿入誤りや置換誤りはユーザの反応に対する誤解を生ずるので, 脱落誤りに比べて深刻な問題になる. そこで, 挿入誤りや置換誤りが少なくなるパラメータを実験的に調べた. その際の認識の結果を表 4 に示す. この際のパラメータは, HMM の状態数が 9, 正解ジェスチャモデルの混合数が 1, 静止モデルの混合数 1, ガーベッジモデルの混合数 16 である.

表 4: 頭部ジェスチャ認識の実験結果 (挿入誤り最小)

	認識結果			
	うなづき	かしげ	首振り	脱落誤り
うなづき	**822**	1	0	557
かしげ	2	**135**	14	381
くびふり	0	5	**286**	41
挿入誤り	112	66	38	

5 音声対話システム

5.1 パラ言語情報の統合

3, 4 節で述べた韻律情報による態度の認識と頭部ジェスチャの認識のそれぞれの結果を, パラ言語情報として応用するためには, 両者の情報を統合することを考えなければならない. 特に, 両者に関して相反する結果が得られた時 (一方が肯定的な結果で, もう一方が否定的な結果の時), どのように判断するかが問題である.

本研究では, 韻律情報から得られる態度と, 頭部ジェスチャの共起によって受ける印象から, 表 5 に示す統合を行った. 例えば, 頭部ジェスチャが「うなずき」で韻律情報による態度が「肯定」の場合は「強い肯定」と解釈するが, 韻律情報による態度が「否定」の場合はユーザは判断に迷って考えていると解釈する.

表 5: 認識結果の統合
(* は対話中ほとんど現れない組合わせを表す)

		頭部ジェスチャ		
		うなずき	かしげ	首振り
韻律	肯定	強い肯定	弱い肯定	思案*
	否定	思案*	否定	強い否定

5.2 音声対話システム

今まで述べたパラ言語情報を用いた音声対話システムを, ヒューマノイドロボット ROBISUKE 上に実装した (図 7). 我々は, 従来から視線や眉毛, 口を使った顔の表情や, 腕を使ったジェスチャを用いた対話を行うヒューマノイドロボットを開発しており [10][11][12], ROBISUKE その最新型にあたる.

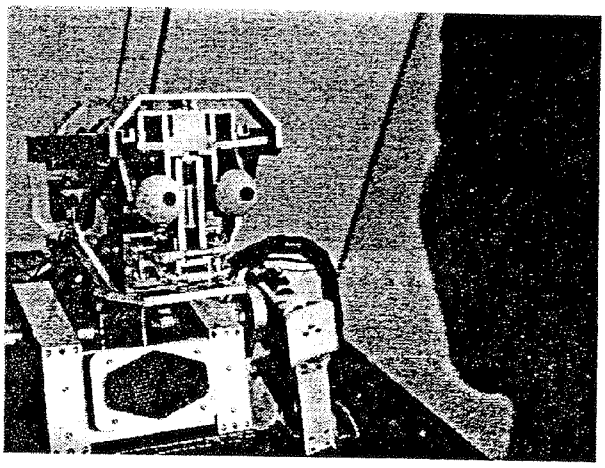


図 7: ROBISUKE

この音声対話システムでは, システムの提案に対するユーザの応答を 5.1 節で述べた手法で統合し, 以下の様に判断, 対応を行う.

- 強い肯定 : 提案はユーザに受け入れられたと判断し, システムは自信を持って詳細な提案や情報を提供する.
- 弱い肯定 : 提案はユーザに大部分受け入れられたと判断し, システムは詳細な提案や情報を提供する.
- 思案 : 決定は見送られ, システムは「うん」や「何?」などを発話しながらうなずき, ユーザにもう一度応答を促す.

- 否定または強い否定：提案は却下されたと判断し，システムは代替案を提供する．

5.3 対話例

対話例を図8に示す．

システムは，提案に対してユーザの応答が否定的な場合は代案を，肯定的な場合はより具体的な提案を行う．

対話例では，「カレーか」「弁当ね」などの言語情報だけでは判断しかねる発話に対し，パラ言語情報の利用によって判断を行い，対話を円滑に進められることが確認できる．

U: お昼ご飯なんだけど， どこかいいところ無いかな?
R: カレーなんてどう?
U: カレーかー (強い否定)
R: それじゃあ，ハンバーガーなんてどうかな
U: あーハンバーガーね (強い肯定)
R: ハンバーガーなら， 近くにマクドナルドがあるよ

図 8: 対話例

U:はユーザの発話，R:は ROBISUKE の発話

6 まとめ

音声対話で利用されるパラ言語情報として，韻律情報による発話態度の認識と頭部ジェスチャの認識を行った．前者の識別能力に関して，人間とほぼ同等の能力を持つことを一致率をもとに確認した．た，パラ言語情報を利用した対話システムの例を示し，対話の進行に有効に作用することを確認した．

今後は，パラ言語情報を用いることによる有効性の評価法，応答以外の発話に関する態度の認識手法などについて検討を行う予定である．

参考文献

[1] Christian Lieske, Johan Bos, Martin Emele, Björn Gambäck, and CJ Rupp, "Giving prosody a meaning," in Proceedings of ISCA EUROSPEECH'97, vol. 3, pp. 1431-1434, 1997.

[2] E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann, "On the use of prosody in automatic dialogue understanding," Speech Communication, vol. 36, pp. 45–62, 2002.

[3] Helen Wright Hastie, Massimo Poesio, and Stephen Isard, "Automatically predicting dialogue structure using prosodic features," Speech Communication, vol. 36, pp. 63–79, 2002.

[4] Shinji Kawato and Jun Ohya, "Real-time detection of nodding and head-shaking by directly detecting and tracking the 'between-eyes'," in Proceedings of Fourth IEEE international conference on automatic face and gesture recognition, pp. 40–45, 2000.

[5] Ashish Kapoor and Rosalind W. Picard, "A real-time head nod and shake detector," Tech. Rep. 544, MIT Media Laboratory Affective Computing Group, 2002.

[6] Hiroshi Kobayashi and Fumio Hara, "Facial interaction between animated 3d face robot and human beings," in Proceedings of 1997 IEEE International Conference on Systems, Man and Cyberntics(SMC97), vol. 4, pp. 3732–3737, 1997.

[7] S. Hayamizu, O. Hasegawa, K. Itou, K. Sakaue, K. Tanaka, S. Nagaya, M. Nakazawa, T. Endoh, F. Togawa, K. Sakamoto, and K. Yamamoto, "RWC multimodal database for interactions by integration of spoken language and visual information," in Proceedings of Fourth International Conference on Spoken Language (ICSLP96), vol. 4, pp. 2171–2174, 1996.

[8] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37-46, 1960.

[9] 江尻康, 松坂要佐, 小林哲則, "対話中における頭部ジェスチャの認識," 電子情報通信学会技術研究報告, Vol.102, No.218, pp.31-36, PRMU2002-61, 2002.

[10] Tsuyoshi Tojo, Yosuke Matsusaka, Tomotada Ishii, and Tetsunori Kobayashi, "A conversational robot utilizing facial and body expressions," in Proceedings of 2000 IEEE International Conference on Systems, Man and Cyberntics(SMC2000), vol. 2, pp. 858-863, 2000.

[11] Yosuke Matsusaka and Tetsunori Kobayashi, "System software for collaborative development of interactive robot," in Proceedings of IEEE-Humanoids2001, pp. 271-277, 2001.

[12] 松坂要佐, 東條剛史, 小林哲則, "グループ会話に参与する対話ロボットの構築," 電子情報通信学会論文誌, Vol.J84-D-II, No.6, pp.898-908, 2001.

社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

音声対話における韻律を用いた話題境界検出

大久保 崇† 菊池 英明†† 白井 克彦†

† 早稲田大学 理工学部
†† 早稲田大学 人間科学部

あらまし 音声対話の話題による自動分割を目指して，話題の切れ目の深さと韻律情報の関係を分析した．これまでにも，談話構造と韻律情報の関係を分析した研究は行われてきたが，本研究ではより詳細な分析を行うため，発話の開始アクセント句と終了アクセント句における韻律情報に関する20種類のパラメータを用いた．その結果，幾つかの韻律情報について話題の切れ目の深さとの相関が得られた．次に，決定木学習を利用した話題の切れ目の深さを判別する実験を行った．判別実験の精度は今後の検討を要するものであったが，得られた決定木から，話題の切れ目の判別に寄与する韻律情報についての考察が得られた．

キーワード 音声対話, 韻律情報, 談話セグメント, 決定木

Detection of Topic Boundary by Prosodic Information in Spoken Dialogues

Takashi OKUBO†, Hideaki KIKUCHI††, and Katsuhiko SHIRAI†

† School of Science and Engineering, WASDEA University
†† School of Human Science, WASEDA University

Abstract We analyzed relationship between topic segments and prosodic information in spoken dialogues, for the purpose of realization of automatic topic segmentation. We introduced twenty types of parameters of prosodic information in initial and final accentual phrases of utterances, and investigated correlation between those parameters and topic boundaries. As the result, it was confirmed that some prosodic features correlated with topic boundaries strongly. Next, we constructed decision tree model to judge strength of topic boundaries and measured its accuracy. As the result, we get some considerations about relationship between prosodic features, discourse markers and speaker-changes.

Key words spoken dialogue, prosodic information, topic segment, decision tree

1. はじめに

本研究では，人間同士の会話を自動的，又は半自動的に話題で分割する技術の実現を目的とする．この技術は，会議の議事録の自動生成や，会話の内容による動画像のシーン分割に応用できる．また，リアルタイムで話題の分割が行えれば，複数のタスクにまたがって自由度の高い会話をするような対話システムにも応用が可能である．

話題境界の検出には，会話の内容理解が有効であることは言うまでもないが，談話構造と音声の韻律情報の関係を分析した研究[1][2]から，韻律情報の利用も有効であると考えられる．[1]では，対話における話題の切れ目の深さについて，基本周波数とパワーとの間に相関が見られることが報告されている．また，より詳細な韻律情報と独話における話題の切れ目の深さとの関係を分析した[2]では，話題の切れ目の前後の発話について，発話間のポーズ，発話末と発話開始位置のアクセントにおけるピッチレンジリセットの程度などが関係を持つことが報告されている．そこで我々は，対話における話題の切れ目の深さと韻律の関係をより詳細に分析し，その分析結果に基づいて決定木を用いた判別実験を行う．

本稿ではまず，対話における話題の切れ目の深さと韻律情報の各パラメータとの関係を分析する．特に基本周波数やパワーに関しては，発話における位置や，話題の切れ目との前後関係に着目する．そして，分析によって得られた話題の切れ目と韻律の関係を考慮して，決定木を用いた話題の切れ目の判別を目指す．以下では，2章で分析方法や分析データについて述べ，3章に分析の結果を示す．そして，4章で決定木を用いた判別実験を行い，5章でまとめとする．

2. 話題境界と韻律の各パラメータの比較

2.1 分析に使用するデータ

2.1.1 対話データ

分析には，人工知能学会「談話・対話研究におけるコーパス利用研究グループ」によって作成されたコーパス[3]を使用した．このコーパスは2名の話者間で行われた課題遂行対話を収録したものであり，「談話タグワーキンググループ」の活動の中で検討された様々なタグ[4]が付与されている．

2.1.2 TBI

本研究では，対話中の話題の切れ目を，コーパスに付与された「談話セグメントタグ」(Topic_Break_Index:以下 TBI) [5]に基づいて決定する．TBI は 1，2 の二段階で話題の切れ目の深さを表し，話題の遷移が大きい場合に 2 を，そうでない場合は 1 をタグ付け作業者の主観で評価し付与する．また，完全に話題が連続している発話には TBI は付与されておらず，本稿ではこのような発話を TBI が 0 の発話として定義する．

2.2 分析方法

まず，コーパスから 5 つの対話（タスクはクロスワード，旅館予約，会議室予約，地図課題）を選び，対話中のすべての発話から話題の遷移に関係ないと思われる「相槌」「フィラー」「言い淀み」のみで発話を構成しているものを分析対象から除外した．表 1 に分析対象となる発話の数と TBI の内訳を示した．

表 1 分析に用いるデータの TBI 毎の発話数

TBI	0	1	2	計
発話数	72	90	38	200

次に，分析対象となる発話に対して，日本語話し言葉音声の韻律ラベリングスキーム X-JToBI [6] を用いて韻律ラベルを付与し，そのラベル情報をもとに開始アクセント句と終了アクセント句（図 1 参照）を抽出することによって，分析対象を発話の開始部分と終了部分に限定した．この時，話者によっては多くの発話が「えー」や「あのー」といったフィラーから始まる場合があり，分析対象の多くがフィラーになってしまう事を防ぐため，最初のアクセント句がフィラーの発話については，フィラー以外の最初のアクセント句を発話の開始アクセント句として抽出した．そして，抽出した区間の韻律情報を ESPS/waves+を用いて測定した．発話速度を除いた韻律情報の各パラメータについて，話者の違いによる絶対的な差を考慮し，話者ごとの平均と標準偏差を用いて標準化した値を使用した．扱う韻律パラメータは以下の通りである．

- 基本周波数の最大 (max)・最小 (min)・レンジ (range)・平均 (ave)
- パワーの最大 (max)・平均 (ave)
- 発話速度 (speed)

以上のような韻律パラメータを用いて，TBI との関係について分析を行う．図 1 に，分析の際に発話のどの部分を対象にしているかを示す．図中の対象発話は TBI の付与されている発話であり，その直前の発話を先行発話と定義する．

まず，各 TBI の対象発話おける韻律的な特徴を調べる．次

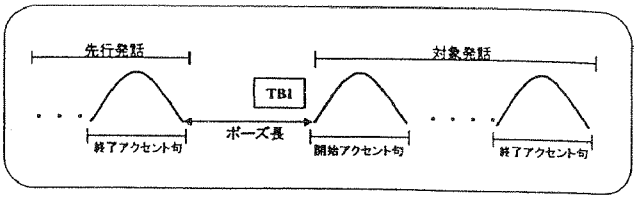


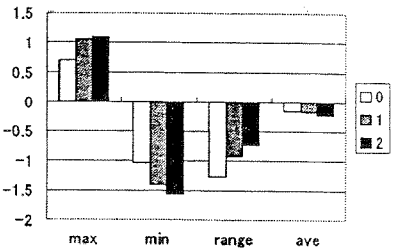
図 1 発話中の分析に用いる部分

に，対象発話の開始アクセント句と先行発話の終了アクセント句との差分をとり TBI との関係を分析する．また，各韻律パラメータだけでなく，図中に示した対象発話と先行発話間のポーズ長と TBI の関係や，話者交代の有無と TBI の関係についても分析する．

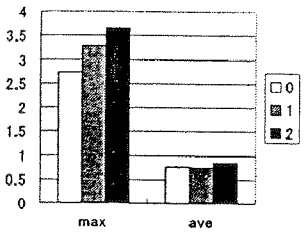
3. 分析結果

3.1 対象発話の終了部分の韻律と TBI の関係

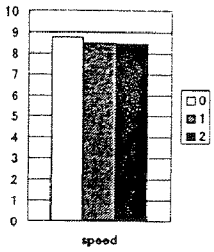
図 2 に，TBI 毎に対象発話の終了部分における各韻律パラメータを平均した結果を示した．このグラフから，まずパワーの最大について TBI と正の相関が見られる．また，基本周波数の最大とレンジについても正の相関が見られる．これらは，TBI が 1 や 2 の発話に，疑問形や半疑問形のイントネーションで終わる発話が多いことが影響していると考えられる．



(a) 基本周波数



(b) パワー

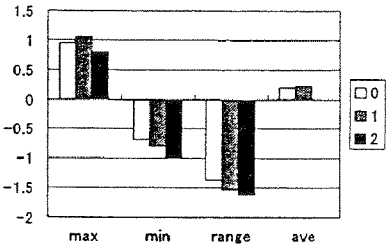


(c) 発話速度

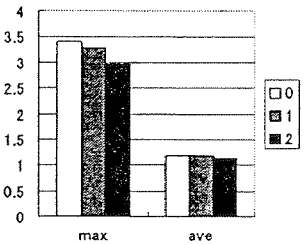
図 2 発話の終了部分の韻律パラメータと TBI の関係

3.2 対象発話の開始部分の韻律と TBI の関係

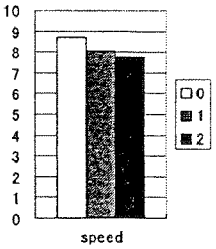
図 3 に，TBI 毎に対象発話の開始部分における各韻律パラメータを平均した結果を示した．結果から開始アクセント句の各パラメータに関して考察する．発話速度に関して，TBI が大きくなると発話速度が遅くなるという傾向が見られる．このことから，話題の切れ目が深くなるほど，最初の部分をゆっくり話す傾向があり，これは，話者が新しい話題の開始を意識して，わかり易くゆっくり話し出す傾向がある事を示していると考えられる．しかし，基本周波数やパワーについては，TBI とは負の相関が見られる．従来の研究では，話題を変える発話の開始部分では，基本周波数やパワーが他の発話に比べて大きくなる傾向が見られていたことを考慮して，発話データの内容からその要因を考えてみたところ，談話標識が関係している可能性が考えられた．そこで，次節で談話標識と TBI の関係を述べ，談話標識を考慮して発話の開始部分の分析を行う．



(a) 基本周波数



(b) パワー



(c) 発話速度

図 3　発話の開始部分の韻律パラメータと TBI の関係

3.3 発話の開始部分と談話標識

談話標識とは「話題の始まり，転換，途切れた会話の再開など，談話同士の対応付けの機能を持つもの」であり，本研究で用いた対話コーパスに対して談話標識の認定を行った研究 [7] では，図 4 に示す語を談話標識と認定している．

これらの語は TBI が 1 や 2 の発話の先頭に多く現れる事が確認されている [7]．図 5 に，TBI 毎に談話標識を含む発話の割合を示した．TBI が 0 の発話には談話標識がほとんど現れず，逆に 2 の発話では約半分の発話が談話標識から始まっていることがわかる．また，図 4 からわかるように，いずれの談話標識も 1 モーラから 4 モーラ程度の短い語が多く，韻律的な特徴が現れにくいと考えられる．また，談話標識には機能語としての特性から，通常の発話の開始アクセント句とは韻律的な特徴が異なることも考えられる．これらのことから，開始アクセント句が談話標識である場合，それを考慮に入れた分析を行う必要があると考えた．そこで，発話の開始アクセント句に談話標識がある発話については，その次のアクセント句（それが談話標識やフィラーの場合はその次のアクセント句）を発話の開始アクセント句として分析の対象とした．図 6 に TBI 毎に各韻律パラメータの平均を求めた結果を示す．

「じゃ」「では」「さいしょ（に）」
「まず」「つぎ（に）」「こんど」
「それでは」「で」「そういえば」
「そしたら」「それで」「さいご（に）」
「それじゃ」「そうすると」「でしたら」
「ま」「ようするに」「つまり」

※ ( ) 内の語は省略も可

図 4　談話標識の候補語

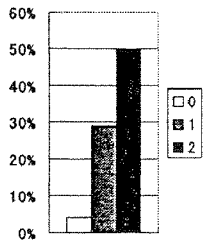
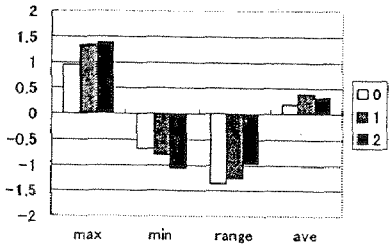


図 5　談話標識から始まる発話の割合と TBI の関係

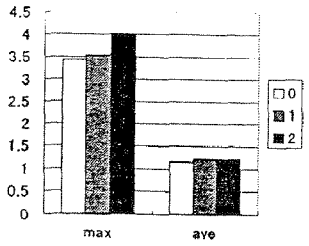
図 6 より，基本周波数のレンジやパワーの最大値に TBI との相関が見られることがわかる．この事から話題の切れ目が深くなるほど，発話の開始部分の声が大きくなる傾向があることが考えられる．基本周波数のレンジについては，アクセントの有る無しを考慮に入れる必要があるのでいちがいにはいえないが，話題の切れ目の深さに比例して基本周波数も大きくなるという事も考えられる．また，発話速度については，談話標識を除外することで，よりはっきりと TBI との負の相関を確認することができた．以上の事から，談話標識を考慮することで，TBI 毎の韻律の差異をより明確にとらえることが考察できる．

3.4 先行発話との差分と TBI の関係

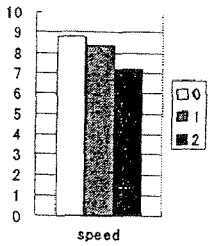
次に，発話の前後関係を考慮にいれるため，TBI 対象発話の開始アクセント句における各パラメータと，先行発話の終了アクセント句における各パラメータの差分を計算し平均を算出した．この時の開始アクセント句は，談話標識を除いたものを採用する．結果を図 7 に示す．



(a) 基本周波数

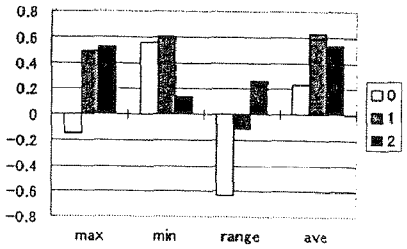


(b) パワー

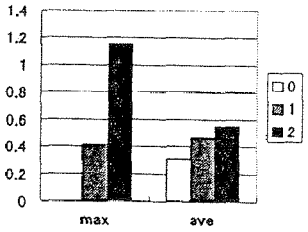


(c) 発話速度

図 6　談話標識除外をした発話の開始部分の韻律パラメータと TBI の関係



(a) 基本周波数



(b) パワー

図 7　先行発話との差分と TBI の関係

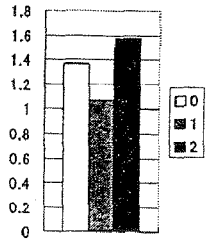


図 8　発話間のポーズ長と TBI の関係

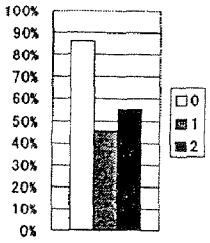


図 9　話者交代の起こる割合と TBI の関係

図 7 より，直前発話との差分をとることで，基本周波数の平均，最大値とパワーの最大値，平均値において，TBI とのより大きな相関を確認することができた．この事から，話者が話題を転換させる場合，直前の発話よりも声を大きく，高くする傾向があることがわかる．

3.5　発話間のポーズ・話者交代と TBI の関係

発話間のポーズ長や話者交代の有無についても TBI との関係を分析する．図 8 に TBI 毎のポーズ長を平均した結果を，図 9 に TBI 毎の話者交代の起こる割合を示した．図 8 から，対象発話と先行発話の話題の切れ目が最も深いとき，ポーズ長が最も長くなることがわかる．図 9 より，TBI が 0 の時が最も話者交代が起こる割合が多い．

以上，話題の切れ目の深さと韻律の関係を分析した．次章では，これらの分析結果に基づいて，決定木を用いた話題の切れ目の深さの判別実験を行う．

4.　決定木学習を用いた話題の切れ目の深さの判別

本章では決定木学習を利用して分析対象である 200 発話を，TBI が 0 か 1 か 2 の 3 カテゴリに分類する．決定木の作成には C4.5 アルゴリズム [8] を利用する．また，冗長な木ができるのを防ぐため，枝刈りを信頼度 1%で行う．用いるパラメータについては次節で改めて説明する．

まず，クローズトなデータに対する決定木の評価を行うため，分類対象となる 200 発話すべてを訓練データとして決定木を作成し，木の精度や分類に有効なパラメータを考察する．次に，10 分割の交差検定を行い，判別率などから決定木学習による判別精度を評価する．

4.1　決定木作成に用いるパラメータ

決定木作成には，3 章の分析に用いた基本周波数，パワー，発話速度の各種韻律パラメータを用いる．これらの韻律パラメータについて，決定木中の表記方法の一覧を表 2 に示した．表の

1行目にある，"開始"は開始アクセント句を，"終了"は終了アクセント句，"差分"はTBI対象発話の開始アクセント句と先行発話の終了アクセント句の差分を表す．

表2 韻律パラメーター一覧

		開始	終了	差分
基本周波数 (f0)	最大	F-f0_max	L-f0_max	S-f0_max
	最小	F-f0_min	L-f0_min	S-f0_min
	レンジ	F-f0_rang	L-f0_rang	S-f0_rang
	平均	F-f0_ave	L-f0_ave	S-f0_ave
パワー (pw)	最大	F-pw_max	L-pw_max	S-pw_max
	平均	F-pw_min	L-pw_min	S-pw_min
発話速度		F-speed	L-speed	

また，表に示した20種の韻律パラメータの他に，3章の分析で取り上げた以下の3つの要素もパラメータとして用いる．

- pause：発話間のポーズ長
- 話者交代：話者の交代が起こったかどうか
- 談話標識：発話の先頭が談話標識であるかどうか

以上の23種類のパラメータから決定木を作成する．パラメータについて，話者交代と談話標識は"有"か"無"の2値の変数とし，それ以外のパラメータは実数である．

4.2 クローズトな決定木の評価と考察

全200発話を訓練データとして用いて決定木を作成した．図10は，作成された決定木の概要である．図中の楕円には分岐に使われたパラメータを表記し，その両脇には分岐の条件を記した．談話標識と話者交代のパラメータに関しては，分岐条件は"有"か"無"かのみであり，その他のパラメータに関しては，C4.5アルゴリズムによって算出された閾値より大きいか小さいかである．図では，右の分岐に閾値より大きいデータを，左の分岐に閾値より小さいデータが分類されるように示した．また，決定木の葉にあたる部分には，分類されたTBIの結果を示した．

決定木の精度評価は，決定木の判別率と判別の信頼度を示すκ値を用いる．表3の上段に，図10に示した決定木の判別率とκ値を示した．この判別率とκ値から，作成された決定木による分類は精度が高く信頼できるものであると考えられる．そのため，作成された決定木の分岐条件からTBIの分類に寄与するパラメータを考察することには十分意味があるといえる．そこで，決定木の分岐に用いられるパラメータを考察していく．

図10から，まず，談話標識有無や話者交代有無といった，発話の韻律以外の情報を用いて分類が行われ，その後幾つかの韻律情報を用いてTBI毎に分類されていることがわかる．また，分岐に使われる韻律情報を見ると，発話の開始部分の情報が多く使われていることがわかる．この事から，発話の開始部分の韻律情報がTBIの判別に寄与していることがわかる．また，韻律以外の情報によって，判別に寄与する韻律情報が変化することも考察される．例えば，話者交代が行われない場合は，その発話の開始部分の韻律を分類に使っているが，話者が交代する場合は，直前の発話との差分，つまり相手の発話との比較を分類に用いていることがわかる．

表3 決定木による判別の精度

	判別率	κ値
クローズトな木の評価	81 %	0.7
交差検定による評価	56.5 %	0.301

4.3 交差検定による判別の評価と考察

次に200発話に対して10分割の交差検定を行い，決定木を用いたTBIの判別精度を調べる．交差検定による判別精度も，さきほどの決定木の評価と同様に，判別率とκ値に基づいて評価する．交差検定による判別率は表3の下段に示した．判別率をみてみると，60%未満であり判別の精度がよいとは言い難い．またκ値についても0.3と低い値しか得られなかった．しかし人手によるTBIの付与実験を行った結果[1]においても，κの平均は0.385とそれほど高くない．この事から，決定木における判別のκがとりわけ低いわけではないことがうかがえる．

決定木による判別の精度がよくなかった原因として，データ量の不足やパラメータの不足が挙げられるが，根本的な判別の精度の向上には韻律情報の特性や談話構造における話題の特性を考慮することが必要だと考えられる．そもそも韻律情報には，談話構造に関する情報以外に，発話に含まれるイントネーションや感情といった情報も含まれている．これらを考慮せずに，韻律パラメータから話題の切れ目の判別を行おうとしたため，イントネーションや感情といったその他の情報の影響を受け，判別がうまくいかなかったといえる．図10の決定木では，談話標識や話者交代といった情報に関する分岐に始まり，その分岐毎に異なる韻律パラメータを用いて話題の切れ目が判別されている．談話標識は話題の始まりなどに関係する言語情報であり，話者交代は対話の構造に関する情報である．このように，話題の遷移に関する情報と，それに関係の深い韻律情報を用いて判別を行っていくことが，判別精度の向上につながると考えられる．

5. まとめ

本稿では，基本周波数やパワーを中心とした韻律情報と話題の切れ目の関係を分析し，決定木による判別を検討した．

分析の結果，いくつかの韻律パラメータにTBIとの相関が見られることがわかった．まず，発話の終了部分に関しては，パワーの最大や基本周波数のレンジに相関が見られた．次に，発話の開始部分に関しては，TBIが大きくなるほど談話標識が現れる割合が多くなる傾向が見られた．また，その談話標識は1モーラから4モーラの短い語であり，韻律的な特徴が現れ難いため，談話標識から始まる発話は，談話標識の次の部分を発話の開始と考えてやることで，韻律とTBIとの間に相関が見られることがわかった．また，発話の開始部分と先行する発話の終了部分との差分を取ることで，発話の前後の韻律とTBIとの関係を分析したが，話題の切れ目が深くなるほど，先行する発話より基本周波数やパワーを大きく発話する傾向があることもわかった．

このように，TBIと韻律の間には判別に有効と思われるいくつかの関係が見られた．そして，実際に決定木を使って判別実

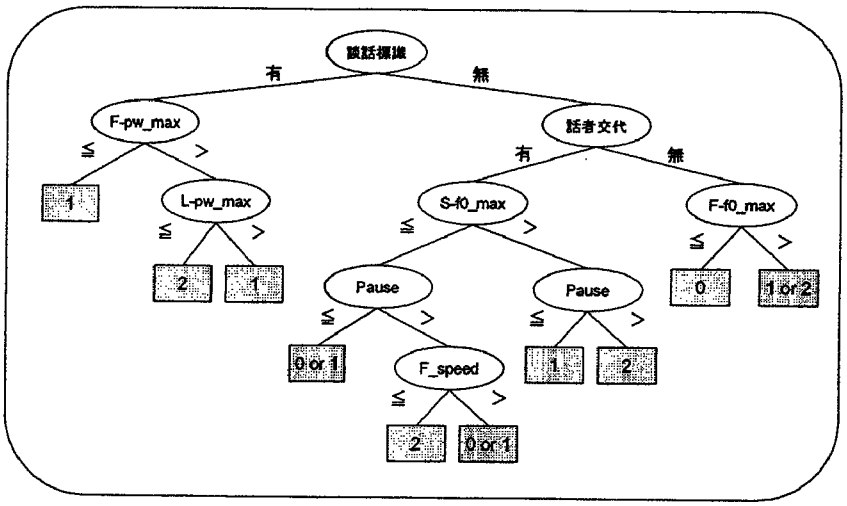


図 10　全 200 発話から作成した決定木

験を行ってみたところ，分析において TBI との相関が見られたパラメータが分類にも寄与していることがわかった．また，分類のされ方を順に見ていくと，最初に談話標識の有無や話者交代の有無で分岐され，その分岐毎に異なる韻律情報を使って TBI が分類されていることが得られた．このように，話題の遷移に関係する言語的な情報や対話構造の情報にともない，話題の切れ目に寄与する発話中の韻律情報が変化することが考察できた．

今後は，話題の遷移との関係が深いと思われる対話の構造や言語的な情報と，それに関係した韻律的な特徴を分析することで，話題の切れ目と韻律のより詳細な分析を行っていく．そして，その結果に基づいて話題境界の判別精度の向上を目指す．

謝　　辞

本研究の一部は，早稲田大学理工学総合研究センターの研究課題「マルチモーダル情報空間における統合的ヒューマンインタフェースに関する研究」によるものである．ここに記して謝意を表する．

文　　献

[1] 村井美智代, 山下洋一 , "談話セグメントと韻律情報の関連について", 第 28 回 SIG-SLUD, pp37-44,1999
[2] 小磯花絵, 米山聖子, 槙洋一, "「日本語話し言葉コーパス」を用いた談話構造と韻律との関係に関する一考察", 人工知能学会研究会資料, SIG-SLUD-A203-P17, pp.139-144, 2003.3. 9
[3] 人工知能学会「談話・対話研究におけるコーパス利用研究グループ」1999 年度版音声対話コーパス , [CD-ROM],1999
[4] 人工知能学会 談話・対話研究におけるコーパス利用研究グループ, "様々な応用研究に向けた談話タグ付き音声コーパス", 第 28 回 SIG-SLUD, pp19-24,1999
[5] 山下洋一, 小磯花絵, 堀内靖雄 , "音声対話に対する談話セグメントタグ方式の検討", 人工知能学会誌 Vol.14 84-91,199
[6] 前川喜久雄, 菊池英明, 五十嵐陽介, "X_JToBI:自発音声の韻律ラベリングスキーム", 情処研報,SLP-39-23,pp25-30, 2002
[7] 中里収, 田本真詞, 菊池英明, 吉村隆, "課題遂行対話における対話潤滑語の認定", 人工知能学会誌, Vol.14, No.5, pp.900-906, 1999.9.
[8] J.Ross Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers

FSTと韻律情報を用いた相槌・復唱機能を持った対話ロボット

Conversation Robot with Feedback/Repetition Function based on FST and Prosody Information

藤江 真也 Shinya Fujie　福島 健太 Kenta Fukushima　柴田 大輔 Daisuke Shibata　小林 哲則 Tetsunori Kobayashi

早稲田大学 理工学部
School of Science and Engineering, Waseda University

Abstract: A dialog robot which can generate a back-channel feedback and repetition of the user's word approprieatly is developed. The contents of these feedbacks depend on the contents of the dialogue partner's utterance, which are provided by the speech recognition. In order to determine the content of the feedback earlier than the beginning of it, we used finite-state tranceducer based speech recognizer which outputs the content of the feedback. And we used prosody information, especially the fundamental frequency(F0) and power of the utterance, to extract the timing of the feedback. Finally, we implemented these modules and applied them to the spoken dialog system on the humanoid robot ROBISUKE.

1. はじめに

1.1 音声対話における相槌・復唱の役割

人同士の対話は, 発話権をやりとりしながら自然に対話を進行している. 従来の音声対話システムでは, ユーザ発話の音声認識の区切りをユーザからの発話権の譲渡, システム発話の音声合成の終了をシステムからの発話権の譲渡と, 暗黙のうちに決定されていると言える. この様なシステム構成の場合, システムは黙ったままユーザの発話の終了を待つことになる. しかしながら, 実際の人同士の対話においては, 話し手が発話を行っている最中に, 聞き手は自分の状態を何らかの形でフィードバックする. それは例えば, 話し手の話を聞いているという合図である相槌であったり, 内容の理解度をフィードバックするための復唱であったりする. 話し手はこのフィードバックを受け取ることにより, 聞き手が自分の話を聞いているか, 理解をしているかを知ることができ, より発話がしやすくなる. 音声対話システムにおいても, システムがユーザの発話に対して適切に相槌や復唱を行うことによって, 対話のより円滑な進行が期待できる.

1.2 相槌・復唱の生成に関する研究

対話システムが適切に相槌や復唱を行うためには, その内容やタイミングを, 直前のユーザの発話から決定する必要がある. 相槌や復唱の生成に関する研究は, 相手の発話における言語情報を利用したもの, 韻律情報を利用したものとに大別される.

連絡先: 藤江 真也, 小林哲則
早稲田大学理工学部コンピュータネットワーク工学科
〒 169-8555 東京都新宿区大久保 3-4-1 55 号館 N 棟 509 号室
Tel: 03-5286-3379, Fax: 03-3205-9381

言語情報を用いるもの

Nakano らは文脈を用いた逐次理解と漸次的な発話生成により豊富なターンテイキングを行うシステムを構築している [1][2][3]. このシステムはユーザの発話の音声認識結果を早期確定し, 漸次的に言語処理や意味解析を行うことにより, ユーザに確認すべき内容を決定する. その内容を, ユーザ発話の復唱として音声合成をする.

この研究では, 発話内容を早期確定することによって適切な内容で復唱を行うことができるが, 韻律情報によるタイミング制御を行わないため, ユーザに不快感を与えるタイミングで復唱を行う可能性がある.

韻律情報を利用したもの

Ward らは低ピッチ区間が一定時間続いたあとに相槌が打たれているという規則に基づき, 相槌応答を生成するシステムを構築している [4]. 岡登らは, 対話コーパスにおいて相槌が打たれた発話の韻律特徴のテンプレートを用いて, 相槌タイミングの検出を行っており, この手法を用いた相槌生成システムも構築している [5]. また向井らは微分方程式で表現した力学系モデルを応答タイミングの制御に用いることを提案している [6]. このモデルでシステムの相槌と発話の意思を制御し, パワー／ピッチでタイミングを制御する. 小磯らは発話断片末の語彙, ピッチパターン, パワーパターンの相関を分析し, その出現確率を線形結合することにより, 発話の継続／終了を予測するモデルを提案している [7].

韻律情報を用いる研究の中でも, 表層的な言語情報まで踏み込んだ例もある. Takeuchi らは, 発話の末尾の韻律情報と末尾の言語表現中の品詞特徴を使い, タイミングを制御している [8]. この研究では, 相槌と話

者交代のタイミングの生成を，決定木によって行う方法を提案している．決定木の素性には，韻律情報と表層的な言語情報を用いている．また野口らは韻律情報，品詞情報が相槌に与える影響を，心理実験を用いて調べており，局所的な韻律情報よりも品詞情報の方が相槌の手がかりになっているとしている [9].

これらの研究では，相槌を打つかどうかやタイミングを適切に決定することが出来る．しかし，相槌をするべきか復唱をするべきか，また復唱の場合にどのような内容を復唱するべきかを決定することは出来ない．

1.3 本研究の手法

本研究では，ユーザの発話途中に適切な内容／タイミングで相槌・復唱を生成する音声対話システムを構築する．前節で述べたように，相槌・復唱の内容の決定は対話相手の発話の言語情報，即ち発話内容に大きく依存し，相槌・復唱のタイミングは発話の仕方に大きく依存すると言える．

発話途中に適切な内容で応答するためには，発話終了を待たずに音声認識結果を確定する，早期確定を行わなければならない．従来の音声認識器で早期確定を実現することは，データ構造の複雑さや信頼性の観点から困難であると言える．本研究では，近年音声認識で用いられるようになってきている有限状態トランスデューサ (FST) を用いて，相槌・復唱内容を決定する．FST は早期確定機能に優れている上，複数のトランスデューサを合成することにより任意の記号を出力することが可能なため，本研究への応用に適している．

相槌・復唱のタイミングは，韻律情報を逐次処理することにより検出する．本研究では，韻律情報としてユーザ発話の過去数ミリ秒の区間の基本周波数 (F0) とパワーを用い，スコア計算を行うことよってタイミング検出を行う．

これら二つの出力を用いて，ユーザの発話に対して適切な内容，適切なタイミングで相槌・復唱を行う音声対話システムを，ヒューマノイドロボット上に実装する．

2. 相槌・復唱の生成の要因

相槌や復唱を生成するかどうか，またその内容を決定するのは主に対話相手の発話内容と，発話形態 (韻律情報など) に依るところが大きいと考えられる．しかしながら，実際の人同士の対話では，音声認識結果や韻律情報など，単体の発話を処理することによって得られるローカルな情報の他に，タイミングや内容は，対話に参加している自分がどのような対話戦略を持っているかにも大きく依存する．

例えば「美味しいものが一食べたいな」と言われた時に，「美味しいものが一」と言われた時点での反応

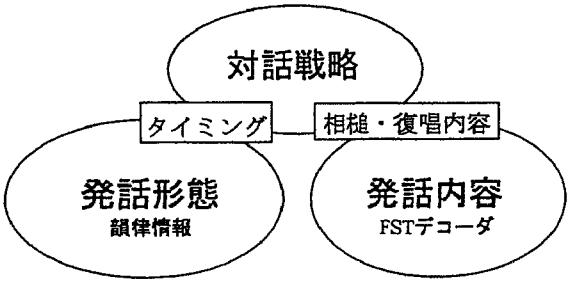


図 1: 相槌・復唱の生成を決定する 3 つの要因

が「はい」というただの相槌なのか，それとも「美味しい」という復唱なのかによって，その役割は大きく異なる．相槌の場合，その行為自体は話を聞いているということのみを表現し，基本的に相手の発話をそのまま促す行為と取れる．復唱の場合は，話を聞いているということを表現することの他に，聞き手が「美味しい」ということを正しく理解したということを確認する役割も持つが，相手の発話を促す効果は相槌に比べると弱い．また，相槌をする頻度も，少ないうちは相手の発話を促す効果があるが，多過ぎると返って相手の発話の終了を促す効果になると考えられる．

したがって，図 1 に示すように，相槌，復唱の内容やタイミングは，発話内容や発話形態によって決まるものが決定的なものであるわけではなく，対話システムが持つ対話戦略にも大きく依存すると言える．よって，相槌・復唱を行う対話システムを構成する場合には，その対話戦略も考慮する枠組が必要である．

3. FST デコーダを用いた相槌・復唱内容の早期確定

3.1 FST を用いた音声認識器

有限状態トランスデューサ (Finite State Transducer; FST) とは，有限状態オートマトンに出力を付与したものである．複数のトランスデューサを合成 (compose) することにより，望みの入出力対のトランスデューサを構成することが可能である．また，最小化 (minimization)，決定化 (determinization) 等の最適化を行うことにより，出力の早期確定機能を実現することが可能である．

大語彙連続音声認識器 (デコーダ) は，ある音素がどのような確率で音声特徴量を出力するかをモデリングした音響モデル，ある単語がどのような音素列で表現されるかを記した単語辞書，ある単語列がどのような確率で出力されるかをモデリングした言語モデルの 3 つのデータから構成される．一般的な音声認識器は音響モデルは隠れマルコフモデル (Hidden Markov Model; HMM)，言語モデルは n 個の単語で構成された単語列の出現確率である n-gram を用いてモデリングしている．

HMM や単語辞書，tri-gram 言語モデルは，いずれも

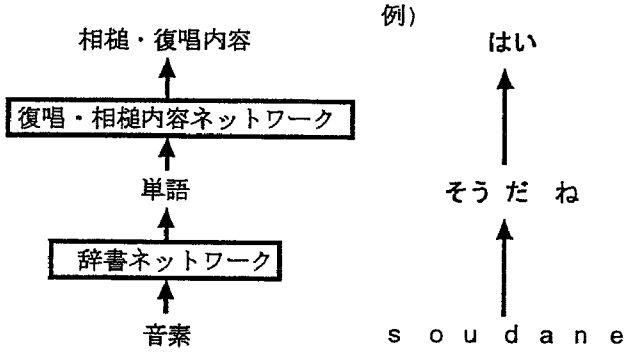


図 2: FST を用いた相槌・復唱内容の早期確定のネットワーク構成

ネットワークで表現することが可能であり, デコーディングはこのネットワーク上を遷移していくことに他ならない. これらのモデルを各々個別の有限状態トランスデューサとして実現し, 合成や最適化を行うことにより, 早期確定機能を持った単一のトランスデューサを実現することが可能である. このネットワークは, 事前に準備できる最適なネットワークであると同時に, 単一のネットワークであるため探索中の様々な情報をスタックする必要が無く, デコーダの実装を単純化することができる. このことから, 近年, 音声認識への FST の応用が注目を集めている [10][11].

3.2 相槌・復唱内容の早期確定

音声認識の最終出力は一般的に単語列であるが, 本研究では FST の早期確定機能を活かして入力文に対する相槌・復唱内容の早期確定機能を実現する.

図 2 に示したように, 音素を入力とした辞書ネットワークの出力である単語列をもとに, その単語列に相応しい相槌・復唱の内容を出力とするネットワークを構成する.

実際の音声認識では, 音響モデルとして音素を前後の音素に依存させてモデリングした tri-phone モデルを用いるため, 図 2 に示した最初の入力である音素は, tri-phone 列が入力されるとそれに対応する mono-phone 列を出力するネットワークの出力である. 辞書ネットワークは, 入力が mono-phone の音素で, 出力が単語である. このネットワークは, 音素列が入力されるとその音素列を持つ単語を出力する. また全ての単語の終了状態から初期状態の遷移を持っており, 巡回している.

なお, FST の合成, 最適化には AT&T のツール [12][13] を使用した.

3.3 入出力の抽象化

図 2 に示した相槌・復唱内容ネットワークは, 単語を入力としている. このようなネットワークを構成するためには, ネットワークを構成する前に, 対象となる文章を全て挙げておく必要がある. しかしながら, 食事にどこへ行くか等を決定するタスクにおいては, 店名や通りの名称などの固有名詞が頻繁に現れ, また, それがしばしば復唱の対象になり得る. 新しい店が対話の対象とされた時に, その店名が入り得る文章を全て生成しなおすことは, 非効率的である. そこで, 入力を単語のみでなく, 復唱の対象となるクラスに分類することによって, 新しい表現が出現したときに対応することが可能になる枠組が必要となる.

本研究では, 手動で生成した規則から, 具体的な文章を生成し, その文章群を用いてネットワークを学習する. 規則の一部を図 3 に示す. 波括弧で囲まれたものが抽象化されたクラス (後ろについているアルファベットは複数出現した場合に対処するためのインデックス), 丸括弧で囲まれたものがそれ以前の入力を受け取った場合に発するべき相槌・復唱の内容である. ただし, 丸括弧内の表現が縦線 ('|') で区切られている場合は, それらの表現がどれもふさわしいということを表している.

例えば「{形容詞 A}もの ({形容詞 A}) が食べたいな (はい)」というルールの場合, 形容詞クラスに「美味しい」という単語が入っている場合, 「美味しいもの (美味しい) が食べたいな (はい)」という具体的な文が生成され, 学習に用いられる.

```
{形容詞 A}もの ({形容詞 A}) が食べたいな (はい)
{形容詞 A}ところが (はい) いいな ({形容詞 A}ところ|はい)
{ジャンル A}が (はい) 食べたい (はい|{ジャンル A})
{店 A}も (はい|店 A) いいんだけど (はい)
```

図 3: 抽象化された表現を含む復唱・相槌内容決定ルールの一部

4. 韻律情報を用いた相槌・復唱タイミングの検出

4.1 相槌・復唱タイミングと韻律

前節で述べた, FST による相槌・復唱内容の出力は, 相槌・復唱を行うとしたときに相応しい内容を出力する. FST の早期確定機能により, 相槌・復唱を行うべきタイミングよりも早く出力されることが期待できるが, 実際にどのタイミングで相槌・復唱を行うのが最適かは決定できない.

実際に相槌・復唱を行うかどうかは, 発話者が何を発話したかよりもどのように発話したかに大きく依存すると考えられる. たとえば, 「辛くて美味しい物が食べたいな」という発話において, 「辛くて」という部分を「辛くてー」のように最後をのばして発話した場合, 「はい」や「辛い」などの相槌・復唱を行うことは自然と言える. 「辛くて美味しい」を一息に発話している場合は相槌・復唱は自然ではなく, 発話者に不快感を与える.

本研究では, F0 とパワーをもとに計算した特徴量を用いて, 相槌・復唱すべきタイミングを検出する.

4.2 対話データの分析

まず, 実際に相槌・復唱生成時の相手側の発話を収集するため, 人同士の対面対話を収録した. 5 分程度のトピックフリーの対話を 9 対話, 映像と音声両方に関して収録した. 動画アノーテショーンツール Anvil[14] を用いて, 収録した動画像全てについて各話者の相槌・復唱にタグ付けを行った.

タグ付けされた相槌・復唱の直前の対話相手の音声を切り出し, F0 抽出, ログパワーの計算を行い, 観察した. 「あー, 一日目でー」という発話の抽出例を図 4 に示す. 笑い声等の F0 抽出が困難になるケースを取り除いた結果を観察すると, 相槌・復唱時点からおおよそ 100msec から 500msec 前の区間に特徴が現れていることが分かった.

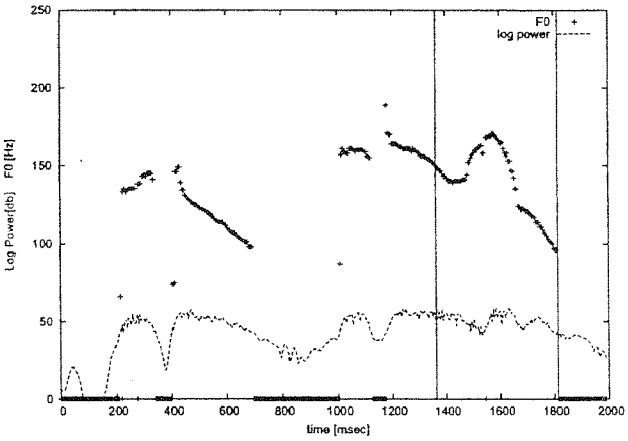


図 4: 発話の F0, ログパワー抽出例

4.3 特徴量の抽出とモデルの学習

特徴量は, 過去 450msec の区間を 150msec ずつの 3 区間に分割し, 各区間の F0, ログパワーの抽出結果を, 最小自乗法を用いて二次曲線

$$y'(t) = at^2 + bt + c \tag{1}$$

に近似した場合の, 係数 a, b と, 平均誤差

$$\frac{1}{N}\sum_i |y_i - y'(t_i)| \tag{2}$$

の三種類の値を用いた. ここで y_i は抽出された F0 またはログパワー, N はその区間で抽出された値の個数を表す. 本研究では, 音声のサンプリング周波数 16kHz, フレームシフト 80sample で分析を行ったため, 1 区間 150msec あたり N は最大で 30 となる. 各区間毎に, F0, ログパワーに関する特徴がそれぞれ 3 種類ずつ計算されるため, 合計で 18 次元の特徴量となる. 図 4 に

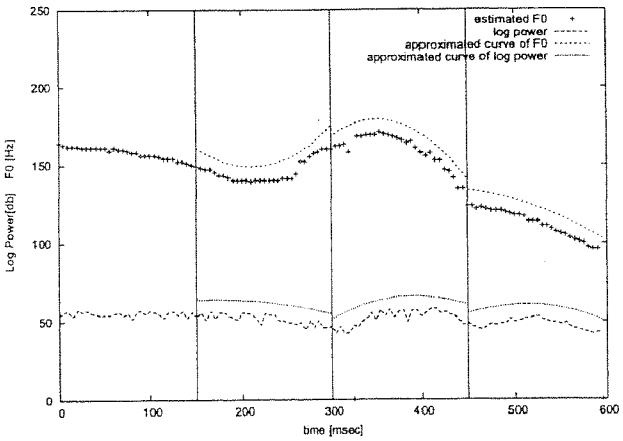


図 5: 抽出された F0 とログパワーの二次曲線近似の例

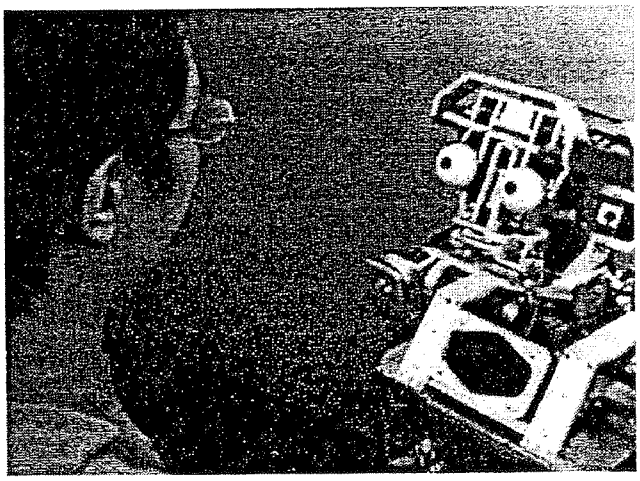


図 6: 音声対話ロボット ROBISUKE

示した区間における F0 とログパワーの二次曲線近似の結果を図 5 に示す.

前節で述べた対話収録により得られた 119 サンプルの発話についてこの特徴量を計算し, 正規分布を学習した. この正規分布が特徴量を出力する尤度がある閾値以上になった時に, 相槌・復唱をするべきタイミングと判断する.

5. 対話ロボット上での相槌・復唱機能の実現

従来から開発している音声対話システムに, 本研究で提案した FST デコーダと韻律処理を加えた. 音声対話システムは, 図 6 に示したヒューマノイドロボット ROBISUKE[15] 上に実装されている. システム構成を図 7 に示す.

ユーザが発話した音声は, 音声認識に必要な MFCC 特徴量を計算するモジュールと同時に, 本研究で提案した韻律処理を行う韻律処理モジュールに入力される. 韻律処理モジュールでは, サンプリング周波数 16kHz の音声を, フレームサイズ 1024sample, フレームシフト 80sample(5msec) で処理を行う. F0 抽出には, 後藤

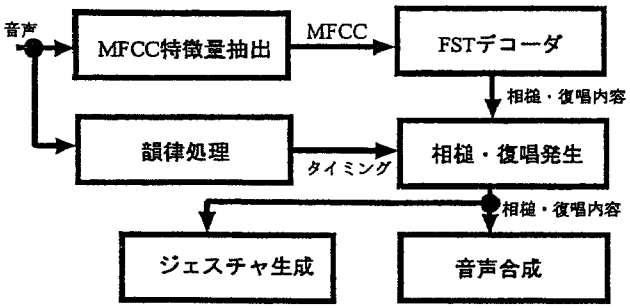


図 7: システム構成

ら [16] が提案した瞬時周波数と combfilter を用いた手法を用いている．この手法は，ESPS や wavesurfer で用いられている DP を用いた手法 [17] に比べると総合的な精度は落ちるが，単一フレームでの推定精度が高いため，リアルタイム性が問われる本システムにおいては，こちらの手法の方が好ましい．韻律処理モジュールではさらに，発話開始から 450msec 経ち，F0，パワーの値が十分揃った状態になると毎フレーム 4.3 節で述べた特徴量の計算を行い，相槌・復唱タイミングの尤度を計算する．計算した尤度が閾値を超えた場合，即座に相槌・復唱発生モジュールに信号を送る．

FST デコーダは，MFCC 特徴量を受け取りデコーディングを行う．デコーディングは仮説を展開する形で進行するため，FST に早期確定機能があったとしても，発話終了まで最終的な出力を得ることができない．そこで本システムでは，一定数の連続フレームで最もスコアの高い出力が変わらなかった場合に，それを一時的に早期確定した出力とすることとした．その後にその出力と違う出力の尤度が高いことがわかった場合は，出力のキャンセルを表す特別な記号 delete を出力し，新たに確定したものを出力する．

相槌・復唱発生モジュールは，FST デコーダから相槌・復唱内容を受け取ると，まずその内容をバッファに保存する．そしてその後，韻律処理モジュールからタイミングが出力されると，バッファに保存された内容を音声合成モジュール，ジェスチャ生成モジュールへと送信する．

音声合成モジュールは，受け取った相槌・復唱内容を発声する．ジェスチャ生成モジュールは，相槌，復唱のどちらの場合も，首を 1 回だけ短時間のうちに縦に振るという動作をロボットに行わせる．

6. まとめ

適切な相槌・復唱とは，相槌と復唱のどちらを行うか，また復唱の場合どのような内容を復唱するかという，適切な内容という側面と，ユーザが不快に感じない，自然で効果的であるための，適切なタイミングという側面の二つの側面がある．これらのどちらか一方が欠けているだけでも，相槌・復唱としては不適切になってしまう可能性がある．本研究では FST デコーダを用いた内容の早期確定，韻律処理によるタイミングの決定手法を提案し，実装した．

また，ヒューマノイドロボット ROBISUKE 上に実装し，適切な内容やタイミングで相槌・復唱を行うシステムを構築した．

対話相手の全く同じ発話に対して，タイミングさえ適切であれば，相槌をすることも復唱をすることも自然である場合が多い．しかしながら，相槌と復唱では果たす役割が大きく異なる．また，適切なタイミングであっても相槌を頻繁にうつ場合とうたない場合では，相手に与える印象は大きく異なる．これらを対話戦略に従って適切に制御することにより，より円滑で効率的な対話が可能にすることは今後の課題である．また，相槌・復唱を適切に出力することにより，対話がどれだけ自然になったか等について評価を行うことも今後の課題として挙げられる．

参考文献

[1] Nakano, M., Dohsaka, K., Miyazaki, N., Hirasawa, J., Tamoto, M., Kawamori, M., Sugiyama, A., Kawabata, T.: "Handling Rich Turn-Taking in Spoken Dialogue Systems," Proc. of Eurospeech99, pp.1167–1170 (1999).

[2] 中野 幹生, 宮崎 昇, 平沢 純一, 堂坂 浩二, 川端 豪: "多重文脈を用いた逐次的な発話理解," 情報処理研究報告, 98-SLP-22-4, pp.21–26 (1998).

[3] 堂坂 浩二, 島津 明: "タスク指向型対話における漸次的発話生成モデル," 情報処理学会論文誌, Vol.37, No.12, pp.2190–2200 (1996).

[4] Ward, N: "Prosodic Features which Cue Back-channel Responses in English and Japanese," Journal of Pragmatics 32, pp.1177–1207 (2000).

[5] 岡登 洋平, 加藤佳司, 山本 幹雄, 板橋 秀一: "韻律情報を用いた相槌の挿入," 情報処理学会論文誌, Vol.40, No.3, pp.469–478 (1999).

[6] 向井 理朗, 関 進, 中沢 正幸, 綿貫 啓子, 三吉 秀夫: "力学系モデルに基くマルチモーダル対話システムの試作," 人工知能学会全国大会 (第 13 回) 論文集, pp.109–110 (1999).

[7] 小磯 花絵, 堀内 靖雄, 土屋 俊, 市川 熹: "先行発話断片の終端部分に存在する次発話者に関する言

語的・韻律的要素について," 電子情報通信学会技術報告, NLC95-72, pp.25–30 (1996).

[8] Takeuchi, M., Kitaoka, N., Nakagawa, S.: "Timing Detection for Realtime Dialog Systems Using Prosodic and Linguistic Information," International Conference, Speech Prosody 2004, pp.529–532 (2004).

[9] 野口 広彰, 片桐 恭弘, 伝 康晴: "心理実験を用いたあいづち応答の手がかり特徴の検証," 人工知能学会研究会資料, SIG-SLUD-A002-13 (2000).

[10] 堀 貴明, 堀 智織, 南 泰浩: "有限状態トランスデューサによる音声認識・文整形・要約処理の統合," 日本音響学会講演論文集, 2-4-4 (2003).

[11] Kanthak, S., Ney, H., Riley, M., Mohri, M.: "A Comparison of Two LVR Search Optimization Techniques," In Proceedings of ICSLP 2002, pp. 1309–1312, 2002.

[12] Mohri, M., Pereira, F., Riley, M.: "Weighted Finite-State Transducers in Speech Recognition," ISCA ITRW, ASR 2000, pp.97–106 (2000).

[13] AT&T Finite State Machine Library, `http://www.research.att.com/sw/tools/fsm/`

[14] Mipp, M: "Anvil - A Generic Annotation Tool for Multimodal Dialogue," In Proceedings of of the 7th European Conference on Speech Communication and Technology, Aalborg, pp. 1367-1370 (2001)

[15] 藤江 真也, 江尻 康, 菊池 英明, 小林 哲則: "パラ言語の理解能力を有する対話ロボット," 情報処理学会研究技術報告, SLP-48, pp.13–20 (2003).

[16] 後藤 真孝, 伊藤 克亘, 速水 悟: "自然発話中の有声休止箇所のリアルタイム検出システム," 電子情報通信学会論文誌 D-II, Vol.J83-D-II, No.11, pp.2330–2340 (2000).

[17] Talkin, D: "A Robust Algorithm for Pitch Tracking(RAPT)," In Kleijin, W.B., Paliwal, K.K. eds., Speech Coding and Synthesis, pp.495–518, 1995.

3—P—4

様々な雑音環境下におけるマルチバンド型音声認識の性能評価*

◎塩田 大輔　△王 大可　大川 茂樹（千葉工大）　白井 克彦（早大理工）†

1 はじめに

音声認識技術の実用化においては，頑健性すなわち雑音や環境の変動に対する耐性が求められる．この問題に対して，人間の聴覚系と類似の原理を提案したFletcherの心理音響的根拠に基づく，部分周波数帯域を独立に処理する，マルチバンド型音声認識が提案されている[1),2),3),4)]．本稿では，このマルチバンド型音声認識において，周波数特性やSNRの異なる様々な雑音環境下での性能を比較検討したので報告する．

2 マルチバンド型音声認識

マルチバンド型音声認識は，何らかの基準に従って分割した複数の周波数帯域に対して，音響特徴量をそれぞれ独立に計算して認識に用いる手法である．例えば，入力音声が特定の帯域に集中する雑音により影響を受けている場合，従来型の音響分析手法で得られる特徴量ベクトルは，全ての要素に雑音の影響が及んでしまう．これに対して，マルチバンド型の手法を用いることにより，雑音の影響を受けている周波数帯域に対応するベクトルの要素のみに影響を封じ込め，他の帯域の情報の損失を最小限に抑えることができる．その他の根拠として，(i) 音響心理的な裏付け，(ii) 雑音に対する頑健性，(iii) 信号処理的な扱いやすさ，(iv) 並列処理のしやすさ，などが挙げられる．

音声認識の際，入力音声に雑音が含まれていない場合は，全帯域の情報をそのまま用いることは妥当であるが，これに対して入力音声が偏帯域性の雑音に影響されている場合，全帯域の情報を考慮することは，しばしば良くない効果をもたらす[3)]．

入力音声に雑音が含まれていることが事前に分かっている場合，または推測することが可能な場合，マルチバンド音声認識は有用である．しかし，我々の日常生活環境における雑音は，通常有色であり一部の周波数帯域に集中して含まれていることが多い．この性質を積極的に利用し，帯域分割により雑音を含む帯域の影響を低減させることができる．

3 特徴量レベルの再統合法

我々はこれまでに，部分周波数帯域の尤度をそれぞれ独立と仮定し，HMMの状態を単位として尤度を統合する，尤度レベルの再統合法（LC法と呼ぶ）と，各帯域に対して音響分析を独立に行い，部分帯域の特徴量ベクトルを連結してモデルを作成する，特徴レベルの再統合法（FC法と呼ぶ）の2つのアプローチを検討し，FC法がLC法よりも良い性能を与えるという結果を得た[3)]．

本研究では，FC法を利用する．認識のフロントエンドには，フィルタバンク分析及び離散コサイン変換（DCT）に基づくメルケプストラム分析を用いる．比較実験のために設計する全帯域システム（従来型の音声認識手法に相当）では，すべてのフィルタバンク係数にDCTを適用してメルケプストラム特徴量を計算する．これに対して，マルチバンド型では，フィルタバンクの出力を幾つかに分割し，それぞれの部分（部分帯域）に対してDCTを独立に適用して複数のメルケプストラム特徴量ベクトルを得る．その後，得られた各部分帯域に対するメルケプストラム・ベクトルを連結して，1つの大きなベクトルを構成する．

4 実験

音声認識実験には，ATR日本語音声データベースに含まれる4名の男性話者による日本語重要語5,240単語を用いた．このうち，3名をHMMの学習に用いて，他の1名を評価に用いた．

付加雑音は，NOISEX-92データベースに含まれる15種類の実雑音を利用し，0～20dBまで5段階に変化させたSNRで計算機上で音声データに加えた．この際，テスト用データには雑音を加えるが，学習用データには加えていない．すなわち，HMMの状態は常に雑音のない条件下で学習されている．

音響モデルは，音素毎に切り出したセグメントより学習した状態数4，混合数4の音素HMMを用い，音響分析は，12kHzで標本化された入力音声をフレーム長21.3ms，フレームシフト長5msのハミング窓で切り出し，FFT計算によりメルフィルタバンク係数を得た後，帯域毎のDCTによりMFCCに変換する．この際，各入力セグメントに対してCMN (Cepstrum Mean Normalization) 法を適用した．全帯域型では，31次のフィルタバンクを用い，特徴量ベクトルは12次のMFCC，フレームのパワー，それらのΔ特徴及び，ΔΔ特徴の計26次とし，マルチバンド型モデルでは，帯域数として2，3，4の3通りについて実験を行った．部分帯域の分割は，フィルタバンクにおける等分点をもとに決定し，MFCCの次数は各帯域のフィルタバンク数に比例させて決定した．

---

* Evaluating multi-band ASR in various noisy environments.

† by *Daisuke Shiota, Dake Wang, Shigeki Okawa (Chiba Inst. Tech.) and Katsuhiko Shirai (Waseda Univ.)*

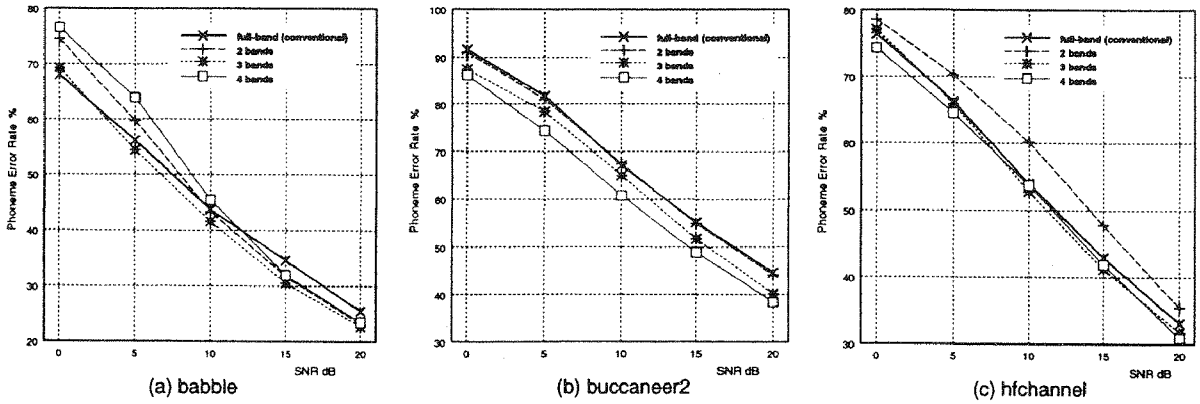


図 1: 3種類の雑音下における SNR に対する音素誤り率の変化

5 結果と考察

図1は，(a)babble, (b)baccaneer2, (c)hfchannel の各雑音下において，SNR に対する音素誤り率の変化を示す．数値が小さい程，認識性能が高いことを表している．雑音の種類により，最適な帯域分割数は異なるが，(b), (c) では，SNR の変化に対して各システムの性能が一様に変化し，マルチバンド型システムが全帯域システムよりも良い性能を与えている．これに対し，(a) では SNR の変化と共に良い性能を与えるシステムが入れ替わり，低い SNR (0dB) では全帯域システムが最も良い性能を与えている．

図2に，15種類の雑音に対して SNR=20, 10, 0 の場合におけるマルチバンド型システムの性能を，全帯域システムに対する相対的な音素認識率（改善率）として示す．babble のように SNR が小さくなるにつれて認識率が低下する場合と，baccaneer2 のようにいずれの SNR でもマルチバンド型システムが有用である場合などが見られた．

以上の結果より，一般に比較的高い SNR においては，雑音の周波数特性に応じてマルチバンド型システムを用いることにより認識性能を大きく改善できるが，SNR が低くなると効果が小さくなる．ただし，一部の雑音環境においては，特定の帯域分割モデルを用いることにより低 SNR でも依然高い性能が保持できる．

今後は，周波数帯域の分割方法と SNR の関係について検討を行う予定である．

文　献

1) H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," *Proc. ICSLP*, 1996.

2) S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," *Proc. ICASSP*, 1997.

3) S. Okawa, E. Bocchieri and A. Potamianos, "Multi-band speech recognition in noisy environments," *Proc. ICASSP*, 1998.

4) C. Cerisara and D. Fohr, "Multi-band automatic speech recognition," *Computer Speech and Language*, 15, 2001.

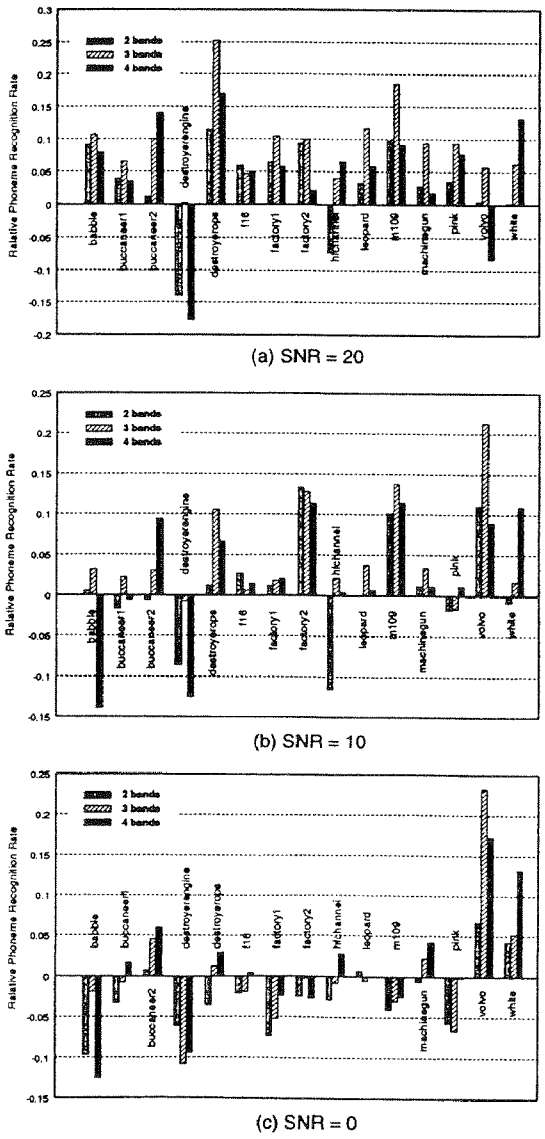


図 2: 様々な雑音及び SNR におけるマルチバンド型音声認識の効果

2C3-03

基本動作制御モデル構築のための歩行動作の分析・生成

An Analysis of Human Walk for Basic Motion Control Model

森野 岳宏*1 Takehiro MORINO　大隈 健史*1 Takeshi OKUMA　長船 大毅*1 Daiki OSAFUNE　村上 真*2 Makoto MURAKAMI　白井 克彦*1 Katsuhiko SHIRAI

*1早稲田大学理工学部 Waseda University　*2東洋大学工学部 Toyo University

In order to simplify the production of the human motion animation, it is required the model which describes human motion as a few rule set. We have analyzed human motion data obtained by motion capture system mainly in interaction with each joint motion. In the result, human walk consists of some motions hierarchically; one motion in a level is independent of another. And each independent motion can be generated from some "key joints" in our experiment.

1. はじめに

現在の CG 作成の現場で用いられている人物動作生成手法（モーションキャプチャ, キーフレームアニメーションなど）を個人の映像コンテンツ制作に適用する場合, コスト高やスキル不足のような問題が考えられる. そこで, これらの問題を解決する為に, 従来とは異なった手法を提案する. それはコンピュータ内の俳優（デジタルアクタ）に対して, 人間が言葉で指示を出し, 頭の中で描いている動作を行うまで演技指導をするという手法である. これにより, 人間同士のコミュニケーションに近い直感的なインターフェースを用いて, 感情豊かな動作を含む映像コンテンツを制作する事が可能となる.

デジタルアクタに対して演技指導を行う為には, あらかじめ動作に応じて, その拘束条件と自由度をモデル化し, それによって自然かつ多様な動作を生成・制御しなければならない. 動作生成におけるモデル構築には, 過去では関数近似を用いたもの [1][2], または人間の筋肉の収縮運動を考慮したもの [3] などがある. 従来の研究は動作における人間の可動部位の運動をそれぞれ独立に考えており, 可動部位の連動性に言及したものはない. 人間の動作は個々の可動部位を連動させて制御される事から, 本研究では, 各部位の空間的及び時間的連動性を考慮し, その連動性の高い部位をまとめて動作生成・制御可能なモデルの構築を試みる. このモデルにより, 動作を制御する際の操作部位を少なくする事ができ, 部位の連動性をもとにリアルな動作表現が可能となる.

2. 基本動作制御モデル

2.1 定義

基本動作制御モデルとは感情を反映させる基本動作を定義し, その動作を制御可能にしたモデルである. 基本動作制御モデルは, 動作における関節の連動性のルールセットからなる. 全ての動作というのは, 個々の関節を連動させて制御されるものであり, その連動性のルールセットによって動作の定義, 制御が可能となる. また, 人物動作は人それぞれの個性が反映されるが, モーションキャプチャによって収録した多人数の動作データを用いて動作をモデル化すれば, 動作者の個性を吸収した動作を生成する事が可能となる.

連絡先: 森野岳宏, 早稲田大学理工学部, 東京都新宿区大久保 3-4-1, 電話番号 03-3200-1399,Fax 番号 03-3200-1399,e-mail:takehiro-morino@suou.waseda.jp

2.2 モデル構築法

本研究では, 基本動作の中でも歩行動作に特化した動作制御モデルの構築を試みる.

まず, 歩行動作の際に, 他の関節と連動性が高いと推測される関節をキー関節として定義する. 動作データで定義されている人体モデルにおいて, 身体関節は全部で 19 個あり, その中から 4 つの関節をキー関節として設定した.

次に, 説明変数にキー関節の動作データ (関節の 3 軸方向回転角度の時系列データ) を前後に数フレームシフトしたデータを複数用い, 目的変数にはキー関節以外の関節の動作データを使用して, 前者のデータから後者のデータを生成する関数群を重回帰分析によって導出する. 説明変数にシフトデータを用いたのは, キー関節とそれ以外の関節の動作データの遅延を考慮するためである. この関数群を動作生成関数群と定義し, 関節間の連動性のルールセットとする.

この動作生成関数群によって, キー関節の新たな動作データを与えれば, 他の関節の新たな動作データを生成する事ができ, 動作全体を生成する事が可能となる. 設定したキー関節と, その関節から他の関節の動作生成の様子は図 1 に示す.

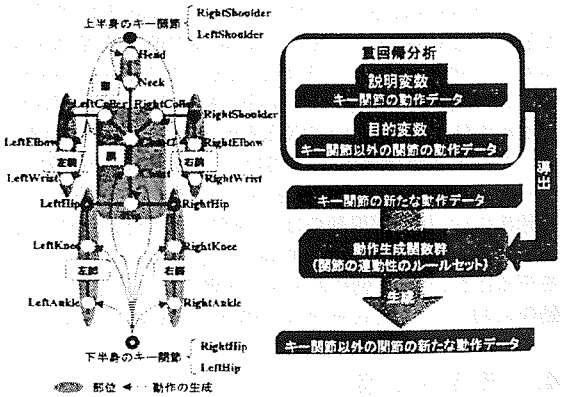


図 1: モデル構築法概念図

3. モデル評価実験

3.1 実験手法

ジャックナイフ法 [4] を用いて, 構築したモデルの評価実験を行う. 31 人の動作データから 1 人を除いた 30 人分のデー

タセットでモデルを構築し，除いた 1 人のデータのキー関節の動作データをこのモデルに適用して動作を生成する．そして，除くデータを替えながら，この作業を 5 回繰り返し 5 つの動作データを生成する．評価の指標としては，生成した動作データと元動作データとの相関係数を用いる．この相関係数を 5 つの生成された動作データについて，各関節の 3 軸方向毎に平均値を算出した．また，この際にキー関節の動作データのシフトフレーム数を前後 18・9・0 フレームの 3 種類を試す．

3.2 結果例

表 1 に，実験結果の 1 つである左脚部位における相関係数表を示す．表 1 の関節動作名の項目において，関節名_3 軸方向とあるが，これはその関節の 3 軸方向の回転動作を意味している．

表 1: 相関係数表－左脚部位

合計シフトフレーム数	動作名（関節_3 軸方向）	3 軸方向毎の平均	3 軸方向毎の標準偏差
前後 18rag	LeftKnee_z	0.7437	0.1931
	LeftKnee_x	0.9578	0.0194
	LeftKnee_y	0.6868	0.1717
	LeftAnkle_z	0.7089	0.1045
	LeftAnkle_x	0.8327	0.0217
	LeftAnkle_y	0.2405	0.4803
前後 9rag	LeftKnee_z	0.7302	0.1381
	LeftKnee_x	0.9614	0.0180
	LeftKnee_y	0.6352	0.2973
	LeftAnkle_z	0.6947	0.1252
	LeftAnkle_x	0.8609	0.0469
	LeftAnkle_y	0.2658	0.4632
0rag	LeftKnee_z	0.4807	0.3003
	LeftKnee_x	0.8146	0.1364
	LeftKnee_y	0.6150	0.2632
	LeftAnkle_z	0.4713	0.3610
	LeftAnkle_x	0.7784	0.0903
	LeftAnkle_y	0.1071	0.2441

3.3 評価実験に対する考察

シフトフレーム数を前後 9 フレームにすれば，動作の再現性が高くなる事が確認された．ここから，キー関節の動きと他の関節の動きが前後 9 フレームの範囲内で相関性を持っていると推測される．

また，キー関節の動きから他の全ての関節の動きが生成されていない事も確認された．動作の再現性の高い関節動作は，キー関節との連動性が高く，動作の再現性の低い関節動作は，キー関節との連動性が低いと推測される．ここで，平均相関係数が 0.5 以上の関節動作を（キー関節に対する）従属関節動作とし，平均相関係数が 0.5 以下の関節動作を（同）独立関節動作とする．すると，全関節動作（19 関節× 3 軸方向＝ 57 動作）からキー関節動作（4 関節× 3 軸方向=12 動作）を除いた 45 の関節動作の内，従属関節動作は 22 動作，独立関節動作は 23 動作あった．これら独立関節動作は，設定したキー関節で構築したモデルにおいては，動作の生成が困難である．従って，独立関節動作において，新たにモデルを構築する必要がある．

4. モデル改良

4.1 独立関節動作のモデル構築

まず，独立関節動作において，連動性があると推測される関節動作を胴部位，頭部位，左腕部位，右腕部位，と 4 つのグループに分け，そのグループ内でそれぞれキー関節動作を設定する．次に，そのキー関節動作からそのグループ内のキー関節動作以外の関節動作を生成する動作生成関数群を重回帰分析によって導出する．設定したキー関節動作と，その動作から生成する関節動作の対応を図 2 に示す．

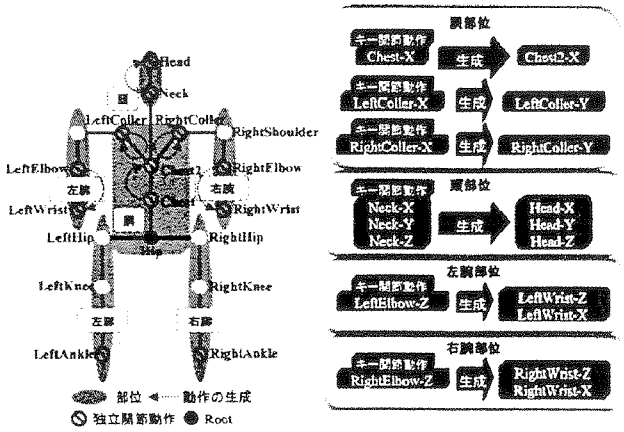


図 2: 独立関節動作のモデル構築法概念図

4.2 再度モデル評価実験及び考察

独立関節動作において構築したモデルの評価実験を再度行う．実験手法は 3.1 節の手法と同様である．表 2 に実験結果を示す．

表 2: 相関係数表－独立関節動作

部位名	動作名（関節_3 軸方向）	平均（モデル改良後）	標準偏差（モデル改良後）
左腕部位	LeftWrist_z	0.2749	0.5044
	LeftWrist_x	0.4325	0.2285
右腕部位	RightWrist_z	0.0972	0.6321
	RightWrist_x	0.1668	0.4412
胴部位	Chest2_x	0.9991	0.0004
	LeftCollar_y	0.8212	0.2059
	RightCollar_y	0.9070	0.0937
頭部位	Head_z	0.4595	0.3610
	Head_x	0.1631	0.5616
	Head_y	0.9106	0.0644

3.3 節と同様に，相関係数の閾値を 0.5 として，関節動作を（キー関節動作に対する）従属関節動作，独立関節動作に分類すると，3.3 節も含め，関節動作を 2 段階で分類した事になる．2 段階目での分類で独立関節動作となった関節動作は，その関節動作のみでモデルを構築する事は動作数からも困難であり，動作生成の際には独自にパラメータを与える事にする．よって，全関節動作（57 動作）を,1 段階目で設定したキー関節動作（4 関節× 3 軸方向=12 動作）,2 段階目で設定したキー関節動作（8 動作）及び最終的に独立関節動作となった関節動作（11 動作）の合計 31 の関節動作で生成する事になる（図 3 参照）．

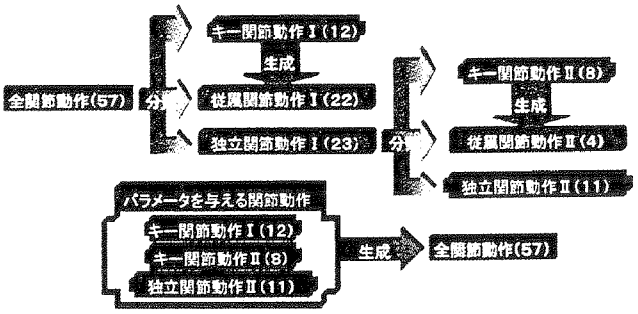


図 3: 関節動作の分類図

最終的に独立関節動作となった関節動作というのは，主に Head,Ankle, 及び Wrist などの身体の末端に位置する関節に

見られる．ここから，歩行動作においては，身体の末端に位置する関節はそれ以外の関節とは独立して動いている事が分かる．一方，身体の末端に位置する関節以外の関節は互いに連動して動いている事が分かる．

5. まとめと今後の予定

2段階にわたる重回帰分析によるモデル構築により，31の関節動作で全関節動作(57動作)を生成可能なモデルを構築した．この事から，歩行動作は幾つかのレベルの独立な動作として構成されており，それぞれの動作は幾つかのキー関節動作から生成可能である事が確認された．

今後は，動作を生成可能なだけではなく制御可能なモデルの構築が必要である．さらに，構築したモデルを実際にデジタルアクタに実装し，その利便性，有効性を確かめなければならない．また，基本動作制御モデルの構築の為には，歩行動作以外の動作についても，同様にモデルを構築する必要がある．

参考文献

[1] 鵜沼宗利，武内良三，"コンピュータアニメーションにおける感情を伴った人間の歩行動作の生成方法，"電子情報通信学会論文誌,D-II,Vol.J76-D-II, No.8,pp.1822-1831

[2] 稲田和彦，佐藤潤一，中川雅通，浜田正弘，前原文雄，"関数近似と逆運動学を用いた人物動作の生成方法，"画像電子学会研究予稿,94-01-02

[3] 西田論司，杉浦彰彦，"アクションユニットを用いた人間の動作の分析・作成，"豊橋技術科学大学

[4] 浅香緑，女部田武史，井上直，岡澤俊士，後藤滋樹，"不正侵入の痕跡と判別分析によるリモートアタックの検出法，"電子情報通信学会論文誌,B,Vol.J85-B,No1,pp60-74

2C3-04

モーションキャプチャデータを用いた歩行動作の分析と制御

Analysis and Control of Human Walk Using Motion Capture Data

長船 大毅*1 Daiki OSAFUNE　大隈 健史*1 Takeshi OKUMA　森野 岳宏*1 Takehiro MORINO　村上 真*2 Makoto MURAKAMI　白井 克彦*1 Katsuhiko SHIRAI

*1早稲田大学理工学部 Waseda University　*2東洋大学工学部 Toyo University

This research aims at development of "Expressing Emotion Model" can create animation easily by directing motions of a CG actor from conversations, gestures, etc. In order to control motions of the actor more simply, we extract parts that generally have same changes from a large number of people's data of motion of walk, and controll motion of walk by changing only the parts.

1. はじめに

近年、CG 技術の進歩により、幅広い分野で CG アニメーションが用いられている。しかし現状では、個人での CG アニメーション作成は容易ではなく、簡単にアニメーション作成を行なえるようなシステムの構築が求められている [1][2]。

本研究は擬人化された CG のアクタに音声やテキストにより動作指示を与えることで、動作指示者の思考を反映した動作アニメーションを生成するディジタルアクタ制御システムの構築を目的とする。ディジタルアクタは、各関節の角度パラメータを変化させることにより動作を制御する。しかし、全関節に対して独立に動作制御を行なうことは現実的には不可能であるため、少ない指示で簡単に動作制御することが重要となる。そこで、モーションキャプチャによって収録した動作データを用いて、動作を変化させた場合の特徴を分析する。特徴は各関節毎の角度パラメータの変化量から算出され、多くの被験者に共通した変化量が算出された部位をその動作の特徴部位とする。そして、その特徴部位のみを制御し、少ない動作指示からの動作制御を可能にする。

2. 特徴部位の分析手法

2.1 動作データ

使用データは、各関節についてのローカルな座標系での回転角度情報（19 関節 ×3 方向）、及び体の中心の位置情報（1 関節 ×3 方向）の時系列データ（計 60 パラメータ × フレーム数）により動作を表現している。ローカル座標系の回転方向は、人に対し横方向軸の回転を X、上方向軸の回転を Y、前方向軸の回転を Z とする。ただし、分析には位置情報は不要なので、それを除く 57 パラメータを用いる。

2.2 分析内容

本研究では、日常的な動作で且つ、規則性のある周期的な動作であるという点から、歩行動作について分析を行なう。特に本稿では、歩行速度を変化させる際に起こる特徴について分析する。具体的には大学生男子 20 人の各々について収録した「普通に歩く」「ゆっくり歩く」「速く歩く」の 3 データを用いる。そして、「普通に歩く」の動作データを基本の動作とし、それに「ゆっくり歩く」及び「速く歩く」を比較することで、動作を変化させた際の特徴を抽出する。

連絡先: 長船 大毅, 早稲田大学理工学部,
osahune@shirai.info.waseda.ac.jp

動作データは各関節の回転角度により定義されており、各データは離散的な波形データ $f(x_n), (n = 0, 1, ..., N-1)$ として扱うことが可能であり、その回転角度の波形は以下の 3 種から構成されると考えられる。

姿勢 直流成分 $\frac{1}{N}\sum_{n=0}^{N-1} f(x_n)$

大きさ 振幅 A

速さ 振動数 f

そこで、歩行動作における上記の 3 種の構成要素の変化について分析する。特徴の表れ方は、姿勢・大きさについては各部位毎に、速さについては全部位共通に特徴抽出を行なう。

2.3 特徴の抽出（姿勢・大きさ）

動作を変化させた際に生じる各部位の姿勢及び大きさの変化量として、変化の度合い K_{ij} を用いる。i を各部位（19 関節 ×3 方向）、j を各人 $(j = 0, 1, ..., J-1)$ とし、「普通に歩く」の各パラメータの値を G_{ij}、「ゆっくり歩く」の各パラメータの値を S_{ij}、「速く歩く」の各パラメータの値を F_{ij} とすると変化の度合いは以下のように表される。

$$K_{ij}(\%) = \frac{(S_{ij} or F_{ij}) - G_{ij}}{G_{ij}} \times 100 \quad (1)$$

さらに、動作者の個人性によらない特徴を導くために変動係数 CV を用いる。CV は平均 μ、標準偏差 σ から、以下のように表される。

$$CV_i(\%) = \frac{\sigma_i}{\mu_i} \times 100 \quad (2)$$

ただし、

$$\mu_i = \frac{1}{J}\sum_{j=0}^{J-1} K_{ij} \quad (3)$$

$$\sigma_i = \sqrt{\frac{1}{J}\sum_{j=0}^{J-1}(K_{ij} - \mu_i)^2} \quad (4)$$

CV は、値の小さいもの程データの分散が小さいことを示すので、CV の小さいものを選出する。選出には閾値を用い、絶対値が 250(%) 以下のものを特徴部位とした。

以上より一般的に同様の変化をする特徴を抽出する。そしてその部位について K_{ij} を分析し、選出された特徴部位のみを制御することが、本研究での動作制御となる。

2.4 特徴の抽出（速さ）

速さの特徴抽出も式(1)によって K_j を算出する（ただし、速さは全部位において共通であるので、i は不要）。これにより、「普通に歩く」から「ゆっくり歩く」及び「速く歩く」へと変化した場合の周波数 f の変化量が算出可能である。

3. 分析結果

「ゆっくり歩く」の結果

表1のように、大きさについては尻・肩・肘・首のX回転、腰・足首・肩・手首・首・頭のY回転が個人差の少ない、共通の変化を持つ部位であることが確認された。姿勢については CV の小さいパラメータが存在しなかった。したがって、以上の部位を「ゆっくり歩く」における特徴部位とする。

表 1: 20 人におけるゆっくり歩くの特徴（大きさ）

	平均 [%]	標準偏差 [%]	変動係数 [%]
腰 Y	29.2	50.8	173.8
左尻 X	-7.1	14.6	-204.3
右尻 X	-9.4	18.4	-195.5
左足首 Y	34.5	63.5	183.9
右足首 Y	23.2	42.2	219.5
左肩 X	-16.6	22.1	-132.6
右肩 X	-14.1	25.3	-179.5
左肩 Y	-25.4	27.6	-108.6
右肩 Y	-21.5	32.6	-151.5
左肘 X	-19.4	33.2	-171.5
右肘 X	-29.5	25.7	-87.1
左手首 Y	-17.6	36.1	-205.3
右手首 Y	-14.6	27.7	-189.2
首 X	34.6	69.9	202.2
首 Y	38.2	87.5	229.0
頭 Y	64.3	149.3	232.1

「速く歩く」の結果

表2、3のように、姿勢については膝・胸のX回転が、大きさについては尻・足首・肘のX回転と肩のY回転が、共通の変化を持つ部位であることが確認された。したがって、以上の部位を「速く歩く」における特徴部位とする。

表 2: 20 人における速く歩くの特徴（姿勢）

	平均 [%]	標準偏差 [%]	変動係数 [%]
左膝 X	19.4	20.2	104.1
右膝 X	16.4	21.9	133.4
上胸 X	22.2	49.7	223.5
下胸 X	22.9	50.1	218.8

表 3: 20 人における速く歩くの特徴（大きさ）

	平均 [%]	標準偏差 [%]	変動係数 [%]
左尻 X	8.2	16.8	204.5
右尻 X	5.9	11.8	201.4
左足首 X	-8.2	17.2	-209.3
右足首 X	-13.3	13.9	-104.3
左肩 Y	14.5	34.4	238.0
右肩 Y	22.4	35.8	160.3
左肘 X	25.5	48.0	188.2
右肘 X	34.1	54.1	158.8

4. 動作制御実験

4.1 実験

ある一人の「普通に歩く」の動作データについて特徴部位に当たるパラメータのみを変化させることで、「速く歩く」と「ゆっくり歩く」の動作制御を試みる。制御方法は以下に示す。

大きさの制御 振幅 A を変化
姿勢の制御 直流成分 $\frac{1}{N}\sum_{n=0}^{N-1} f(x_n)$ を変化
速さの制御 振動数 f を変化

4.2 実験結果

以下に、収録した歩行動作と動作制御により生成した歩行動作の連続画像を示す。この収録データはある一人の歩行動作のデータであり、開始フレームは全て右尻と左尻のX方向の値が最も近い、つまり一番両腿が重なって見える姿勢のフレームに設定した。連続画像は5フレームおきに撮影した。

上から、収録した「普通に歩く」、収録した「ゆっくり歩く」、「普通に歩く」に「ゆっくり歩く」の特徴を用い制御した「ゆっくり歩く」、収録した「速く歩く」、「普通に歩く」に「速く歩く」の特徴を用い制御した「速く歩く」、となっている。

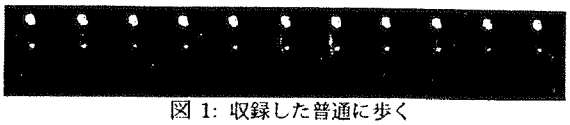


図 1: 収録した普通に歩く

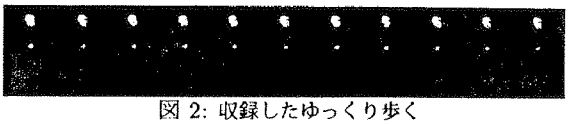


図 2: 収録したゆっくり歩く

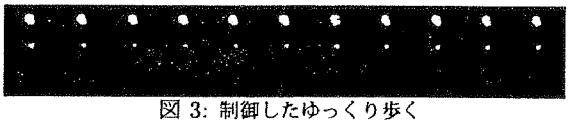


図 3: 制御したゆっくり歩く

図 4: 収録した速く歩く

図 5: 制御した速く歩く

5. まとめと課題

本稿では、まず20人のデータから「普通に歩く」と「ゆっくり歩く」及び「速く歩く」の変化の度合い K_{ij} を算出した。次に変動係数 CV を用いてその K_{ij} を20人分比較することで、特徴部位を決定した。そしてその特徴部位での平均的な特徴の表れ方を示した。最終的に、抽出した特徴を用いて基本となる「普通に歩く」の波形を変化させ、「ゆっくり歩く」「速く歩く」の動作を制御した。しかし、2種間の動作の違いから特徴を抽出し動作制御しただけなので、制御により得られた動作と収録した動作の類似度を示すような評価が必要となる。

本研究は平均値から抽出した一般的な特徴量のみに注目した分析である。したがって、変化の度合いの分散や最大値・最小値から特徴部位内での個人性も抽出できれば、特徴の内容の発展に繋がる。さらに、本手法は汎用性があり、走る動作やジャンプのような違う動作についての分析も可能である。そのように基本動作の幅を広げることも、本研究の発展に繋がる。

上記の様にある動きを制御するための特徴量がより幅広く算出できれば、感情表出モデルに動作指示をするためのインタフェースの機能向上が可能となる。そして、その特徴を考慮したモデルを作成することにより、基本動作制御モデルから容易に動作を生成することも可能となる。

参考文献

[1] 鵜沼宗利　竹内良三, "コンピュータアニメーションにおける感情を伴った人間の歩行動作の生成法"
電子情報通信学会,J76-D-II,No.8,pp.1822-1831

[2] 河崎雷太　北村喜文　岸野文郎, "スポーツにおける運動の個性の抽出と異なる体型への適用"
日本バーチャルリアリティ学会,Vol.7,No.2,pp.267-274

F-029

波形重畳型音声合成における心理的態度の再現性評価
Reproducibility evaluation of mental attitude in concatenative speech synthesis

大久保 雅史[†] 望月 亮[†‡] 蓑輪 利光[‡] 小林 哲則[†]
Tadashi Okubo Ryo Mochizuki Toshimitsu Minowa Tetsunori Kobayashi

1. はじめに

PSOLA[1] 型の音声合成器において，発話者の心的態度を韻律の制御のみによってどの程度表現できるかについて検討した。

人間同士の対話において，発話には言語情報に加え心的態度に代表されるパラ言語情報が含まれ，それらが総合的に作用して発話者の意図を伝えている。高度な対話システムを実現するためには，音声合成器もまた言語情報のみならず発話者の心的態度を表現する必要がある。

合成音声によって心的態度を表現するためには，一般には分節的特徴と韻律特徴の双方を制御する必要があることが予想されるが，現在規則合成の代表的手法であるPSOLA 型の音声合成器においては，分節的特徴を変形することは難しく，可能な限りこれを抑えることが望まれる。効率的な規則合成器を作るためには，分節的特徴の制御が不可避となる条件について明らかにする必要がある。本研究ではこれを知るための手始めとして，原音声の分析の結果得られる韻律パターンを用いるとき，分節的特徴の変更を伴わないでどの程度原音の心的態度を再現できるかについて検討する。

2. 心的態度の表現

本研究では，心的態度の表現空間として，Russell[2] の提案する「快－不快」,「覚醒—不覚醒」の 2 軸を用いる。

従来音声合成における，パラ言語情報の付与に関しては，怒り，悲しみ，といった感情の扱いが中心であった[3][4]。しかしながら，実際の対話の場面で言語に加えて必要となる情報としては，相手発話に対する注目度・共感度，自発話に対する確信度，あるいは状態に対する満足度といった，対話状況に対する発話者の内的な評価であることが多く，それらは，基本感情とは必ずしも整合しない。一方，Russell の軸は，こういった内的評価と完全に一致するものではないものの，それぞれに関係を与えることができ，我々の目的には適している。

3. 評価実験

3.1 実験方法

発話内に４種類のフレーズを含む様々な態度を故意に含ませた音声を録音し、それぞれの音声に手動で音素ラベリングを行った。この音素ラベルに基づき音素時間長、音素内４点における基本周波数、パワーの韻律情報を抽出した。これらの自然な韻律情報を用いて、合成音声を作成した。合成音声作成の流れを図１に示す。

合成に使用した音声素片は、音韻バランスを考慮して約３時間の平静な読み上げによる音声コーパスを使用した。

†早稲田大学理工学部, Sch. of Sci. and Eng., Waseda Univ.
‡松下電器産業 (株), Matsushita Electric Ind. Co., Ltd

４種類の発話それぞれについて、原音声から１つ最も平静な状態の読み上げと思われる音声を選び、そのフレーズについての基準の音声とした。またすべての原音声、合成音声を評価対象音声とした。ランダムに評価対象音声を選択し、その発話の基準音声、評価対象音声という順に９人の評価者に聞かせた。

基準音声に対する評価対象音声の態度を、「快-不快」、「覚醒-不覚醒」の二軸において各軸それぞれ５段階で評価するよう指示した。

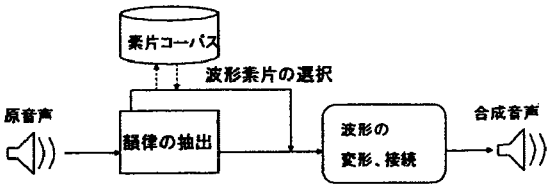


図 1: 合成音声作成の流れ

3.2 音声試料

発話内に自然に心的な態度が含まれるよう、応答の発話に決まったフレーズが含まれるような一対の対話文を複数作成した。フレーズは、対話の応答に現れるもの、フレーズの長さのバリエーションという観点から表２に示す４種類を選び、対話文は各フレーズにおいて、様々な態度が現れるような対話文を複数作成した。収録テキストの例を表１に示す。応答の発話の話者は固定として二人のプロの声優に読ませ録音した。

応答の発話の、決まったフレーズを含む部分を切り出して原音声試料とした。試聴実験に使用した音声試料を表１に示す。

表 1: 音声試料

発話内容	原音声の数
「自然のままの方がいい。」	14
「それは違うでしょ。」	14
「そうです。」	13
「つまりね。」	13

表 2: 収録テキストの例 (つまりね)

指定態度	発話内容
不快／覚醒	「ど、どういうことなの？(怒り)」 「つまりね、あんたがいけないの！」

3.3 実験結果

「快-不快」、「覚醒-不覚醒」それぞれに関して、各音声から合成音声を作成したときの、原音声の評価値の平均値に対する合成音声の評価値の平均値をそれぞれ図２、図３に示す。また、図中の直線は、評価値の平均の一次近似直線を示す。

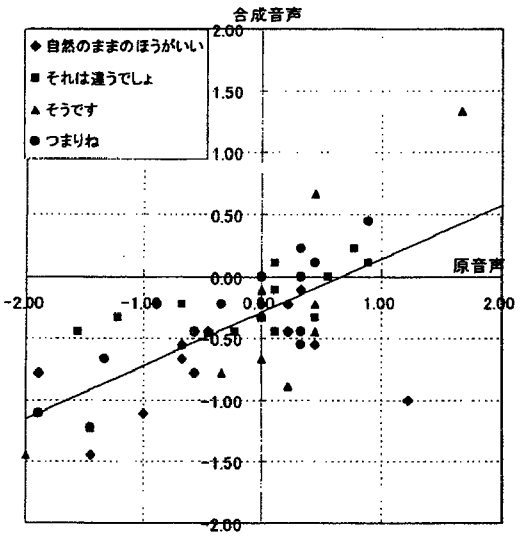


図 2: 快-不快に関する評価の変化

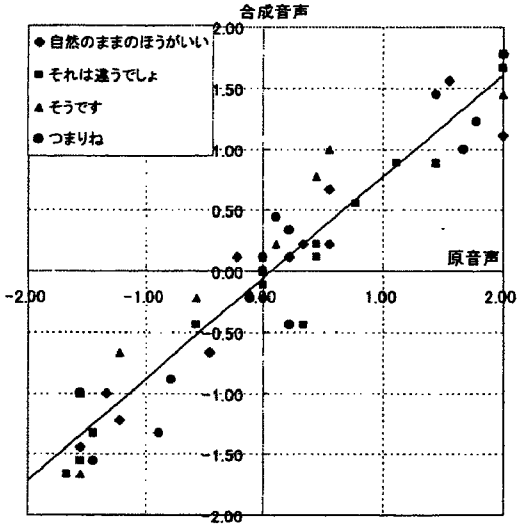


図 3: 覚醒-不覚醒に関する評価の変化

実験結果より、発話文の違いや、基準音声の違いによる評価への影響は多少あるものの、全体的にほぼ同じような傾向を示したと言える。原音声での「覚醒-不覚醒」の評価値は、その韻律で作成した合成音声においてもよく再現されているが「快-不快」の評価値は合成音声においては弱まっている。

図 4 に原音声の快、覚醒の評価値と韻律パラメータとの関係を示す。これらより、「快-不快」の評価値には「覚醒-不覚醒」ほどの韻律の影響はないものと考えられる。

これらより、「快-不快」の合成音声は、韻律情報のみから作成することが難しいと言える。

4. むすび

心的態度を表現できるような高度な対話システムを作ることを目的とし、PSOLA 型の音声合成器において、韻律情報のみから心的態度をどの程度表現可能か検討を行った。心的態度の表現尺度として、「快-不快」「覚醒-不覚醒」の二軸を用いた。

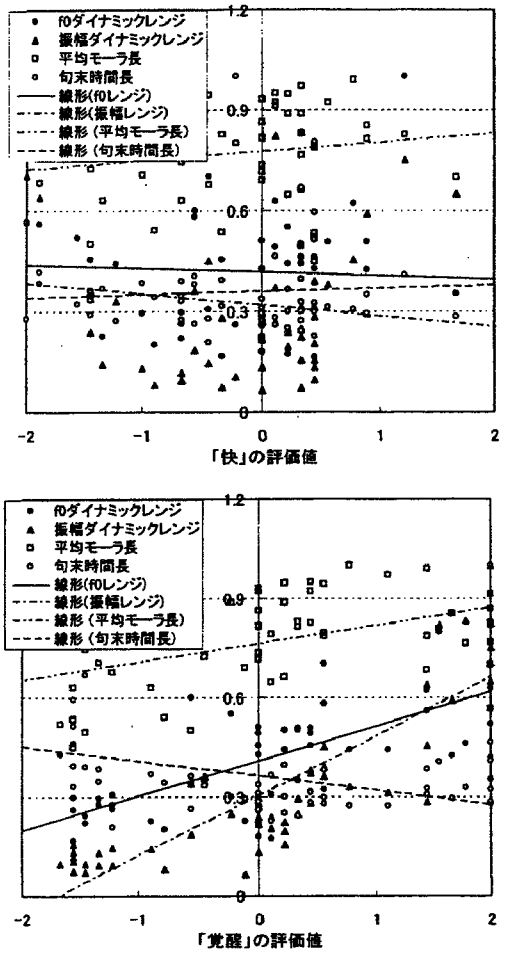


図 4: 各評価値と韻律パラメータとの関係

様々な態度を含む音声から韻律を抽出して合成音声を作成し、原音声と合成音声を試聴実験により二次元平面上に評価させ、どのような態度がどの程度合成音声において再現出来るか試聴実験を行った。

結果として合成音声においては、「覚醒-不覚醒」については十分表現可能だが、「快-不快」については表現が薄らいでしまうというデータが得られた。これより、「快-不快」の表現のためには、それぞれの態度に特化した素片の作成が必要であると考えられる。今後それらの素片を使用した上での実験も行う必要がある。

参考文献

[1] Charpentier F., Moulines E. "Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", Pros. Eurospeech '89,1989.

[2] Russell J.A. "A circumplex model of affect.", Journal of Personality and Social Psychology,39,1161-1178,1980.

[3] Y.Kitahara,Y.Tohkura "Prosodic Control to Express Emotions for Man-Machine Speech Interaction.",IEICE Trans. Fundamentals,E75-A,No.2,155-163,1992.2.

[4] Iida A. Campbell N. Iga S. Higuchi F. and Yasmura M. "A Speech Synthesis System with emotion with assisting communication.", In Proceeding of the ISCA Workgroup on Speech and Emotion,167-172,2000.

パラ言語情報を用いた音声対話システム*

◎藤江 真也† △八木大三† 菊池 英明‡ 小林哲則†
(†早大・理工 ‡早大・人間科学)

1 はじめに

パラ言語の理解機能を持たせることにより自然な対話を行うことのできる音声対話システムについて検討する.

従来の対話システムの多くは, 人間とシステムが発話をやりとりする際に, 主にその発話が持つ言語的な情報を手がかりとして対話を進めるものが多い. 人間同士の対話では, 言語的な情報だけでなく, 視覚的, 聴覚的に言語的な情報を補助する情報を用いて, 円滑に対話を進行していると言える. ここで言語的な情報を補助する情報をパラ言語情報と呼ぶ.

我々はこれまでに, 対話中の視覚的なパラ言語情報として, うなずき等の頭部ジェスチャの認識手法を提案した [1]. 本研究では, 聴覚的なパラ言語情報として, 発話に含まれる韻律情報から, 発話が肯定的であるか否定的であるかを識別する手法を検討した. また, これらのパラ言語情報を音声対話システムに応用し, その有効性を確認した.

2 対象とするタスク

本研究で扱う音声対話は, ある事柄に関して決断を迷っているユーザに対して, ユーザの要求を聞き出し, 提案を行う対話を想定している. 本研究では, このような対話を相談型対話と呼ぶ. 今回は, 昼食に何を食べるか, またどの店へ行くかを迷っているユーザの相談に答えるタスクを対象としている.

このような対話にどのような特徴が現れるかを観察するため, 事前に人同士の模擬対話を収録した. この観察によると, システムの提案に対するユーザの応答に関して, 言語的な情報だけでは得られない肯定的な態度, 否定的な態度がよく現れることが分かった.

システムの提案に対するユーザ応答の音声データの収録を行った. 今回は, 肯定的な態度, 否定的な態度のデータを数多く収集するため, システムの提案発話を音声合成装置を用いて合成し, その応答を肯定的, 否定的な態度で発話するという形で収録を行った. 収録した発話の種類を表 1 に示す. 収録は研究室の男子学生 20 人分 (計 2000 発話) について行った.

例えば, 「ラーメン」「○○かー」「否定的」なデータを収録する場合, まず合成音で「ラーメンなんてどうかな」という音声が流れる. その後ユーザが「ラーメンかー」という内容を否定的な気持ちを込めて発話するという形で行った.

*Spoken Dialogue System Using Para-Linguistic Information, by Fujie Shinya†, Yagi Daizo†, Kikuchi Hideaki‡, Kobayashi Tetsunori† (†Sch. of Sci. and Eng., Waseda Univ.: ‡Sch. of Human Sci., Waseda Univ.)

表 1: 収録した発話の種類

カテゴリ・店 (10 種)		言い回し (5 種)	態度 (2 種)
ハンバーガー	マクドナルド	○○かー	肯定的
ラーメン	味源	○○ねー	否定的
弁当	夢民	○○	
カレー	ボカ弁	いいんじゃない	
学食	そばの実	そうだね	

3 韻律情報による発話態度の認識

3.1 特徴量と識別手法

収録した発話に対し, 音素アライメントの実行と基本周波数 (F0) の抽出を行った. 各音素ラベルに対する F0 の観察と検討を行った結果, 図 1 に示すように, 以下の 3 次元の特徴 $\boldsymbol{x} = (x_1, x_2, x_3)$ を識別に用いることとした.

x_1 : 第 1 モーラの母音部分の F0 の傾き

x_2 : 発話全体の F0 レンジ

x_3 : 最終モーラの継続長

ここで, 第 1 モーラの F0 の傾きは最小自乗法で求めた. また, 発話の最初に適当なフィラーを言うことを許容したため, 特徴量の抽出はその部分を除いて行った.

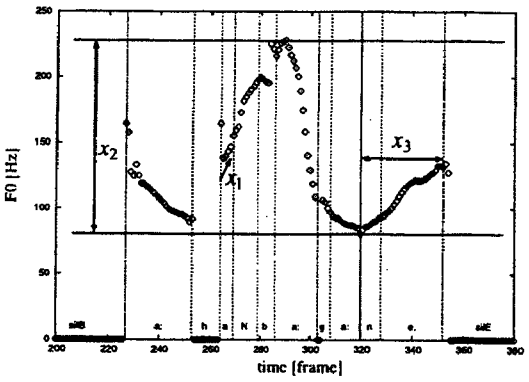


図 1: 特徴量の抽出例

抽出した特徴量を用いて, 肯定的, 否定的の 2 カテゴリについて単一正規分布を最尤推定により学習した. 識別はこの分布を用いたベイズ識別によって行う.

3.2 実験結果

収録した 20 人分計 2000 発話のデータを, 4 セット (1 セットあたり 5 人分計 500 発話) に分けた. 3 セットで学習を行い, 残る 1 セットの識別を行う. これを各セットにつき計 4 回行い, 各セットの正解率の平均を取った. 結果を図 2 に示す.

言い回しによる違いを見るため, 最終モーラの継続長が常に短くなる「いいんじゃない」を除いた実験,

また傾向が似ている「○○かー」「○○ねー」のみを用いた実験，「○○かー」のみを用いた実験をそれぞれ行った．

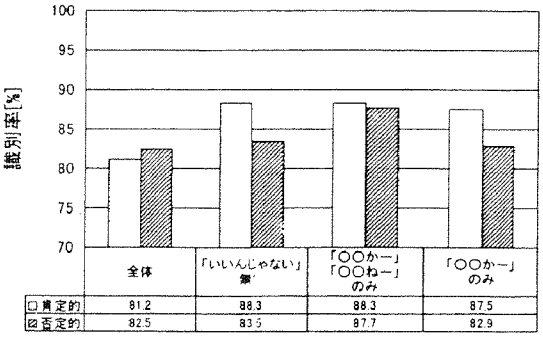


図 2: 識別実験結果

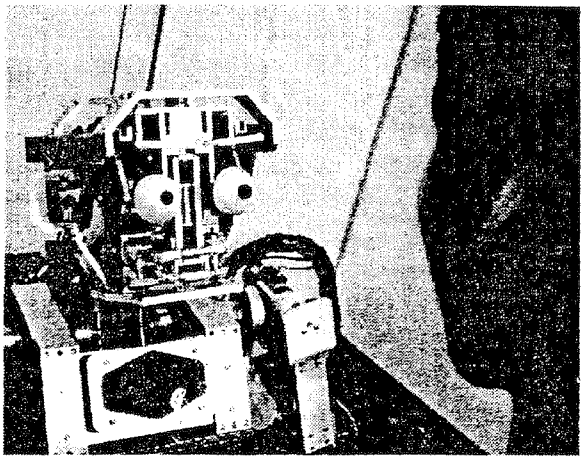


図 3: ROBISUKE

4 人による認識との比較

前節での実験は，発話者が収録時に想定した態度を正解としている．ここでは，人間の持つ識別能力との比較のため，収録した発話を人が識別した結果と，前節の手法で識別した結果との一致率を見る．

4.1 人による識別結果の収集

収録したデータのうち，態度が肯定的，否定的のものからそれぞれ 20 発話，計 40 発話をランダムに選択した．この 40 発話を 5 人の作業者 (A〜E) に順不同に聞いてもらい，それぞれ否定的か肯定的かを判断させた．

4.2 一致率の計算方法と結果

一致率の計算には，対話データのタグ付けの評価等に用いられる Cohen の κ を用いた．5 人の作業者について，他の 4 人の作業者との間で κ を求め，その最大値，最小値，平均値を求めた．また同様に，態度識別システムの判定結果 (M) と 5 人の作業者の間で κ を求めてその最大値，最小値，平均値を求めた．結果を表 2 に示す．

平均値を見ると，今回の識別結果は 6 人中 5 位であるが，最小と最大の一致率は人間のものと近い．また，Cohen の κ は，$0.60 < \kappa < 0.75$ の範囲にあるときに良好な結果とされており，その点からも十分人間に近い識別能力を持つと言える．

5 パラ言語情報を用いた音声対話システム

これまでに述べた韻律情報による態度の認識と，頭部ジェスチャの認識 [1] の結果をパラ言語情報として利用する音声対話システムを，対話ロボット ROBISUKE(図 3) 上に構築した．対話例を図 4 に示す．

システムは，提案に対してユーザの応答が否定的な場合は代案を，肯定的な場合はより具体的な提案を行う．対話例では，「カレーか」「弁当ね」などの言語情報だけでは判断しかねる発話に対し，パラ言語情報の利用によって判断を行い，対話を円滑に進められることが確認できる．

表 2: 一致率 κ の計算結果

	最小値	最大値	平均値
M	**0.52**	**0.80**	**0.66**
A	0.52	0.85	0.72
B	0.52	0.79	0.62
C	0.65	0.80	0.72
D	0.65	0.85	0.76
E	0.61	0.79	0.72

U: お昼ごはんなんだけど，
　どこかいいところ無いかな
R: カレーなんかどう
U: カレーか (**否定的**)
R: それじゃあ，弁当なんかどうかな
U: ああ，弁当ね (**肯定的**)
R: 弁当なら近くにホカ弁があるよ

図 4: 対話例
U はユーザ，R は ROBISUKE の発話

6 まとめ

音声対話で利用されるパラ言語情報の一つとして，韻律情報による発話態度の識別を行った．識別能力に関して，人間とほぼ同等の能力を持つことを一致率をもとに確認した．また，パラ言語情報を利用した対話システムの例を示し，対話の進行に有効に作用することを確認した．

今後は，パラ言語情報を用いることによる有効性の評価法，応答以外の発話に関する態度の認識手法などについて検討を行う予定である．

参考文献

[1] 江尻康，松坂要佐，小林哲則，“対話中における頭部ジェスチャの認識” 電子情報通信学会技術報告，Vol.102，No.218，pp.31-36，PRMU2002-61.

[2] 藤尾茂，ニックキャンベル，樋口宣男，“韻律を用いた発話タイプの識別”，音講論，pp.199-200，Sep.，1995.

[3] 田本真詞，川森雅仁，川端豪，“文型と音調によるユーザ発話行為理解”，音講論，pp.35-36，Sep.，1998.

テキスト音声合成における感情表現の付加

谷津 慶* 村上 真* 杉本 富利* 米山 正秀*

Addition of the emotional expression for text-to-speech synthesis

Kei Yazu*, Makoto Murakami*, Futoshi Sugimkoto*, Masahide Yoneyama*

Abstract - Representing emotional expressions on text-to-speech synthesis has been a future subject. We aim to add the emotional expressions to the synthetic speech. This system recognizes the emotion based on the words in a sentence and outputs the emotional speech.

Keywords: prosody, text-to-speech synthesis and morphological analysis

1. はじめに

近年、テキスト音声合成の技術が発達して比較的自然な合成音声を得ることができるようになったが、その合成音声に感情を付加する課題については未解決な問題が多くある。本研究では合成音声に感情表現を付加することを目的としている。

2. 感情表現

2.1 感情表現の過程

心理学的には感情表現の過程は Fig.1 のように定義できる。[1] 本研究ではそれを Fig.2 のように応用している。

Fig.1 一般の感情表現過程

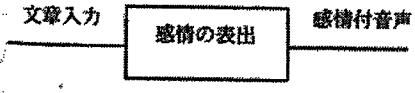


Fig.2 本研究の感情表現過程

2.2 対象感情の制限

本研究では、音声を強く特徴づける感情と思われる基本感情である「喜び、悲しみ、怒り、驚き」と各感情の中心となる「無感情」についてのみ感情表現の対象とする。

*: 東洋大学工学部情報工学科
*: Dept. of information and computer sciences, Toyo University

また、日本語には多くの単語があるが語数を限定するため IPAL[2]を用いた。その形容詞に関してはすでに感情分類が行われている。

2.3 感情付加表現のシステム

本研究での感情付加の概略方式を Fig. 3 に示す

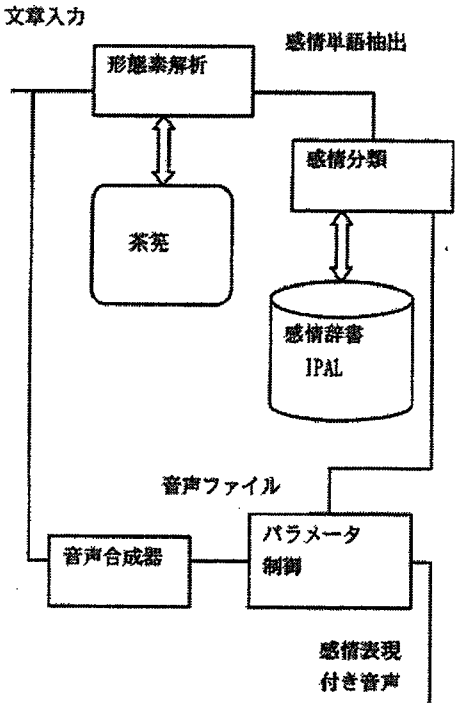


Fig. 3 感情表現の過程のシステム

茶筅[3]による形態素解析によって入力文章から抽出された感情を持った単語が持っている感情に基づいて入力文章の感情を決定する。まず、文章を入力した後に形態素解析を行って感情を持つ単語のみを抽出する。つぎに IPAL を基にした感情辞書の中の形容詞 137 語中本研

究に該当する感情の形容詞 31 語を参照して文章の感情を決定付ける。その後、各感情ごとのパラメータを音声ファイルに付加して感情付き音声を出力する。

3. 感情と韻律の関係

3.1 パラメータの種類

感情を付加するためのパラメータは主に次の３つである。

- テンポ（時間）
- ピッチ（基本周波数）
- パワー（振幅）

3.2 感情と音響学的特徴

感情を表現するための韻律パラメータ（ピッチ、振幅、速さ）と表現感情との関係については既に研究がなされており、それは Table 1 に要約したとおりである。[1]

無感情を基準として各パラメータを比較してある。なお、現在のところ「怒り」に関しては明らかではない。

Table 1 感情と音響学的特徴の統計

	無感情	喜び	悲しみ	怒り	驚き
ピッチの高さ	普通	高い	やや低い	高い	高い
パワー	普通	大きい	小さい	大きい	普通
テンポ	普通	速い	遅い	速い or 遅い	速い

4. 韻律パラメータの数値化の実験

4.1 実験方法

Table 1 の定性的表現を数値化するためにテキスト音声合成及び波形編集ソフト「Visual Speech Creator」(NTT-IT)を用いて試聴実験を行い各感情のパラメータを決定した。

被験者は１０人の学生によって行われた。被験者達には、例文「サッカー日本代表がロシアに勝ちました」をテキスト音声合成ソフトで試聴しながら、ピッチ、振幅の大きさ、速度の３つのパラメータを各感情が最も表出できるように文の全体部分を一括で調整するように依頼した。さらに、全体だけで満足できなかったら、音素ごとに波形を調節してもらった。

一つの文で多くの感情を表すことは、まず実際には考えられないことだが、ここでは被験者に意味を考えないで韻律だけで判断してもらうようにした。

4.2 実験結果

各パラメータ毎の全被験者の平均値を Table 2 に示す。なお、単位の後のカッコ内は無感情と比べた倍率である。

Table 2 実験結果

	無感情	喜び	悲しみ	怒り	驚き
ピッチの高さ	280 (Hz)	234 (Hz) (0.8 倍)	358 (Hz) (1.2 倍)	395 (Hz) (1.4 倍)	354 (Hz) (1.2 倍)
パワー	50	33 (0.6 倍)	80 (1.6 倍)	80 (1.6 倍)	65 (1.2 倍)
テンポ	3.765 (s)	4.314 (s) (1.2 倍)	3.381 (s) (0.9 倍)	3.021 (s) (0.8 倍)	3.381 (s) (0.9 倍)

括弧の中の数字は無感情と比べたパラメータの倍率である。また、「怒り」は語尾の音素のピッチとパワーを上げるという人も４人いた。

5. まとめと今後の課題

実験により得られた数値は Table 1 とほぼ同じ結果となり感情音声合成がうまくできたが、怒りについては感情表現が難しかった。本研究では形容詞だけで感情判定を行ってきたが、今後は名詞、動詞、さらにそれらの組み合わせも含めて感情を決定する方法を検討している。そのためには名詞、動詞の組み合わせによる感情判定のルール作りが必要である。また、本研究では文章全体に同一のパラメータ設定を適応しているが、今後は単語単位や音素単位でのパラメータ制御を検討してより人間の話し方に近づける必要がある。

参考文献

[1] 計算機用日本語形容詞辞書 IPAL, 情報処理振興事業協会技術センター,(1990).

[2] 日本語形態素解析システム「茶筅」, http://cl.aist-nara.ac.jp/lab/nlt/chasen.html, 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座, (1999).

[3] 感情心理学への招待 感情・情緒へのアプローチ, 濱 治世 他, サイエンス社, 第１章,第４章,(2001).

3—8—10

韻律情報を用いた肯定的／否定的態度の認識*

◎八木大三† 藤江 真也† 菊池 英明‡ 小林哲則†
(†早大・理工 ‡早大・人間科学)

1 はじめに

韻律情報を用いて, 発話者の肯定的／否定的態度を認識する.

人間同士の対話では, 言語的な情報だけでなく, 様々な視覚的, 聴覚的情報を補助的に用いて, 円滑に対話を進行している. 我々はこれまでに, 発話に含まれる韻律情報を用いて発話態度の自動識別を行い, 対話システムに応用してきた. そこでは, 基本周期のダイナミックレンジ, 語末のモーラ長, 語頭の基本周期の傾きを特徴量として用い, 人間に近い能力を持つ識別システムを実現した [1]. しかしながら, ここで用いた特徴量はシステムの設計者が少数の例の観察に基づいて選択したものであり, 最適なものではない. より高い性能を与えるためには, どのような特徴量を用いることが適当かについて検討することが望まれる.

一方,Dellart らは, "喜び", "悲しみ", "怒り", "恐怖" の 4 つの感情の認識に有用な韻律特徴を,Forward Selection という手法により選択することを試みている [2]. 本研究では,Dellart らの方法を, 発話態度の識別問題に適用し, そこでどのような韻律情報が有効となるかを調べる. また, 同じ問題を決定木学習の結果から検討し, 両者の結果を比較する.

2 データの収録

ここで検討対象となる問題は, 相談者と提案者の対話において, 提案者の提案を相談者が肯定的に受取ったか否かの判断を, 提案を受けて相談者が発した第一発声の韻律から行おうとするものである.

そこで必要となる肯定的な態度, 否定的な態度のデータを数多く収集するため, 提案発話を音声合成装置を用いて合成し, その応答を肯定的, 否定的な態度で発話するという形で収録を行った. 収録した発話の種類を表 1 に示す. 収録は研究室の男子学生 20 人 × カテゴリ・店 (10 種)× 言い回し (5 種)× 態度 (2 種)= 計 2000 発話について行った.

例えば, 「ラーメンかー」(否定的な態度で発話) を収録する場合, まず合成音で「ラーメンなんてどうかな」という音声が流れる. その後ユーザが「ラーメンかー」という内容を否定的な気持ちを込めて発話するという形で行った.

3 特徴量

ここで検討対象とする韻律の特徴量は, 以下の 11 種類である. これらの特徴量は F0 平均値を除いて, 話者ごとに平均値を求め, その値を引くことにより話者による違いを除いてある. F0 の抽出には ESPS/waves+ を使用し, 音素アライメントには当研究室で開発したデコーダを使用した.

表 1: 収録した発話の種類

カテゴリ・店 (10 種)	ハンバーガー, ラーメン, 弁当, カレー, 学食, マクドナルド, 味源, 夢民, ホカ弁, そばの実
言い回し (5 種)	○○かー, ○○ねー, ○○ いいんじゃない, そうだね
態度 (2 種)	肯定的, 否定的

- x_1 : F0 平均値
- x_2 : F0 中央値
- x_3 : F0 最大値
- x_4 : F0 最小値
- x_5 : F0 レンジ
- x_6 : アクセントモーラでの F0 最大値
- x_7 : 句末モーラでの F0 最小値
- x_8 : アクセント位置のモーラ長
- x_9 : 第 1 モーラのモーラ長
- x_{10} : 句末モーラのモーラ長
- x_{11} : 平均モーラ長

4 Forward Selection による特徴量選択

4.1 Forward Selection

Forward Selection は, 多次元の特徴量の中から, 重要な特徴量を選択していく方法である. まず, それぞれの特徴量を 1 次元の特徴量として識別を行い, その中で最も良い識別率を与えた特徴量を 1 次元目の特徴として採用する. 次に, ここで選んだ特徴量に加え, 残りの特徴量をそれぞれを加えた 2 次元のベクトルによって識別を行い, 最良の結果を与えた特徴量を 2 次元目の特徴量として採用する. 以下同様にして, 順番に各次元に用いる特徴量を選択していく.

本実験においては,Forward Selection の各ステップにおいて, 肯定的, 否定的の 2 カテゴリの識別を混合正規分布 (GMM) を用いて行った.GMM の混合数は, 各ステップ毎に 1, 2, 3, ···, 10, 16, 32, 64 の 13 通り変化させ, 最高値を与えた結果を採用した. データは, 先に述べた 20 人分計 2000 発話であり, これを 4 セット (1 セットあたり 5 人分計 500 発話) に分け,3 セットで学習に, 残る 1 セットを識別に用いた. この操作を組み合わせを変えて計 4 回行い, 各セットの正解率の平均を取った.

4.2 実験結果

Forward Selection を用いた識別結果を表 2 に示す.

表 2 から, x_1(F0 平均値), x_{10}(句末モーラのモーラ長), x_5(F0 レンジ), x_8(アクセントモーラのモーラ長), x_7(句末モーラでの F0 最小値), x_6(アクセントモーラ

*Recognition of positive/negative attitudes using prosodic infomation by Daizo Yagi†, Shinya Fujie†, Hideaki Kikuchi‡, Tetsunori Kobayashi† (†Sch. of Sci. and Eng., Waseda Univ.: ‡Sch. of Human Sci., Waseda Univ.)

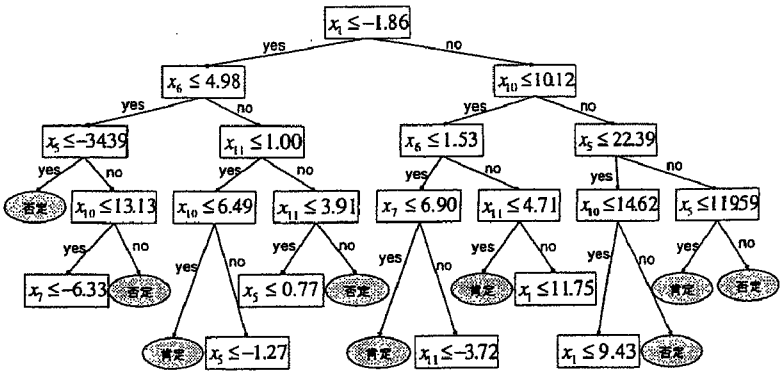


図 1: 決定木学習の結果

での F0 最大値), x_{11}(平均モーラ長), x_2(F0 中央値) の 8 次元を用いた場合に最も良い識別率が得られることがわかった.

表 2: Forward Selection を用いた特徴量選択

次元数	特徴	GMM 混合数	識別率 [%]
1	x_1	7	83.60
2	x_{10}	10	87.95
3	x_5	64	89.80
4	x_8	10	90.40
5	x_7	32	90.65
6	x_6	6	90.80
7	x_{11}	7	91.10
8	x_2	5	**91.15**
9	x_9	3	90.75
10	x_4	16	89.90
11	x_3	4	89.35

5 決定木による学習

各特徴量の有効性を見るために, 決定木による学習を行った. 決定木学習には C4.5 を使用した. 決定木による学習の結果の一部を図 1 に示す. 図 1 を見ると, x_1(F0 平均値), x_{10}(句末モーラのモーラ長), x_6(アクセントモーラでの F0 最大値), x_5(F0 レンジ), x_{11}(平均モーラ長) が上位ノードに現れている.

6 人による認識との比較

前節での実験は, 発話者が収録時に想定した態度を正解としている. ここでは, 人間の持つ識別能力との比較のため, 収録した発話を人が識別した結果と, 前節の手法で識別した結果との一致率を見る.

収録したデータのうち, 態度が肯定的, 否定的のものからそれぞれ 20 発話, 計 40 発話をランダムに選択した. この 40 発話を 5 人の作業者 (A〜E) に順不同に聞いてもらい, それぞれ否定的か肯定的かを判断させた.

一致率の計算には, 対話データのタグ付けの評価等に用いられる Cohen の κ[3] を用いた. 5 人の作業者について, 他の 4 人の作業者との間で κ を求め, その最大値, 最小値, 平均値を求めた. また同様に, 4 節の実験で最も良い識別率が得られた特徴量, 混合数において, 態度識別システムの判定結果 (M) と 5 人の作業者の間で κ を求めてその最大値, 最小値, 平均値を求めた. 結果を表 3 に示す.

表 3: 一致率 κ の計算結果

	最小値	最大値	平均値
M	**0.65**	**0.80**	**0.73**
A	0.52	0.85	0.71
B	0.52	0.79	0.65
C	0.65	0.80	0.74
D	0.65	0.85	0.76
E	0.71	0.79	0.75

表 3 より, 今回の識別結果は人間のものと近い. また, Cohen の κ は, $0.60 < \kappa < 0.75$ の範囲にあるときに良好な結果とされており, その点からも十分人間に近い識別能力を持つと言える.

7 まとめ

多次元の韻律情報の特徴量を用いて, 肯定的／否定的態度の識別を行った. Forward Selection による特徴量選択を行った結果, 肯定的／否定的態度の認識には x_1(F0 平均値), x_{10}(句末モーラのモーラ長), x_5(F0 レンジ), x_8(アクセントモーラのモーラ長),x_7(句末モーラでの F0 最小値), x_6(アクセントモーラでの F0 最大値), などが重要であることがわかった. また, 決定木による学習を行った結果, Forward Selection により選択された特徴量が上位ノードに現れ, その特徴量の重要性が確認された.

今後は, 応答以外の発話に関する様々な態度の認識手法などについて検討を行う予定である.

参考文献

[1] 藤江真也, 八木大三, 菊池英明, 小林哲則, "パラ言語情報を用いた音声対話システム", 音講論,pp.39-40, Sep., 2003.

[2] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," Proc. of ICSLP'96, (1996).

[3] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, no. 1, 37~46, (1960).

2—P—23

心的態度表現における韻律的／分節的特徴の影響*

◎大久保 雅史† 望月 亮†,‡ 小林 哲則†
†早稲田大学理工学部 ‡松下電器産業株式会社

1 はじめに

合成音声に付与する韻律的特徴及び分節的特徴が，合成音声の心的態度表現にどの程度寄与するか調査した．

人間同士の対話において，発話には言語情報に加え心的態度に代表されるパラ言語情報が含まれ，それらが総合的に作用して発話者の意図を伝えている．このため高度な対話システムを実現するためには，音声合成器もまた言語情報のみならず発話者の心的態度を表現する必要がある．我々は音声合成によって態度表現を行なうことを目的に，韻律によって話者の心的態度がどの程度制御できるか調査したが [1]，より高度に態度表現をモデル化するためには，韻律的特徴のみらず，分節的特徴の振舞についても調査する必要がある．そこで本研究では Russell[2] の提案する心的態度表現空間を用いて，「快ー不快」，「覚醒ー不覚醒」という２つの態度が韻律的特徴及び分節的特徴によってどの程度表現できるのか調査した．本報告では，上述の韻律的特徴，及び分節的特徴がそれぞれ合成音声の心的態度表現にどの程度寄与するのか調査した試聴評価実験とその結果について報告する．

2 心的態度の表現空間

従来の音声合成において，パラ言語情報の付与に関しては，怒り，悲しみ，といった感情の扱いが中心であった [3][4]．しかしながら，我々は人間とシステムが協調しながらタスクを遂行する対話システムの実現を目的にしており，このような対話では，対話に対する注目度・共感度，自発話に対する確信度，あるいは状態に対する満足度といった，対話状況に対する発話者の内的な評価が必要であることが多く，それらは基本感情とは必ずしも整合しない．そこで本研究では，話者の内的な評価を表す手法として，Russell の提案する「快ー不快」，「覚醒ー不覚醒」の２軸を用いて話者の心的態度を表現することにした．

3 合成音声の心的態度表現力評価

3.1 音声試料

プロのナレータが対話の中で意図的に心的態度を含んで発声した発話を収録し，評価に用いた．対象となる発話の収録は，その前後の対話のコンテキストを擬似的に用意し，意図した態度が収録しやすいようにして対話形式で収録した．今回の評価では発話内容（文の音韻並び）が同じものについて，心的態度が異なると思われる音声を１３～１４種類用意した．以下に示す評価では，このような発話セットを合計４セット使用した．

3.2 評価方法

収録した各発話セットから，「快ー不快」，「覚醒ー不覚醒」の両軸において中立的な態度の音声を選択し，評価の際の基準音声として用いる．韻律的特徴，及び分節的特徴がどの程度態度表出に寄与するかを調べるために，上述の要領で収録した音声について，図１に示すように２種類の合成音声を作成した．

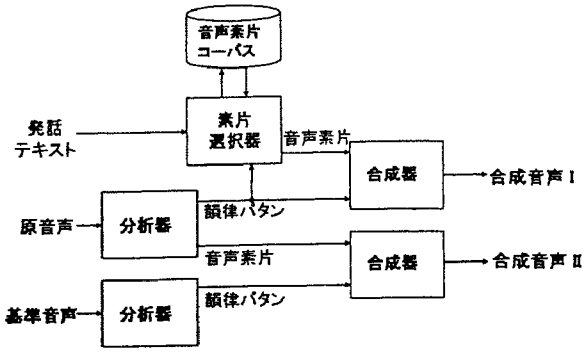


図 1: 合成音声作成の流れ

- 【合成音声 I】

  原音声から抽出した韻律パタンを利用．音声素片は同一ナレーターが発声した約３時間分の音声素片コーパス（中立的な音声素片）を使用して合成した音声．

- 【合成音声 II】

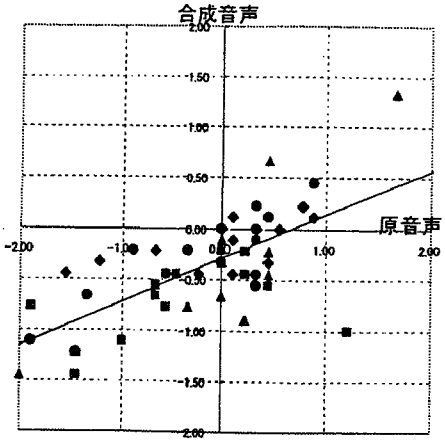
  原音声から抽出した音声素片を利用．基準音声から抽出した韻律パタンに合わせて合成した音声．

ここで，合成音声 I は，中立的な態度の音声素片を用いることで，自然音声の韻律的特徴のみでどの程度心的態度が表現できるのかを調査するために，また合成音声 II は，中立的な態度の韻律パタンを付与することで，分節的特徴のみでどの程度心的態度が表現できるのかを調査するために用いた．音声合成には，PSOLA[5] 法を用いた．評価は，原音声，及び上述の合成音声 I，II を評価対象音声とし，この評価対象音声を基準音声と一対にして提示して，「快ー不快」，「覚醒ー不覚醒」の各軸について５段階（＋２～－２）で評価した．なお，評価者は９人である．

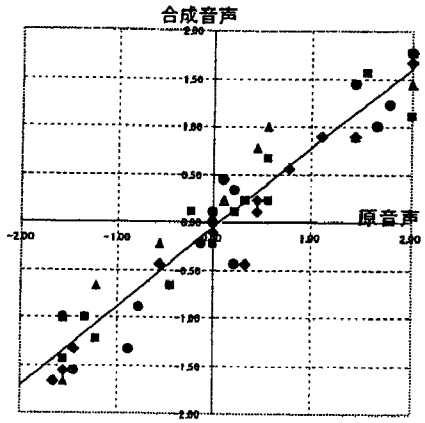
3.3 評価結果

原音声の心的態度と，上述の合成音声 I，及び合成音声 II での心的態度の表現力の関係をそれぞれ図 2，及び図 3 に示す．各図の (a) は「快ー不快」，(b) は「覚醒ー不覚醒」についての評価結果である．ここで図中のプロットは，評価者９人の平均値であり，横軸は原音声の評価値，縦軸はその原音声をもとに各条件で合成した音声の評価値を示す．図中の直線はプロットの近似直線を示し，この近似直線の傾きが１に近いほど，合成音声においても態度が表出されていることを意味する．表１に，各図における近似直線の

*Prosodic and spectral effect on affected speech expressed by speech synthesis: by Tadashi OKUBO† , Ryo MOCHIZUKI†,‡ and Tetsunori KOBAYASHI† (†Sch. of Sci. and Eng., Waseda Univ., ‡Matsushita Electric Ind. Co., Ltd.)



(a) 「快－不快」



(b) 「覚醒－不覚醒」

図 2: 韻律的特徴による表現力評価

表 1: 近似直線の傾き

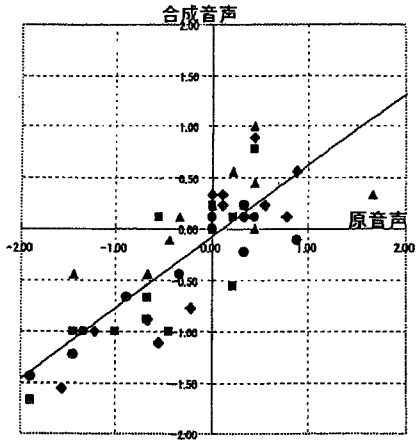
合成音声 I		合成音声 II	
(a)	(b)	(a)	(b)
0.43	0.83	0.69	0.56

傾きを示す．また，図中のプロットの形状は，発話内容の種類を表している．

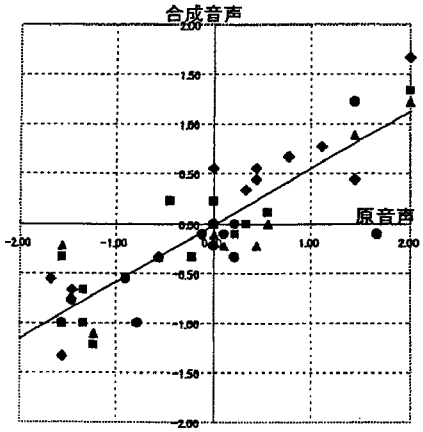
この結果から，合成音声 I による「快－不快」の表現においては，合成音声の評価値が中立側に縮退しており，表現力が弱っているといえる．しかし，「覚醒－不覚醒」の表現においては合成音声の評価値が原音声の評価値とかなり近い値を示しており，合成音声においても心的態度が表現できていることが分かった．また，合成音声 II による「快－不快」の表現においては，比較的合成音声の評価値と元の原音声の評価値が近い値を示しており，合成音声においてある程度の心的態度が表現できていることが分かった．

4 まとめ

対話システムにおいて心的態度表現が可能な音声合成器を実現することを目的に，今回は合成音声に付与する韻律的特徴及び分節的特徴が，心的態度表現にどの程度寄与するか調査した．自然音声の韻律パタンを用いて合成音声を作成した場合，また自然音声の分節的特徴を用いて韻律パタンに中立的な態度のものを用いた場合，どの程度合成音声に原音声



(a) 「快－不快」



(b) 「覚醒－不覚醒」

図 3: 分節的特徴による表現力評価

の心的態度を再現できるか試聴評価を行った．実験結果より，「覚醒－不覚醒」については韻律の影響が支配的であり，「快－不快」については分節的特徴の影響が支配的であることが分かった．

参考文献

[1] 大久保雅史, 望月亮, 蓑輪利光, 小林哲則, "波形重畳型音声合成における心的態度の再現性評価", 情報科学技術フォーラム (FIT), F-029, 2003.

[2] J.A. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, vol.39, pp.1161-1178, 1980.

[3] Y. Kitahara and Y. Tohkura, "Prosodic Control to Express Emotions for Man-Machine Speech Interaction," IEICE Trans. Fundamentals, E75-A, No.2, pp.155-163, Feb.1992.

[4] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasmura, "A Speech Synthesis System with emotion with assisting communication," In Proceeding of the ISCA Workgroup on Speech and Emotion, pp.167-172, 2000.

[5] E.Moulines and F.Charpentier, "Pitch-synchronous waveform processing teschniques for text-to-speech synthesis using diphones," Speech Comm., Vol.9, pp.453-467, 1990.

韻律情報を利用した文章入力システムのための韻律制御モデル*

◎大久保 崇（早大・理工）　菊池 英明（早大・人科）　白井 克彦（早大・理工）

1　はじめに

近年、パソコンや携帯電話の普及により、メールやチャットなどの文字によるコミュニケーションが盛んに行われるようになってきた。これに伴い、やりとりされる文章も単純な文章から、楽しく円滑なコミュニケーションを行うために、字体の変化や顔文字などを含めた文章へと変化してきている。そこで本研究では、音声による文章入力においても、このような多彩な文章表現を簡単に入力できるインタフェースの実現を目指し、従来の音声ディクテーションに韻律情報を利用したシステムの提案と構築を行う。

韻律情報を利用する音声入力インタフェースはこれまでにもいくつか提案されてきた。音高の変化や有声休止といった韻律的な情報をキーボードの Shift キーや Tab キーの代わりに利用するインタフェース [1][2] などは、音声に含まれる韻律的な情報を利用することで、より使い易いインタフェースが実現可能であることを示している。

本研究では、声の高さや大きさ、速さといった複数の韻律情報を複合的に利用することで、複数の文章表現を韻律制御のみで正しく入力できるシステムの構築を目指す。本稿では、2 章で目標とするシステムの概要と操作に必要な韻律制御方法を示し、3 章ではその韻律制御方法に基づいて発声された音声の判別実験について述べ、4 章でまとめとする。

2　システムの概要と韻律制御方法

本研究で構築するシステムでは、韻律を制御する事で次のような表現を加えた文章の入力を可能とする。

- 強調表現（フォント変更）
- 段落変更
- 顔文字 ([怒](-_-X), [喜]＼(^0^)／, [悲](T-T))

ユーザは音声入力時に、あらかじめ決められた韻律制御方法に従って意図的に発声する事で、これらの文章表現を入力する事ができる。本研究では、その韻律制御方法を表 1 のように定めた。この制御方法は、顔文字や強調表現 (太文字) を施した 40 文を、被験者 9 名に意識的に韻律を制御して読み上げてもらった際の韻律制御方法を参考に設定したものである。これにより、少しでもユーザが習得し易く、極端な不自然さを感じ難い韻律制御方法を目指した。

例えば、「頑張れ＼(^0^)／」と入力する時は、「頑張れ」を高く大きな声で発声してもらい、最終モーラの「れ」を通常よりも比較的長く発声してもらう。このように発話 (句) 全体と最終モーラの 2 箇所に韻律的な変化を置くことで、5 種類の韻律制御パターンをシステムが判別し易くなるようにした。

表 1: 韻律制御方法

	全体	最終モーラ
普通	―	―
強調表現	高く+大きく+ゆっくり	―
段落変更	―	強く
顔文字 (怒)	高く+大きく	低く → 高く
顔文字 (喜)	高く+大きく	伸ばす
顔文字 (悲)	低く+小さく	伸ばす

3　判別実験

表 1 で示した韻律制御方法に基づいて発声された音声を収録し、どの文章表現を入力するための音声かを決定木学習により判別する実験を行った。

3.1　収録データ

被験者は男子学生 11 名で、以下の 8 つの文章を韻律制御方法を変えてもらい、1 人につき 82 文、計 902 文の音声を収録した。

- 日曜日は 12 時にハチ公前ね
- 今からタクシーで帰ります
- その本は面白いと思うよ
- チャーハンよりラーメンが食べたい
- 早稲田大学理工学部の皆様
- 誕生日のプレゼントが届いた
- 今日は朝から雨だって
- 大きなみかんをもらいました

発話文は、アクセント型による f0 パターンの違いを考慮し、すべて有核句の文章とした。また、メールなどへの利用を想定して、文末表現をですます調や口語調、体言止と多様化した。また、収録の際に被験者には、表 1 の韻律制御方法を示しながら口頭で説明し、それに加えて、発声例の音声を数回聞かせてから発声してもらった。

3.2　判別に用いる韻律パラメータ

本研究で構築するシステムは、一定時間長のポーズ毎に区切られた音声区間の韻律情報を抽出し、どの文章表現が入力されたのかを判別する。そこで本実験では、ポーズで区切られた音声区間全体とその最終モーラに関する全 20 種類の韻律情報を判別のパラメータに用いた (表 2)。表 2 中の F_f0_rise は最終モーラ末の f0 値と F_f0_min との差分をとった値であり、D_pw_ave は F_pw_ave と pw_ave との差分をとった値である。また、判別に使うパラメータはその話者の基本周波数やパワー、発話速度の平均値と標準偏差を用いて正規化した。

*"Prosody Control Model for Dictation System Using Prosodic Information" by Takashi OKUBO, Hideaki KIKUCHI, and Katsuhiko SHIRAI (University of WASEDA)

表 4: 決定木による判別結果 (数値の単位は κ 値以外すべて%)

話者	正解率 (κ 値)		普通		強調表現		段落変更		顔文字 (怒)		顔文字 (楽)		顔文字 (悲)	
	closed	open	closed	open	closed	open	closed	open	closed	open	closed	open	closed	open
A	91.8(0.87)	92.4(0.87)	94.8	96.9	83.3	92.9	87.5	25.0	87.5	100.0	100.0	87.5	100.0	100.0
B	88.4(0.78)	82.9(0.66)	93.7	91.5	78.6	61.9	62.5	75.0	75.0	75.0	87.5	62.5	87.5	75.0
C	86.0(0.76)	88.2(0.81)	93.3	94.2	78.6	81.0	75.0	100.0	62.5	62.5	62.5	62.5	87.5	87.5
D	86.8(0.77)	82.5(0.68)	92.2	97.4	78.6	73.8	75.0	87.5	87.5	0.0	62.5	0.0	87.5	75.0
E	91.0(0.80)	85.2(0.68)	96.2	93.4	95.2	97.6	0.0	37.5	100.0	37.5	62.5	0.0	62.5	12.5
F	87.6(0.78)	81.1(0.68)	92.8	91.0	83.3	71.4	62.5	50.0	50.0	12.5	87.5	75.0	100.0	100.0
G	89.3(0.82)	88.2(0.80)	94.7	97.3	78.6	61.9	75.0	87.5	62.5	87.5	100.0	87.5	100.0	100.0
H	88.3(0.81)	91.1(0.85)	93.4	96.2	83.3	95.2	62.5	87.5	75.0	62.5	87.5	75.0	87.5	50.0
I	82.2(0.69)	84.5(0.74)	92.3	88.5	76.2	88.1	25.0	25.0	75.0	75.0	62.5	75.0	50.0	100.0
J	93.9(0.89)	94.4(0.90)	98.4	100.0	88.1	83.3	87.5	100.0	75.0	100.0	75.0	100.0	100.0	50.0
K	96.6(0.94)	96.0(0.93)	98.1	100.0	100.0	92.9	75.0	75.0	87.5	100.0	100.0	75.0	87.5	100.0
All	89.3	87.7	94.6	95.0	84.0	81.8	62.5	68.2	76.1	64.8	80.7	63.6	88.1	76.2

表 2: 判別に用いた韻律パラメータ

		全体	最終モーラ
基本周波数 (f0)	最大値	f0_max	F_f0_max
	最小値	f0_min	F_f0_min
	平均値	f0_ave	F_f0_ave
	レンジ	f0_range	F_f0_range
	句末の上昇度	—	F_f0_rise
パワー	最大値	pw_max	F_pw_max
	最小値	pw_min	F_pw_min
	平均値	pw_ave	F_pw_ave
	レンジ	pw_range	F_pw_range
	句末の上昇度	D_pw_ave	
発話速度		speed	F_speed

表 3: データ数と内訳

普通	強調表現	段落変更	顔文字			計
			怒	喜	悲	
1299	462	88	88	88	84	2109

3.3 決定木学習による判別

収録した音声データを決定木学習によって判別した。決定木学習のアルゴリズムは C4.5[3] を用い、判別対象となるデータ数とカテゴリの内訳は表 3 に示した。

判別の結果を表 4 に示した。表中の正解率はテストデータに対する正解データの割合を表し、κ 値は偶然性を排除した判別の一致率を、それ以外の項目はカテゴリ毎の正解率を表す。また、判別は closed と open の 2 種類行った。closed はその話者のデータのみを用いて 10-fold の cross validation を行った結果で、システムに実装した際に、ユーザが事前に数種類の文章を決められた韻律制御で読み上げるエンロール作業を行った場合を想定している。open は、その話者以外の全データを訓練データとして作った決定木を用いて判別を行った結果であり、システム利用時にエンロールを一切行わなかった場合を想定している。

3.4 考察

まず全体の正解率を見ると、closed と open の両方の場合において、全ての話者で 90%前後の正解率が得られており、今回実験した韻律制御方法である程度の判別が可能である事が示されている。また、open な判別の正解率を話者別に見ると、最も高い話者 (K) が 96.0%である事から、ユーザが韻律制御に慣れてくれば、高い精度でユーザの意図した文章表現の入力が可能であるといえる。

次に、カテゴリ別に見ると段落変更の判別精度が全体的に低い。段落変更は強調表現として誤判別される事が多く、この 2 つのカテゴリを区別するためのパラメータを追加する必要がある。しかし、段落変更に関しても話者によっては 100%の判別ができているので、ユーザが韻律制御に慣れていけば解消される問題とも考えられる。また、顔文字 (怒) と顔文字 (楽) の open な判別では、0%とまったく判別されない話者がいた。これは、顔文字の入力に設定した制御方法が複雑なため、話者によって得手不得手がわかれた事が要因と考えられる。これに関しては、0%の話者も closed な判別では精度が上がることから、韻律制御がうまくできず、判別率が極端に低いユーザに対しては、エンロールを行ってもらう必要があると考えられる。

4 まとめと今後

本稿では、複数の韻律情報を組み合わせた韻律制御の方法を考案し、その方法に基づいて発声された音声が正しく判別されるかどうかの実験を行った。その結果、話者毎で正解率に多少のばらつきは見られたが、いずれの話者においてもある程度の判別が可能であることから、今回実験した韻律制御方法は正しく判別できる事がわかった。

今後は、文末表現やアクセント型と判別精度の関係を調べて、韻律制御方法あるいは判別モデルの改善を行い、そのうえでシステムへの実装を行う予定である。

謝辞

本研究の一部は、早稲田大学理工学総合研究センターの研究課題「マルチモーダル情報空間における統合的ヒューマンインタフェースに関する研究」によるものである。ここに記して謝意を表する。

参考文献

[1] 尾本 幸宏, 後藤 真考, 伊藤 克亘, 小林 哲則, "音声シフト: " SHIFT " on Speech," 情処研報 音声言語情報処理 2002-SLP-40-3, pp.13-18, 2002

[2] 後藤 真考, 伊藤 克亘, 速水 悟, "音声補完 : " TAB " on Speech" 情処研報 音声言語情報処理 2000-SLP-32-16, pp.81-86, 2000

[3] J.Ross Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers

音声対話における発話の感情判別*

◎小林 季実子 (早大・理工)　菊池 英明 (早大・人科)　白井 克彦 (早大・理工) †

1 はじめに

本研究では、対話音声に現れる感情判別モデルの構築を目標としている。感情判別モデルを統計的に学習するアプローチを念頭に置いた場合、学習データを準備する上で安定性や再現可能性の高い主観的感情評定方法が必要である [1]。そこで、まず適切な主観的感情評定方法を検討する。具体的には「音声データの呈示方法」と「時間の経過」が感情評定に対して与える影響を検討する。また、その考察に基づいて設定した感情評定方法を用いて複数の被験者が行った評定の結果を正解データとして決定木学習を試みる。

2 感情評定方法

感情評定に影響を与える要因を調べるため、以下の実験を行った。感情評定を行った被験者は研究室の男子学生 5 名である。

2.1 実験

本研究における感情の評定には *Mehrabian* が提案する感情の 3 次元 [2] のうち、対話音声によく現れると感じられる「快-不快軸」、「覚醒-無覚醒軸」の 2 軸を指標として用いた。以下にその説明を示す。

快-不快軸 : 肯定的あるいは否定的な感情
覚醒-無覚醒軸 : 能動的あるいは受動的な感情

また、各軸ごとにその状態の度合いを中心から両端に向かって「0:どちらとも感じない、1:少し感じる、2:はっきり感じる」の 5 段階で評定をし、記述するものとする。

使用した対話データは研究室で以前作成した対話コーパス [3]「旅行プランニングタスク」(学生同士の自然な雑談に近いもの) 等 2 種類、声優の対話による演技が収録されている *CD* 音声の計 3 種類である。各対話データのうち、それぞれ連続した 10 発話ずつを抜粋し 3 種類で 30 発話、計 90 発話の音声データを用いた。

音声データはすべてヘッドホンから呈示し、呈示方法は以下の 2 通りとした。

- コンテキスト呈示 (*C* 呈示:対話中の出現順に呈示)
- ランダム呈示 (*R* 呈示:対話をランダムに呈示)

音声データは *PC* 上で被験者自身が再生し、繰り返しての再生は許さないものとした。

また、評定機会は 1 つの呈示方法につき 3 回とした。まず 2 通りの呈示方法による評定を 2 回続けて行い、各呈示方法の残り 1 回ずつの評定は 1 週間以上後に行なうこととした。

2.2 実験結果

2.2.1 感情評定の安定性

まず評定の安定性を見るために、被験者全員の評定機会間の一致率を表 1 に示す。一致率の計算には *Cohen* の κ 値を用いた。表中で A は 5 段階評定を、$B1$ は「(2,1) 0 (1,2)」という分け方の 3 段階評定、$B2$ は「(2) (1,0,1) (2)」という分け方の 3 段階評定として 3 段階への再構成を行った場合を表す。

表 1: 評定における評定機会間の一致率 κ 値

(*a*) 快-不快軸

評定機会	C 呈示			R 呈示		
	A	$B1$	$B2$	A	$B1$	$B2$
1,2 回目	0.74	0.78	0.81	0.63	0.68	0.73
1,3 回目	0.61	0.74	0.64	0.50	0.63	0.59
2,3 回目	0.57	0.69	0.65	0.54	0.64	0.67

(*b*) 覚醒-無覚醒軸

評定機会	C 呈示			R 呈示		
	A	$B1$	$B2$	A	$B1$	$B2$
1,2 回目	0.65	0.67	0.77	0.39	0.49	0.57
1,3 回目	0.57	0.65	0.65	0.41	0.50	0.56
2,3 回目	0.57	0.56	0.75	0.46	0.54	0.66

快-不快軸・覚醒-無覚醒軸どちらにおいても、音声データをコンテキスト呈示した方がランダム呈示よりも一致率が高い。また、3 回目の評定機会との一致率は全体に低下の傾向が見られる。更に、評定の種類としては $B2$ が最も安定性が高いが、今後取り組む予定であるアプリケーションへの応用をふまえて以降では $B1$ を用いることにする。

2.2.2 感情評定の再現可能性

次に、評定の再現可能性を調べるために、評定機会ごとに 5 名の被験者間の評定結果の一致率を求めた。その結果を表 2 に示す。

*Discrimination of Emotion under Spoken Dialogue
†KOBAYASHI Kimiko, KIKUCHI Hideaki, SHIRAI Katsuhiko (Waseda University)

表 2: 評定機会ごとの被験者間の一致率 κ 値

評定機会	快-不快軸		覚醒-無覚醒軸	
	C 呈示	R 呈示	C 呈示	R 呈示
1 回目	0.49	0.39	0.27	0.30
2 回目	0.53	0.52	0.34	0.39
3 回目	0.42	0.43	0.30	0.29

快-不快軸・覚醒-無覚醒軸とも 1 回目より 2 回目の一致率が上昇し、3 回目には減少が見られる。2 回目の一致率が最も高くなっているのは、初めて評定を行った場合に 1 回目に比べて 2 回目の方が慣れにより安定するためと考えられる。

以上より、本研究では感情判別モデル構築への利用を考慮した上で安定性や再現可能性を高くするために、評定の練習を行った後コンテキスト呈示で 1 回だけ評定する方法を用いることにする。

3 感情判別のための決定木学習

2 章で述べた感情評定方法を用いて評定を行い、その結果より感情の自動判別を試みた。

3.1 データ

使用した音声データは、2.1 節で示した対話コーパス「旅行プランニングタスク」である。コーパス内各組の対話の中から比較的感情が分かりやすく現れていると感じられるものを 3〜4 発話ずつ連続して抜き出し、計 100 発話分使用した。

感情評定の際には、音声はヘッドホンから呈示した。感情評定を行った被験者は研究室の男子学生 11 名で評定には 2.2.1 項に示したように、快-不快軸・覚醒-無覚醒軸とも 5 段階の記述をしてもらった上で、各軸とも「2・1」の段階と評定されたものは同じ評定結果としてまとめ、「(2, 1) 0 (1, 2)」という 3 段階評定として再構成した。また、使用した 100 発話中 13 発話は相槌であったため分析の対象から除外し、最終的に計 87 発話を分析対象として用いた。各人の評定結果より多数決をとり、その結果を正解データとした。

3.2 特徴量

決定木学習のための特徴量の抽出には、*ESPS/waves+* を用いた。また、発話の最初と最後の 1 モーラ区間を抽出するため *julius* による自動アライメント結果を利用した。

抽出した特徴量は、1 発話全体・発話の第 1 モーラ ($S_$)・発話の最終モーラ ($L_$) それぞれに対して

- F0 最大値
- F0 最小値
- F0 平均値
- F0 レンジ
- パワー最大値
- パワー最小値
- パワー平均値
- パワーレンジ
- 発話速度 (モーラ数 / 秒)

の 9 種類ずつ計 27 種類で、発話者ごとに正規化した値を用いた。

3.3 決定木による判別

決定木学習には $C4.5$[4] を用い、信頼度 25% で行った。学習した決定木により *closed* なデータの判別を行い、得られた判別率と有効な特徴量を表 3 に示す。

表 3: 決定木学習による判別率と κ 値、有効な特徴量

	快-不快軸	覚醒-無覚醒軸
判別率	91.95 %	96.55 %
κ 値	0.87	0.94
特徴量	S_パワーレンジ S_F0 レンジ L_パワー平均値 L_パワー最小値	S_パワー最小値 パワー最小値 S_パワー最大値 パワーレンジ

快-不快軸・覚醒-無覚醒軸とも高い判別率、κ 値が得られている。感情判別を行う場合、快-不快軸では発話の第 1 モーラや最終モーラの特徴量が有効であり、覚醒-無覚醒軸では発話全体や発話の第 1 モーラのパワーが有効であることがわかった。

4 まとめ

感情判別モデルの統計的学習において重要な主観的感情評定方法に影響を与える要因を実験により確認した。また、その結果に基づいて設定した主観的感情評定方法を用いて評定した結果により決定木学習を行い、高い判別率が得られることを確認した。今後はデータ数を増やして実用的な感情判別モデルの構築を行い、ペット型ロボットに実装する予定である。

謝辞

本研究の一部は早稲田大学理工学部総合研究センターの研究課題「マルチモーダル情報空間における統合的ヒューマンインタフェースに関する研究」によるものである。ここに記して謝意を表する。

参考文献

[1] 森大毅, 相澤宏, 粕谷英樹 “対話音声のパラ言語情報ラベリングの安定性に影響を与える要因について”, 人工知能学会研究会資料,pp.45-50,SIG-SLUD-A401-08(2004-6)

[2] 福井康之著,“感情の心理学”, 川島書店,pp.83-85

[3] 岩野裕利, 杉田洋介, 松永美穂, 白井克彦 “対面および非対面における対話の違い〜頭の振りの役割分析〜”, 情報処理学会音声言語情報処理研究会資料,pp.105-112,15-19(1997-2)

[4] J.Ross Quinlan,“C4.5:Programs for Machine Learning Morgan Kaufmann Publishers”

モーションキャプチャデータを用いた多様な歩行動作制御手法の検討
A Study on Human Walk Controll Model Using by Motion Capture Data

板野　貴之[†]　森野　岳宏[†]　長船　大毅[†]　誉田　雅彰[†]　村上　真[‡]　白井　克彦[†]
Takayuki Itano Takehiro Morino Daiki Osafune Masaaki Honda Makoto Murakami Katsuhiko Shirai

1. はじめに

今日、コンピュータの著しい発達に伴い、映像コンテンツ制作の多様な手法が確立されている。映像コンテンツ制作における人物動作においては、動作を記述し、さらに演出による効果を付随する必要がある。特に演出による効果は、現在モーションキャプチャやキーフレームアニメーションなどを用いてCGアーティストが手作業で行っている。しかし、この手法では、コスト高やスキル不足、さらに仕事量や時間量といった問題が考えられる。

コンピュータ上の人物動作をより容易に表現可能とするためには、あらかじめ動作に応じてその拘束条件と自由度をモデル化し、それによって自然かつ多様な動作を生成・制御させる必要がある。動作生成におけるモデル構築には、過去では関数近似を用いたもの[1]、または人間の筋肉の収縮運動を考慮したものなどがある[2]。従来の研究は動作における人間の可動部位の運動をそれぞれ独立に考えており、可動部位の連動性に言及したものはない。人物動作はこの可動部位を連動させて制御させることから、本稿では、まず歩行動作に特化し、関節間の連動性を考慮し、多様な歩行動作を表現可能な歩行動作制御モデルを構築する。

2. 歩行動作制御モデル

2.1 定義

歩行動作制御モデルとは、簡単な操作で多様な歩行動作を表現可能としたモデルのことである。歩行動作において、「手を大きく振って歩く」や「速く歩く」といった物理的な変化を加えた歩行動作においては皆ほぼ共通の動きをするが、「嬉しそうに歩く」や「怒って歩く」といった感情を付加した歩行動作においては、人や状況に応じて様々な動作表現が考えられる。そこで、「普通に歩く」から「速さ」「大きさ」「姿勢」といった物理的な変化を付加することで、感情を含めたすべての歩行動作を表現可能なモデルの構築を試みた。

2.2 構築手法

歩行動作制御モデルは、2段階に分けて構築する。まず、各関節動作の連動性を分析することにより、歩行動作における人物の関節構造がどのようになっているのかを明確にする。次に、関節間の連動性を考慮し、モーションキャプチャで収録した物理的な変化を加えた歩行動作と「普通に歩く」とのスペクトルの差分を用いることにより、4つの動作制御特徴量を抽出する。歩行動作は、「普通に歩く」に強さを制御した4つの動作制御特徴量を加えることにより制御可能とした。

†早稲田大学理工学部, Waseda University, Department of Computer Science
‡東洋大学工学部, Toyo University, School of Engineering

2.3 収録歩行動作

歩行動作制御モデルを構築する際に必要となる歩行動作を、モーションキャプチャを用いて収録した。収録する歩行動作は以下の8種類であり、35人の被験者に対して収録を行った。

- 普通に歩く
- ゆっくり歩く
- 速く歩く
- 大きく手を振って歩く
- 大股で歩く
- 細かく歩く
- 胸を張って歩く
- うつむいて歩く

モーションキャプチャで収録された動作データには各関節の1フレームごとのローカルな座標系での回転角度（19関節×3軸方向）の情報が含まれており、本稿ではそのデータを扱う。

歩行動作における関節動作は全て周期的な運動をするため、それらの時系列データは周期信号として捉えることが可能である。よって、全関節動作の時系列データは式(1)のようにフーリエ級数で近似できる。

$$\Theta(t) = \sum_{n=-\infty}^{\infty} C_n e^{i\frac{2\pi n}{T}t} \tag{1}$$

また、過去の研究より、人物動作における関節動作は、全て第3高調波までのスペクトル情報で十分動作を再現可能であることが示されている[1]。従って、本稿では、全関節運動の時系列データを第3高調波までのスペクトル情報に集約しモデルを構築する。

3. 連動性の分析

人間の関節構造を簡略化するために、連動性の分析を行った。連動性の分析には、重回帰分析を用いる。図1に銃回帰分析を用いた連動性の分析手法を示す。

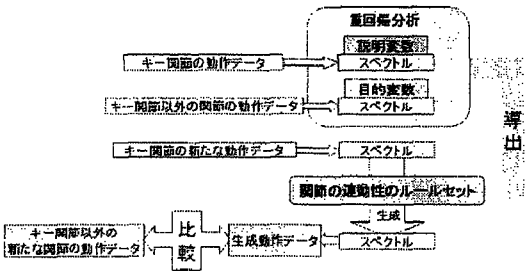


図1: 重回帰分析を用いた連動性の分析

オープンなデータを用いて、連動性のルールセットによる関節動作の再現度を相関係数を用いて導出した。再現度の高い関節動作ほど、キー関節動作との連動性が高いことが言える。表1に、分析結果の1つである右脚部位における再現度（相関係数）表を示す。

表 1: 再現度表 - 右脚部位

関節動作名	再現度
RightKnee-Z	0.8553
RightKnee-X	0.9410
RightKnee-Y	0.7048
RightAnkle-Z	0.6801
RightAnkle-X	0.8541
RightAnkle-Y	0.3531

ここで閾値を 0.5 とし、連動性の高い関節動作と連動性の低い関節動作に分類した。連動性の低い関節動作においては、その中でもう一度分析を行い、新たに分類を行った。その結果を図 2 に示す。

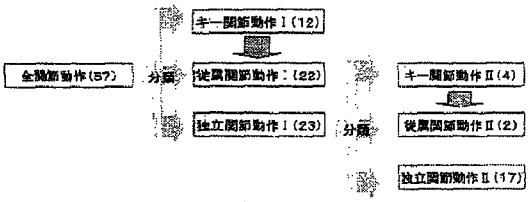


図 2: 関節動作の分類図

この結果を基に、連動性の高い関節動作をまとめると、上半身を制御する関節動作群と下半身を制御する関節動作群に分類できた。また、連動性の低い関節動作は主に身体の末端に位置する関節動作であることがわかった。

4. 歩行動作制御モデルの構築

4.1 動作制御特徴量の抽出

歩行動作を制御可能な動作制御特徴量の抽出法を表 2 に示す。まず、収録した各歩行動作の平均化を行う。そして、物理的な変化を加えた歩行動作と「普通に歩く」とのスペクトルの差分を導出する。各関節動作のスペクトルの差分の中で、物理的な変化に影響される関節動作のみを制御可能としたのが動作制御特徴量となる。

表 2: 特徴量の抽出

特徴量	歩行動作名	制御する関節動作グループ
速さ	ゆっくり歩く	上半身グループ + 下半身グループ
	速く歩く	上半身グループ + 下半身グループ
手の振り	細かく歩く	上半身グループ
	大きく手を振って歩く	上半身グループ
脚の振り	細かく歩く	下半身グループ
	大股で歩く	下半身
姿勢	うつむいて歩く	上半身グループ + 下半身グループ
	胸を張って歩く	上半身グループ + 下半身グループ

4.2 動作制御特徴量の制御による歩行動作の生成

動作制御特徴量の強さを制御可能とし、歩行動作制御モデルを構築した。歩行動作制御モデルによる生成歩行動作の動作データは (2) 式のように求まる。

$$\Theta(t) = \sum_{n=-3}^{3} \{C_n + \sum_{k=1}^{4} c_{nk}\gamma_k\} e^{i\frac{2\pi n}{T}t} \qquad (2)$$

5. モデル評価実験

本稿で構築した歩行動作制御モデルの評価を行った。このモデルの構築に使用していないオープンな動作データを用いて、このモデルからの再現性を調べた。

まず、計 33 動作（11 種類× 3 人）について、モーションキャプチャで収録した歩行動作と構築したモデルによって生成した歩行動作との類似度を、20 人の大学生に 5 段階（1. 全く似てない、2. 似てない、3. どちらとも言えない、4. 似ている、5. 非常に似ている）で心理評価してもらった。表 3 にその結果を載せる。

表 3: 各歩行動作における心理評価の平均

歩行動作名	心理評価
普通に歩く	4.200
ゆっくり歩く	4.300
速く歩く	4.267
大きく手を振って歩く	4.183
大股で歩く	4.017
細かく歩く	3.967
胸を張って歩く	3.933
うつむいて歩く	4.100
怒って歩く	2.017
楽しそうに歩く	2.433
悲しそうに歩く	3.067

次に、モーションキャプチャで収録した歩行動作と構築したモデルによって生成した歩行動作に対して、計 55 動作（11 種類× 5 人）における各関節動作の相関係数の平均を導出した。表 4 に左脚部位における関節動作の再現度を載せる。

表 4: 左脚部位における関節動作の再現度

関節動作名 (左脚部位)	再現度	
	物理動作	感情動作
LeftHip_ z	0.690	0.466
LeftHip_ x	0.980	0.945
LeftHip_ y	0.876	0.654
LeftKnee_ z	0.818	0.707
LeftKnee_ x	0.947	0.845
LeftKnee_ y	0.848	0.601
LeftAnkle_ z	0.844	0.728
LeftAnkle_ x	0.896	0.639
LeftAnkle_ y	0.309	0.369

表 3 から、物理的な変化を加えた歩行動作に関しては、ほぼ再現可能であることがわかる。しかし、感情を付加した歩行動作に関しては、十分に再現できたとは言い難い。表 4 を見ると、感情を付加した歩行動作を表現する際、再現度が大幅に低下した関節動作があることがわかる。これらは、以下の 3 点であることがわかった。

- 腕（Elbow・Wrist）の Y 軸方向
- 脚（Hip・Knee・Ankle）の Z,Y 軸方向
- 胸（Chest・Chest2）の X 軸方向

6. まとめと今後

本稿では、少ない指示により、多様な歩行動作を表現可能な歩行動作制御モデルの構築を試みた。モデルの構築にあたり、まず、各関節の 1 週期分の時系列データをフーリエ変換により第 3 高調波までのスペクトル情報に集約した。歩行動作において、連動性の高い関節群はまとめて制御可能とするために、重回帰分析を用いて全関節動作をいくつかのグループに分類した。この結果を基に、4 つの動作制御特徴量を抽出した。そして、その強さを制御することにより、動作制御特徴量の内挿・外挿表現、また、4 つの動作制御特徴量を補間することにより、様々な物理的な変化を持った歩行動作が表現可能となった。評価実験から、4 つの動作制御特徴量だけでは、歩行動作を十分に表現できないことがわかった。今後は、再現度の低下した関節動作を制御可能な動作制御特徴量を抽出し、再度モデルの評価を行う予定である。

参考文献

[1] 鵜沼宗利, 竹内良三, "コンピュータアニメーションにおける感情を伴った人間の歩行動作の生成方法", 電子情報通信学会論文誌,D Vol.J76-D pp.1822-1831,1998-08

[2] 西田論司, 杉浦彰彦, "アクションユニットを用いた人間の動作の分析・作成", 豊橋技術科学大学学会論文誌,J82-D-II,No.10,pp.1655-1665