■S3 群（脳・知能・人間）－ 感性・マルチモーダル情報処理

1 章 音の感性情報処理

執筆中

■感性・マルチモーダル情報処理 － 1 章

1-1

執筆中

■感性・マルチモーダル情報処理 － 1 章

1-2 音声の感性情報処理

（執筆者：藤澤隆史）［2009 年 9 月 受領］

ヒトの発声による音声情報は(1)言語情報，(2)パラ言語情報，(3)非言語情報の三つに大別される [1]．言語情報とは主に書き言葉（テキスト）として表される成分を指し，単語と統語によって文（発話文）を形成し状況を記述する．音声言語には言語情報以外にも様々な成分が含まれており，例えば声の大きさや高さなどがこれに当てはまる．これらの成分はパラ言語情報もしくは非言語情報と呼ばれる．パラ言語情報とは，意図や態度がそうであるように，発話者がある程度は意識的に制御することが可能な部分の情報のことである．それに対して非言語情報とは，性別や年齢がそうであるように，発話者が意識的に制御することができない部分によって表現されている情報のことである．パラ言語情報と非言語情報の区別は便宜的に有用な側面があるが，意識的制御の可能性による分類が困難な場合がある点についても注意が必要である．例えば，感情や性格，発話スタイルなどの特性は，社会的な相互作用場面の場合，対話者に対する印象制御の必要性によっても左右される．本節では，パラ言語情報と非言語情報を便宜的に用いつつも，両者を文脈に応じて厳密には区別せずに使用し，韻律（プロソディ）と感性情報の関連性及び周辺技術について概説する．

1-2-1 プロソディとパラ言語情報

音声言語には言語情報以外にも様々な成分が含まれており，代表的なものとしては，声の大きさ（ラウドネス），声の高さ（ピッチ），話速（スピーチレート），抑揚（イントネーション），間（ポーズ）などがあげられる．音声情報処理では，上であげたようなテキスト上では表現されえない情報をもっている諸成分及びその全体を指して韻律（プロソディ）と呼ぶ．言い換えれば，プロソディとは音の強弱，高低，長短などから構成される時系列パターンのことであり，それぞれは大まかに，音圧，基本周波数（F0），有声音区間（非無声音区間）といった音響特徴に対応づけられる．

プロソディによって表現される種々のパラ言語情報において，発話意図はその最も代表的なものとしてあげられる．例えば，言語情報レベルでは「そうですか」という単一表現であったとしても，語尾を上げればと疑問や非同意を表し，語尾を下げると平叙や同意を表すといったように，パラ言語レベルでは異なった意図をもって表現され得る．またプロソディは発話者の態度も表現する．例えば，目上の人と話すときには声が高くなるといった現象にもみられるように，全体的なピッチ水準の高低は，対話者に対する礼儀正しさ（ポライトネス）と関連し，社会的な地位関係を調節する機能をもっていることが明らかにされている [2]．また Ohara [3] は，上記であげたようなピッチの上昇下降，高低による意図や態度の表現には文化・民族によらない普遍性がある（サウンドシンボリズム）とし，生物学的な基盤との関連性について検討している．

1-2-2 プロソディと感情

表出された音声の音響特徴と感情カテゴリの対応関係については，1980 年代における Scherer [4] の検討を中心に組織的な検討がなされた．例えば，「怒り」の音声はピッチの水準

が高く，レンジが広く，ラウドネスが大きく，スピーチレートが速いのに対して，「悲しみ」ではほぼ真逆の傾向であることが分かっている．しかしながら，顔表情では物理的特徴量から基本 6 感情を識別することがほぼ可能であるのに対して，音声ではそれが困難であり，識別に至る決定的な音響特徴を見出せていないというのが現状である．今後，感情カテゴリを識別するための音響的手がかりとしては，ピッチ曲線（イントネーション）と声質という二つの音響特徴がその候補としてあげられよう．これまでの先行研究では，イントネーションや声質と各感情カテゴリの対応関係について検討した論文はごく少数にとどまっている [5]．イントネーションや声質とはパターンであり，定量的に評価することが困難であるという点から，これまで系統的な努力がなされてこなかった [6]．藤澤ら [7] は「協和－不協和」といった音響音楽理論に基づいた特徴定義を援用し，複数のピッチの相対的な関係から特徴づけられる「和音性」を新たな音響的手がかりとして，感情音声を定量的に評価するという試みを行っている．

1-2-3 対話における「間」の機能

発話の目的はコミュニケーションが原則であり，その点で発話の基本的単位は二者関係による対話である．対話において，ポーズ（間）は重要な機能をもつことが明らかにされている．対話において間がもつ基本的な機能としては，主に発話内容の区切りと発話交代（ターンテイキング）の表示の二つがあげられる．発話内容の区切り機能について説明すると，間のない会話が理解不可能であることからも分かるように，人の感覚記憶の容量には限界があるため，チャンク化による情報圧縮プロセスを必要とする．チャンクとは人が記憶する情報の単位であり，例えば，数列において 7426369863 という区切りのない提示は 9 チャンクであるのに対して，742-6369-863 という区切りのある提示は 3 チャンクへと縮約されるとされる．後者が前者よりも記憶が容易であることは明らかであり，発話における「間」はちょうど数列におけるハイフンの役割に相当する．

次に「間」のターンテイキング機能について説明すると，「間」は対話におけるターンテイキングを促す役割をもつことが明らかにされており [8]，ターンテイキングの規則を無視したり，交代のタイミングを乱すような話者は対話者として好まれない．更に近年の研究から，より協調的な対話では「間」や発話長の時間が二者間で同じような長さとなる（同調する）ことも明らかにされており [9]，円滑なコミュニケーションを促す感性システムの実現において，適切な「間」や同調を制御する技術開発は重要なポイントである．

1-2-4 音声の操作技術と感性研究

感性の法則性を明らかにするうえで，試料（ここでは音声）を用いた主観評価実験は必要不可欠であるため，その物理的特徴を操作する技術は同様に重要である．なかでもモーフィング技術は，その技術開発によって顔領域研究が飛躍的に進展したように，試料の物理的特徴と心理的属性の対応関係について明示的な知識をもつことなく中間的な性質をもつ新たな試料を提供するという点で，感性研究の核となる技術であるといえる．元音声の品質を損なうことなしに音声操作やモーフィングを可能とした代表的技術として Kawahara ら [10] の STRAIGHT があげられる．STRAIGHT では，音声を(1)基本周波数，(2)非周期成分から構成される音源情報と，滑らかに変化する(3)スペクトル包絡情報の 3 要素へと分解し，各要素に

対して操作を加えた後，再合成を行っている．篁ら[11]は STRAIGHT を用いて，3 感情（喜び，怒り，悲しみ）と無感情である平静の音声モーフィングを行い，それらを用いた認知実験によって各感情の弁別閾に関する検討を行っている．藤澤ら[12]もほぼ同様の手法を用いて，子どもにおける感情の弁別閾に関して検討を行っている．また豊田ら[13]は素人とプロの歌声のモーフィングを行い，素人の声質であるにも関わらずプロの歌い回しである歌声や，それとは逆にプロの声質でもあるにも関わらず素人の歌い回しである歌声であるデモンストレーションを実現している．内田[14]は，話者の基本周波数やスピーチレート，イントネーションの大きさや変化パターンを操作することで，聴取者が推測する話者の性格に対する印象がどのように変化するのかについて定量的に検討している．

■参考文献

1) 藤崎博也, “韻律研究の諸側面とその課題,” 音講論集秋季, pp.287-290, 1994.
2) S. W. Gregory and S. Webster, “A Nonverbal Signal in Voices of Interview Partners Effectively Predicts Communication Accommodation and Social Status Perceptions,” Journal of Personality and Social Psychology, **70**, pp.1231-1240, 1996.
3) J. J. Ohala, “The Frequency Code Underlies the Sound-symbolic Use of Voice Pitch,” In L. Hinton, J. Nichols, and J, J. Ohala(Eds.), “Sound Symbolism,” Cambridge University, pp.325-347, 1994.
4) K. R. Scherer, “Vocal Affect Expression: A review and a model for future research,” Psychological bulletin, **99**, pp.143-165, 1986.
5) L. Cosmides, “Invariances in the Acoustic Expression of Emotion During Speech,” Journal of Experimental Psychology: Human Perception and Performance, **9**, pp.864-881, 1983.
6) 宇津木成介, “音声による情動表出と非言語的な弁別手がかり,” 異常行動研究会(編), “ノンバーバル行動の実験的研究,” 川島書店, pp.201-217, 1993.
7) 藤澤隆史, 高見和彰, N. D. Cook, “感情的発話における音楽性: 基本周波数を用いた和音性の定量化について,” 認知心理学研究, **1**(1), pp.25-34, 2004.
8) 近藤富英, “ノンバーバル・コミュニケーション行動としてのポーズの機能と役割への一考察,” 信州大学人文科学論集＜文化コミュニケーション学科編＞, **40**, pp.129-136.
9) 長岡千賀, “対人コミュニケーションにおける非言語行動の 2 者間相互影響,” 対人社会心理学研究, 6, pp. 101-112, 2006.
10) H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, “Restructuring Speech Representations using a Pitch-adaptive Time-frequency Smoothing and an Instantaneous-frequency-based F0 Extraction: Possible role of a repetitive structure in sounds,” Speech Communication, **27**, pp.187-207, 1999.
11) 篁 一彦, 曽我部優子, 河原英紀, “表情と感情音声の知覚,” 信学技報, TL2005-13, pp.31-38, 2005.
12) 藤澤隆史, 時津裕子, 土居裕和, 篠原一之, “感情認知の感受性と性格傾向の関連性－音声刺激を用いた検討－,” 日本生理学雑誌.
13) 豊田健一, 片寄晴弘, 河原英紀, “STRAIGHT による歌声モーフィングの初期的検討,” 情処研報, 2006-MUS-64, pp.59-64, 2006.
14) 内田照久, “音声中の抑揚の大きさと変化パターンが話者の性格印象に与える影響,” 心理学研究, 76(4), pp.382-390, 2005.

■感性・マルチモーダル情報処理 － 1 章

1-3

執筆中

■感性・マルチモーダル情報処理 － 1 章

1-4 音楽の感性情報処理

（執筆者：藤澤隆史）［2009 年 9 月 受領］

音楽情報処理において，その基礎となる技術の柱の一つは，計算機に楽曲構造を理解させるという認識技術である．認識技術の進展は 1990 年代頃より隆盛となり，今もなお続いている．特に近年では音楽音響信号からの認識に関する技術の進展が目覚ましい．音楽における感性情報処理では，音楽情報処理に関する基盤技術を前提としながらも，人間が音楽から受けとるより高次の感性を対象としなければならない．音楽情報処理技術の詳細については 2 群 9 編を参照してほしい．本節では，まず音楽の構成要素と感性（主観的イメージ）の関連性について概観する．次に，人の感性と計算機による自動処理のインタラクションを利用した音楽の検索技術や操作技術について概観する．

1-4-1 音楽の構成要素とイメージ

楽曲の印象を決定づける構成要素としては様々なものが考え得るが，例えば低次の特徴量としては，音高（ピッチ），音の大きさ（ラウドネス），音色などがあげられ，また高次の特徴量としては，低次特徴の組合せによって，旋律（メロディ），和音（コード），和声（ハーモニー），テンポ，律動（リズム），調性などがあげられる．上述の音楽的特徴が，人にどのような印象を与えるのかについては，**表 1･4･1** に示されているように心理実験を通じて古くから検討されている [1]．近年では，Gabrielsson ら [2] が先行研究のメタ分析を行っており，同様の傾向が確認されている．

表 1･4･1　音楽の構成要素と印象の関連性

音楽的要素	品位のある ／厳粛な	悲しい ／重い	夢のような ／感傷的な	静かな ／優しい	優雅な ／輝かしい	嬉しい 明るい	興奮した ／高揚した	力強い ／雄大な
モード	短調 4	短調 20	短調 12	長調 3	長調 21	長調 24	－	－
テンポ	遅い 14	遅い 12	遅い 16	遅い 20	速い 6	速い 20	速い 21	速い 6
音 高	低い 10	低い 19	高い 6	高い 8	高い 16	高い 6	低い 9	低い 13
リズム	堅調 18	堅調 3	変調 9	変調 2	変調 8	変調 10	堅調 2	堅調 10
和 声	単純 3	複雑 7	単純 4	単純 10	単純 12	単純 16	複雑 14	複雑 8
旋 律	上昇 4	－	－	上昇 3	下降 3	－	下降 7	下降 8

※数字はそれぞれの感情カテゴリにおける相対的な重要性を示している．

音色による特徴は，音楽の場合，その大部分は用いられた楽器に由来すると考えられ，楽曲の印象を左右する大きな要因である．例えば，トランペットによる演奏は快活で能動的な印象を与えがちであるのに対して，オーボエによる演奏はのどかで牧歌的な印象を与える傾向にある．音色がもつ印象については，Kitamura ら [3] による 1960 年代からの一連の実験より 3 因子から構成されることが明らかにされている．第 1 因子は「美しい－きたない」，「うるおいのある－カサカサした」などの美的因子，第 2 因子は「迫力のある－もの足りない」，「大きい－小さい」などの迫力因子，第 3 因子は「金属性の－深みのある」，「高い－低い」など金属性因子であるとされている．音色の特徴を決定づける物理的要因は，一般的に，音

に含まれている倍音の振幅比と，発音から消音までの時間的変化によるとされているが，定量的な対応関係については不明な点も多く，その点を明らかにすることが音色研究における今後の課題である．

最後に，全体としての音楽作品としての印象次元については，谷口 [4] により 5 因子であることが明らかにされている．谷口は，先行研究で用いられた 50 語の形容詞を収集して 5 段階の評定尺度を作成し，複数の楽曲について 209 名の被験者に評定させた．その結果，第 1 因子として「明るい」，「楽しい」などの高揚因子，第 2 因子として「いとしい」，「優しい」などの親和因子，第 3 因子として「猛烈な」，「刺激的な」などの強さ因子，第 4 因子として「落ち着きのない」，「浮かれた」などの軽さ因子，第 5 因子として「崇高な」，「厳粛な」などの荘重因子が見出された．また，それぞれの因子について因子付加量の高かった順に 4 項目ずつを選び出し，計 24 項目からなる音楽作品の感情価測定尺度を構成している．

1-4-2 人の感性を利用した音楽の検索や操作

近年の急速な情報処理技術の進展に伴い，一昔前では困難であったが，莫大な情報量をリアルタイムで処理することが可能となってきた．処理のリアルタイム化は，人間と計算機が対話形式で協調して複雑な処理を行うという新たな情報処理形態を生み出し，ヒューマンコンピュータインタラクション（Human Computer Interaction : HCI）と呼ばれる分野の形成に至っている．以下では，音楽検索や音楽操作の場面においてユーザの感性という高次情報を利用した，HCI による音楽情報処理システムについて概観する．

人の感性情報を利用した音楽検索システムにおいて最も代表的なものとしては，歌唱やハミングを検索キーとして楽曲を検索するシステム（QBH : Query by Humming）があげられる [5]．QBH では，ユーザは音楽に関する明示的な知識をもつことなく希望の楽曲を検索することが可能となっている．また池添ら [6] は，楽曲を「音楽感性空間」と呼ばれる主観的な印象空間へとマッピングし，感性語を利用することでユーザが指定したイメージに近い楽曲を検索するシステムを構築している．次に音楽推薦システムでは，従来の推薦システムは他のユーザの評価を参考に評価を行う「協調フィルタリング」と楽曲の属性に基づいて評価を行う「内容に基づくフィルタリング」の二つに大別できる．協調フィルタリングによる推薦システムは商用などで既に実現されており，また内容に基づくフィルタリングでは音色やリズム [7]，楽器構成 [8] など音楽音響信号の類似度に基づいた検索技術が既に確立しているが，吉井ら [9] は両者の併用によるハイブリッド型の音楽推薦システムを開発している．

音楽操作では，Hashida ら [10] による事例参照型の演奏デザインシステムがあげられる．DTM による楽曲制作の最大の問題点は，楽譜情報の入力のみによって再現された演奏の不自然性にある．従来のシステムでは，ユーザが微妙なニュアンスをもつ自然な楽曲制作を実現するためには細部を手作業で修正していくというプロセスが不可欠であった．本システムでは，フレーズの指定はユーザが，フレーズ構造の階層化は計算機が行い，更に演奏表情については専用データベースから事例参照を行うことでユーザの負担を大幅に軽減しつつも細部の作りこみが可能となっている．同様に，歌声合成の分野においても，従来のシステムは DTM と同様の問題点を抱えていたが，中野ら [11] はユーザによる歌唱の入力を事例として参照することで，より自然で豊かな感情表現が可能な歌声合成システムを実現している．

■参考文献

1) K. Hevner, “The affective Value of Pitch and Tempo in Music,” American Journal of Psychology, **48**, pp.246-268, 1937.
2) A. Gabrielsson and E. Lindstr?m, “The Influence of Musical Structure on Emotional Expression,” In P. N. Juslin and J. A. Sloboda (Eds.), “Music and emotion: Theory and research,” Oxford University Press, New York, pp.223-248, 2001.
3) O. Kitamura, S. Namba, and R. Matsumoto, “Factor Analytical Study of Tone Color,” Proc. 6th Int. Congr. on Acoustics, A-5-11, 1968.
4) 谷口高士, “音楽作品の感情価測定尺度の作成および多面的感情状態尺度との関連の検討,” 心理学研究, **65**(6), pp.463-470, 1995.
5) A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, “Query by humming: Musical information retrieval in an audio database,” Proc. ACM Multimedia, **95**, pp.231-236, 1995.
6) 池添　剛, 梶川嘉延, 野村康雄, “音楽感性空間を用いた感性語による音楽データベース検索システム,” 情処学論, **42**(12), pp.3201-3212, 2001.
7) G. Tzanetakis and P. Cook, “Musical Genre Classification of Audio Signals,” IEEE Trans. Speech Audio Process., **10**, pp.293-302, 2002.
8) T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, “Instrogram: Probabilistic representation of instrument existence for polyphonic music,” IPSJ Journal, **48**(1), pp.214-226, 2007.
9) 吉井和佳, 後藤真孝, 駒谷和範, 尾形哲也, 奥乃　博, “ユーザの評価と音響的特徴との確率的統合に基づくハイブリッド型楽曲推薦システム,” 情処研報, 2006-MUS-66, pp.45-52, 2006.
10) M. Hashida and H. Katayose, “Mixtract: A directable musical expression system,” Proc. of Affective Computing and Intelligent Interaction (ACII2009), pp.xx-xx, 2009.
11) 中野倫靖, 後藤真孝, “VocaListener: ユーザ歌唱を真似る歌声合成パラメータを自動推定するシステムの提案,” 情処研報, 2008-MUS-75, pp.49-56, 2008.