解説

# 音声に含まれる感情情報の認識

## ——感情空間をどのように表現するか——*

赤 木 正 人（北陸先端科学技術大学院大学）**

43.71.−k; 43.72.−p

### 1. ま え が き

音声対話などの音声によるコミュニケーションでは，「何を話しているか」という言語情報だけではなく，これ以外の情報，例えば個人性（性別，年齢），感情・健康状態，声質などの言語以外の情報が多数送受される。これらの情報を多分に含む音声は，Expressive Speech と呼ばれている [1]。音声によるコミュニケーションでは，言語情報だけではなくこれらの情報にも重要な役割が含まれていると言われており，音声対話の精緻な解析のためには，これら双方を考慮する必要がある。本稿では，工学よりの目的（機械による感情の認識）を設定した上で，音声及び聴覚分野においてこれまでに得られた言語以外の情報の知覚に関する知見を取り混ぜながら，機械による感情の認識という目的に向かうための基本的考え方をどのように構成すれば良いかについて，思想まで踏み込んで解説する。

### 2. 言語以外の情報：パラ言語情報，非言語情報

まず，ことばの定義から始めよう。Fujisaki [2] は，音声に含まれる情報を次のように分類した。

**言語情報**：言語によって表記できるあるいは文脈によって一意に推測できる離散的情報

**パラ言語情報**：言語情報を変形あるいは補完するために話者によって付加される離散的もしくは連続的情報

**非言語情報**：話者の感情，性別，年齢のような話者によって一般には制御できない情報

### 3. 感情音声の認識

音声には，上述のように，言語情報以外に，パラ言語情報，非言語情報が含まれる。音声コミュニケーションではこれらが送受されている。このため，人–人の対話解析に基づいて人–機械のインタフェースを構築しようとする場合，言語情報だけではなく，話し手の特徴，特に感情がどのように変化しているかという情報は重要な要素となる。

近年，一層の国際化が進むにあたり，言語・民族・文化を越えた（グローバルな），また，言語・民族・文化のみならず老人，幼児，あるいは障害者との障壁のない（ユニバーサルな）コミュニケーションの重要性が増している。その中でも，感情の認識は，重要な要素となっている [3]。

現在，感情認識の研究は，音声関係で権威ある国際会議（ICASSP，InterSpeech 等）で多く発表されるようになってきた。2009 年度の InterSpeech では，チュートリアル及びスペシャルセッションで感情音声認識のセッションが組まれ，1 日以上このテーマが議論された [4]。

### 4. 感情認識に求められるもの

#### 4.1 感 情 空 間

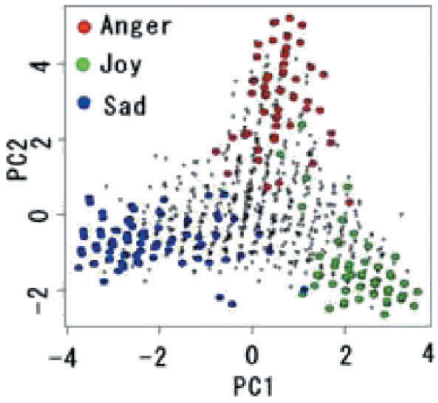
機械による感情の認識を考えるために，ヒトによって知覚された感情の性質についてまとめておこう。

感情を含む音声の聴取結果から構成した 2 次元感情空間の典型的な例を図–1 に示す [5]。この図は，次のような実験の結果が描かれている。

実験では，アニメ「ポケットモンスター」のキャラクタであるピカチュウの泣き声を基に櫻庭らが作成したデータベース [6] 中の，“怒り”，“悲しみ”，“喜び” 3 感情を意図した発話音声からなる 85 データを用いて，その音声をランダムに呈示した。聴取者は日本人大学院学生 17 名であり，発話ごとに

* Emotion recognition in speech: How do we describe an emotion space?

** Masato Akagi (Japan Advanced Institute of Science and Technology, Nomi, 923–1292)
e-mail: akagi@jaist.ac.jp




図–1　2 次元感情空間の例

含まれる “怒り”，“悲しみ”，“喜び” の感情成分についてそれぞれ独立に 5 段階評定をさせた。各音声サンプルに対する全被験者の “怒り”，“喜び”，“悲しみ” の認知感情を変数として相関行列を求め主成分分析を行い，累積寄与率 67%を占める 2 位までの主成分を抽出した結果が図–1 である。図では，上に “怒り”，右下に “喜び”，左下に “悲しみ” が分布した 3 角形状となっている。点数の高い評定を大きなドットで表現してあり，3 角形の頂点付近に分布している。

この図からも分かるように，

(1) 知覚された感情には度合いがある：

聴取者が容易に 5 段階の評定を行うことができたのが，「知覚された感情には度合いがある」何よりの証拠である。同じ感情であっても，その受ける印象，度合いは異なっている。ただし，評定された結果は心理量であり，絶対的な数値としての意味を持つものではなく，連続で曖昧な値となっている。“ちょっと”，あるいは，“かなり” 怒っているなど表現されるのが相応しい。

(2) 一つの発話に複数の感情が含まれる：

一つの単語，文から複数の感情が知覚されている。感情の度合いが強ければ一つの感情が知覚されるが，弱ければ複数の感情が知覚されている。

(3) 感情知覚空間は連続である：

高い評定を得ている音声データはカテゴリを形成しているように見えるが，他の多くの音声データは明確なカテゴリ構造を持っておらず，広く連続的に分布している。

という性質を持っている。

しかし，従来の感情認識の研究では，感情を言語情報と同様に離散的なカテゴリととらえ，従来型のパターン認識技術，すなわち音声認識・文字認識等で使用されてきた「入力を各感情カテゴリに振り分ける技術」が用いられてきた。特に音声認識では，Hidden Markov Model（HMM），Gaussian Mixture Model（GMM），Artificial Neural Network（ANN）等が用いられてきたため，感情認識においてもこれらを流用した研究が多く発表されている。

### 4.2 感情認識の特殊性

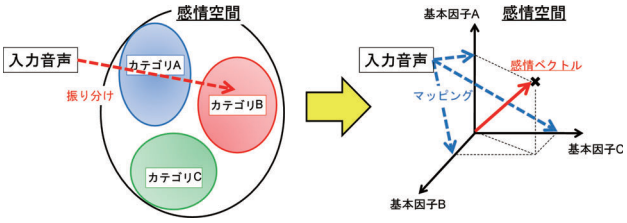
ここであらためて次の問いを発したい。「感情（例えば “怒り”，“喜び”，“悲しみ”）はカテゴリか？」もしカテゴリならば，従来から音声認識に用いられている HMM，GMM などの手法が効率よく使用できるはずである。しかし，これらの方法が感情認識本来の目的を達成しているかどうか甚だ疑問である。

上述したように，人が音声中の感情を知覚する場合，同じ感情（例えば怒り）でも「少し怒っている」から「かなり怒っている」というように知覚された感情の程度は連続的に変化し，しかも，一つの発話文から「怒っているけど悲しそうだ」などのように複数感情が同時に知覚されることもありうる。このことは，感情認識においては，各感情は従来のパターン認識が対象としているような単純なカテゴリ構造を持っておらず，現有の感情認識システムのように感情を有限個のカテゴリとして捉えることはかえって感情認識の本質を捻じ曲げてしまうことを意味する。このため，機械による感情認識においては，複数の感情を同時にその程度までを含めて認識するシステムを構築する必要がある。「同時に複数の感情の度合いを含めた認識」を実現するためには，従来のカテゴリ判別器ではなく，新しい発想の認識手法を考えなければならない。

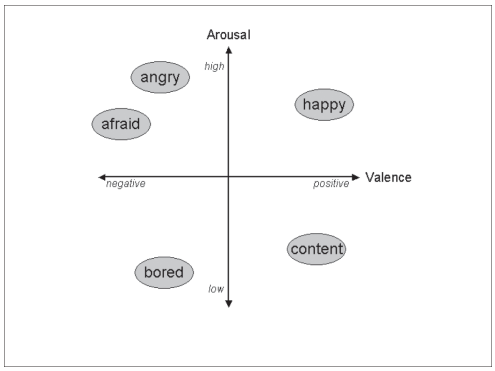
## 5. 感情空間の表現法

本章では，感情空間の新たな表現法である，感情基本因子ベクトルの合成ベクトルとして感情を表現する手法について解説する。

図–2 に概念図を示す。従来の感情認識システムが感情をカテゴリとして捉えていた（図–2 左）のとは異なり，感情空間は多数の感情基本因子ベクトルによって張られる連続した多次元空間として捉える（図–2 右）。そして，音声に含まれる物理




図–2　感情空間の再定義及び認識方略の変更
基本因子が張る空間として感情を定義。




図–3　Arousal（Activation）–Valence（Evaluation）空間の概念図。典型的な感情が上書きされている。

的音響特徴から個々の感情基本因子ベクトルへのマッピングを行い，感情基本因子ベクトルの合成ベクトルとして感情を表現する。このためには，感情空間を張る元（感情基本因子ベクトル）をどのように見つけるか，また，入力音声から抽出された音響特徴をどのように基本因子にマッピングするのか，を考察する必要がある。

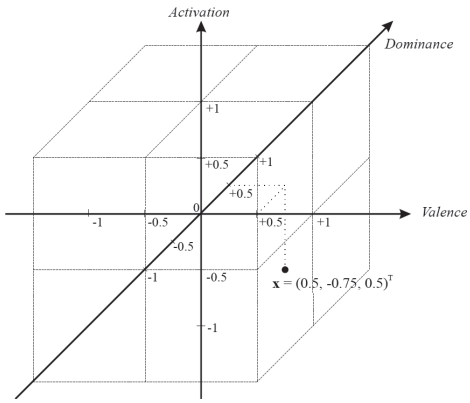
### 5.1　感情空間—2 次元空間の場合—

心理学者である Schlossberg は，1954 年に，顔表情の知覚に関する検討から，“Three dimensions of emotions” と題する論文を発表している [7]。この中で，第 1 次元は Sleep–Tension の次元であり，Tension が大きくなるときに感情が知覚され，残りの 2 次元（Pleasantness–Unpleasantness 及び Attention–Rejection）で様々な感情が説明できるとした。

Cowie らは，Schlossberg のモデルを受け，感情空間を Arousal（あるいは Activation）の次元と Valence（あるいは Evaluation）の次元の 2 次元空間と考え，Activation–Evaluation 空間と呼んだ [8, 9]。Vogt らがまとめた Activation–Evaluation 空間の概念図を図–3 に示す [10]。

### 5.2　感情空間—3 次元空間の場合—

Grimm らは，Cowie らが提案した Activation–



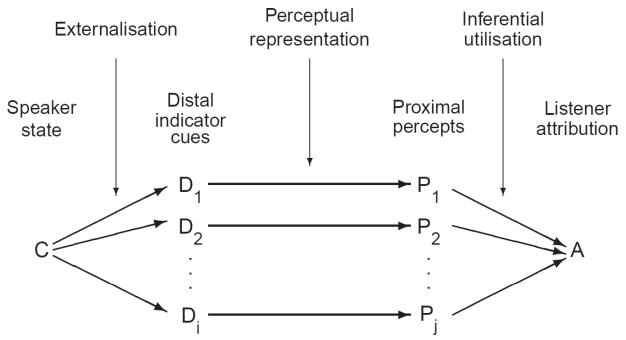

図–4　Activation–Evaluation–Dominance 空間の概念図

Evaluation の 2 次元空間では “怒り” と “恐れ” の違いをうまく表現できないため，新たに Dominance（Strong vs. Weak）の次元を加え [11]，これらの次元をお互いに直交する（無相関である）と見立て，図–4 に示す直方体として感情空間を表現している [12]。

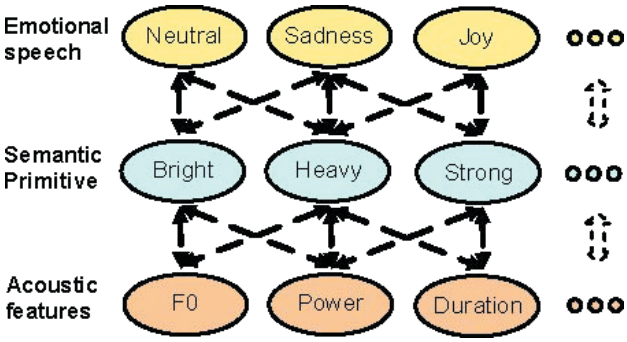
Schroeder は，表現豊かな音声の特質を扱う目的で，“怒り”，“恐れ”，“喜び” などのラベルではなく，感情空間の表現として三つの次元を用いることを提案している。それは，Cowie らが提案した Activation の次元と Evaluation の次元に加えて，支配，優越，社会的地位などの社会とのかかわりに関係する Power の次元である [13, 14]。Power の次元は，Grimm が提案している Dominance の次元とほぼ同じものである。

感情音声を認識する場合，入力音声の音声特徴とそれぞれの次元の関係を明らかにしておく必要がある。ここでは，ホルマント周波数と Activation–Evaluation–Dominance の次元との関係を論じた論文を紹介する。

Goudbeek らは，様々な感情をこめて発話された母音/a/，/i/，/u/について，Activation，Evaluation，Dominance それぞれの聴取実験による評価結果と第 1，第 2 ホルマント周波数（$F_1$，$F_2$）の関係を議論している [15]。Arousal（Activation）が高い場合はすべての母音で $F_1$ が高くなり，特に/a/の場合は $F_2$ が低くなる。Valence（Evaluation）が正の方向となると $F_2$ が上昇する。また，Power（Dominance）が大きい場合は/a/と/i/の $F_1$ が上昇し/u/の $F_2$ が下降する。このように，Activation–Evaluation–Dominance それぞれの次元で，ホルマントの特徴的な変化が観測できる。




図–5 Brunswik のレンズモデルを参考とした Scherer の感情知覚モデル。[13] から引用。




図–6 感情音声知覚の多層構造モデル

### 5.3 感情空間—多次元空間の場合—

Scherer は，Brunswik のレンズモデルを参考に，話し手から聞き手への感情の伝達を検討している [16]。図–5 に，Scherer のモデルを示す。Scherer のモデルでは，話し手の感情が知覚されるとは，話し手の感情に含まれる多数の手がかり（Distal indicator cues）が，聞き手の主観的な知覚（Proximal percepts）として表現され，それらが統合されることで，聞き手の属性（この場合は知覚された感情）が決まる。聞き手の主観的な知覚の例として，ピッチとか声質の知覚がある。すなわち，Scherer のモデルでは，個々の手掛かりから知覚された Proximal percepts が感情基本因子となり，これらが表す多次元空間として感情空間が表現されている。

### 5.4 多次元空間としての感情空間の構成例

感情空間を多次元として表現する一つの方法として，個々の基本因子を形容詞によって表現し，その統合として感情空間を記述する方法がある。本節では，Huang と赤木が行った，多次元空間としての感情空間の構成についての研究 [17] を紹介する。

#### 5.4.1 怒った声はどんな声？

例えば，「怒った声はどんな声？」と聞かれたときに，読者の方々はどのように答えるだろうか？怒った声は，高域パワーが○○ dB 大きくなった声と答えるだろうか？ 確かにこの答えは正しいかもしれないが，声質を正しく反映した答えとは言いがたいし，誰もこのようには答えないだろう。恐らくは，大きな声とか甲高い声とか答えるのではないだろうか。このように声質はことばで表現されることが多いため，“怒った声” などの感情と入力音声の音声特徴との間は，ことばを介して結びつけるのが自然である。ただし，どのようなことばでも良いかというとそうではない。感情にふさわしい形容詞を選び出し，この形容詞と “怒った声” の関係，及び，この形容詞と音声特徴の関係を考える必要がある。更にことばの対応関係の曖昧性をも表現できるモデルとするべきである。

#### 5.4.2 感情知覚の多層モデル

上記の仮定をもとに，次のような感情知覚の多層モデルを提案した。概念図を図–6 に示す。モデルは，(1) 感情（Natural, Sad, Joy etc.）を形容詞で表現された聞き手の主観的な知覚（semantic primitives）で説明すると共に，(2) 形容詞と音声特徴の関係を説明し，(3) 感情と音声特徴を関連付ける，というコンセプトで構成されている。

#### 5.4.3 モデルの構築

目的としている感情を形容詞で表現される semantic primitive に分解し，これらの関係を記述する。ここでは多次元尺度構成法と多重回帰分析を用いて形容詞を選択する方法，及び，Fuzzy Logic を用いて関連性を記述した例を示す。

A) モデルの構築：第 1 層から第 2 層へ

5 種類（Normal, Joy, Sad, Cold-Anger, Hot-Anger）の感情を意図してプロの声優により発話された日本語感情音声データベース（富士通研究所作成）を用意した。聴取者にこれらの音声がどのくらい感情を表しているかについて点数付けを行ってもらい，各感情で最高，中間，最低の点数を得た音声，計 15 個を刺激音声として採用した。これらの音の対比較実験結果に多次元尺度構成法（MDS 分析）を適用して知覚的距離空間を構成する。聴取者はすべて日本人である。

形容詞を選択するために，過去の音質表現語の研究結果 [18] から 34 個の形容詞を用意し，MDS

で構築した知覚的距離空間へ多重回帰させることにより相関が高い 17 個（英語表記：bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast and slow）を semantic primitive として採用した。すなわち，感情空間を 17 次元基本因子で表現したこととなる。

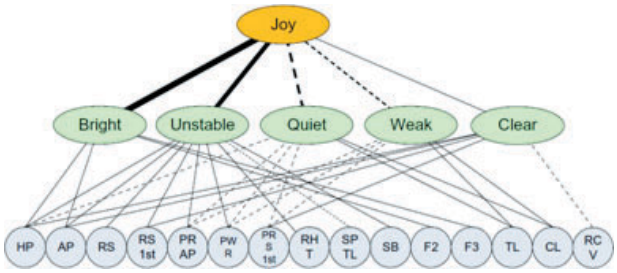
B) 感情音声モデルへの Fuzzy Logic の導入

基本的な心理特徴はことばで表現されているが，モデルの構築に際しては，関連性を数学的に記述する必要がある。そこで，形容詞と感情の関係を記述可能な Fuzzy Logic を用いることとする。実際には Fuzzy Logic Interface System（FIS）を用いて，各形容詞と感情の関係を記述する。FIS を用いれば，ある形容詞の印象が強まったときに，出力である感情がどのように変化するかが予測可能となり，結果として，どの形容詞が感情と強い関係（正及び負の関係を含む）を持つのかが推定できる。

C) モデルの構築：第 2 層から第 3 層へ

形容詞と音響特徴の関係を記述する。音声特徴候補を選択するために，$F_0$ 包絡，パワー包絡，パワースペクトラム及び発話長を分析し，それぞれ 8 個，8 個，7 個，3 個の計 26 個の音響特徴候補を用意した。各音声特徴と形容詞の印象の強さの相関値を計算し，0.6 を超えるものについて，その音声特徴が形容詞に関係していると判断した。結果として，16 個の音声特徴を採用した。$F_0$ 包絡（$F_0$ の最大値：HP，$F_0$ の平均値：AP，$F_0$ 上昇の傾きの平均値：RS，第 1 句での $F_0$ 上昇の傾き：RS1st），パワー包絡（アクセント句でのパワーレンジの平均値：PRAP，パワーレンジ：PWR，第 1 句でのパワー上昇の傾き：PRS 1st，3 kHz 以上での平均パワーと全周波数での平均パワーの比：RHT），パワースペクトラム（第 1 ホルマント周波数：$F_1$，第 2 ホルマント周波数：$F_2$，第 3 ホルマント周波数：$F_3$，スペクトルの傾き：SPTL，スペクトルの重心：SB），発話長（文の時間長：TL，子音の区間長：CL，子音と母音の区間長の比：RCV）である。

本章で示した手法を感情 “喜び（Joy）” に適用した例を示す（**図–7**）。図では，実線が正の関係，破線が負の関係を表している。また，線幅は関係の強さを表している。Joy 音声は主に 5 次元で表




**図–7** 感情音声 Joy のモデル構築結果
実線が正の関係，破線が負の関係を表している。また，線幅は関係の強さを表している。

現され，bright，unstable，clear，であり，quiet 及び weak ではない音声となる。

本モデルを感情音声の合成に適用した結果，感情を連続的に，しかも，その度合いを制御できる合成手法が実現できた [17]。

## 6. 感情空間表現法の感情認識への応用

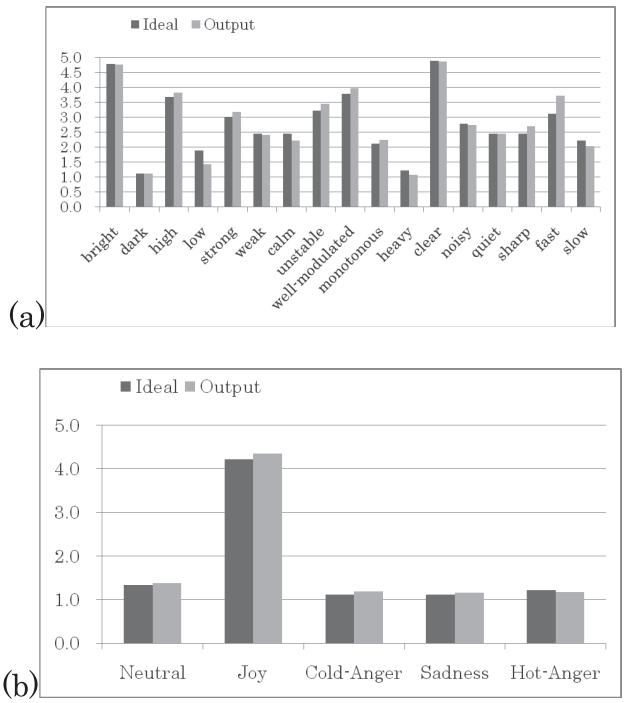
感情空間を表現する次元の属性が見つけられたとしても，感情認識を行う場合には，入力音声の特徴量からそれぞれの軸上の対応する値へのマッピングを行う必要がある。すなわち，音声特徴をどのように各次元の軸へマッピングするかが問題となる。線形重回帰モデル，非線形マッピングを考慮したニューラルネットワークなどが試されているが，ここでは，ことばの対応関係の曖昧性を表現できるマッピング法として，ファジィ論理（FIS）を用いた手法を紹介する。

Grimm ら [12] は，入力音声から基本周波数，発話速度，パワー及び声質に関わる 46 個の音声特徴を抽出し，主成分分析によって次元を圧縮したのち，FIS を用いて，3 次元感情空間 Activation–Evaluation–Dominance それぞれの次元で，メンバシップ関数を設計している。メンバシップ関数を表現する形容詞は

Activation: calm; neutral; excited
Evaluation: negative; neutral; positive
Dominance: weak; neutral; strong

である。

感情認識に多次元での感情空間表現を応用した例として，感情知覚の多層モデル [17] を用いた赤木らの研究 [19] を紹介する。赤木らは，図–7 に示す感情知覚の多層モデルについて，入力音声から得られる 16 個の音声特徴から 17 個の semantic primitive それぞれを予測する 17 組の FIS を構




図–8 (a) 17 個の semantic primitive の予測結果，(b) 5 感情の予測結果。縦軸は 5 段階評価（1〜5）の評点。

築し，その後，17 個の semantic primitive から五つの感情を予測する 5 組の FIS を構築した。入力が “喜び（Joy）” であった時の結果を図–8 に示す。図–8(a) が 17 個の semantic primitive の予測結果であり，図–8(b) が 5 感情の予測結果である。それぞれの項目で左の棒線が聴取実験の結果，右の棒線が予測結果である。かなりの一致を見ている [19]。このシステムでは，各感情を独立に，また，連続値として予測できるので，一つの発話に複数の感情が含まれる場合にも，個々の感情の度合いも含めて推定が可能である。

## 7. ま と め

本稿では，機械による感情の認識に向けて，音声及び聴覚分野においてこれまでに得られた言語以外の情報の知覚に関する知見を取り混ぜながら，感情空間をどのように表現するのかを中心に解説を行った。

感情認識は，従来の記号（カテゴリ）へのマッピングとは多くの点で異なる。この解説が，注意喚起の一助となれば幸いである。

### 文 献

[1] D. Erickson, “Expressive speech: Production, perception and application to speech synthesis,” *Acoust. Sci. & Tech.*, **26**, 317–325 (2005).
[2] H. Fujisaki, “Prosody, information, and modeling — With emphasis on tonal features of speech,” *Proc. Speech Prosody 2004 Nara*, pp. 1–10 (2004).
[3] 総務省情報通信審議会,「我が国の国際競争力を強化するための ICT 研究開発・標準化戦略」, 情報通信審議会答申, 平成 20 年 6 月 27 日 (2008).
[4] *Proceedings of InterSpeech 2009*, Brighton, UK, CD-ROM (2009).
[5] K. Sawamura, J. Dang, M. Akagi, D. Erickson, A. Li, K. Sakuraba, N. Minematsu and K. Hirose, “Common factors in emotion perception among different cultures,” *Proc. ICPhS 2007*, 2113–2116 (2007).
[6] 櫻庭京子, 今泉 敏, 筧 一彦, “「ぴかちゅう」にこめられた感性情報,” 音声研究, **8**, 77–84 (2004).
[7] H. Schlosberg, “Three dimensions of emotion,” *Psychol. Rev.*, **61**, 81–88 (1954).
[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. Taylor, “Emotion recognition in human-computer interaction,” *IEEE Signal Process. Mag.*, **18**, 32–80 (2001).
[9] R. Cowie and R. Cornelius, “Describing the emotional states that are expressed in speech,” *Speech Commun.*, **40**, 5–32 (2003).
[10] T. Vogt, E. André and J. Wagner, “Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realization,” in *Affect and Emotion in HCI*, C. Peter and R. Beale, Eds. (Springer, Berlin/Heidelberg, 2008), pp. 75–91.
[11] M. Grimm, K. Kroschel, E. Mower and S. Narayanan, “Primitives-based evaluation and estimation of emotions in speech,” *Speech Commun.*, **49**, 787–800 (2007).
[12] M. Grimm and K. Kroschel, “Emotion estimation in speech using a 3D emotion space concept,” in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. (I-Tech Education and Publishing, Vienna, 2007), Chap. 16.
[13] M. Schröder, “Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis,” *Doct. thesis, Phonus 7, Res. Rep. Inst. Phonet., Saarland Univ.* (2004).
[14] M. Schröder, “Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions,” in *ADS 2004*, E. André *et al.*, Eds. (Springer, Berlin/Heidelberg, 2004), pp. 209–220.
[15] M. Goudbeek, J.P. Goldman and K.R. Scherer, “Emotion dimensions and formant position,” *Proc. Interspeech 2009*, pp. 1575–1578 (2009).
[16] K.R. Scherer, “Personality inference from voice quality: The loud voice of extroversion,” *Eur. J. Soc. Psychol.*, **8**, 467–487 (1978).
[17] C-F. Huang and M. Akagi, “A three-layered model for expressive speech perception,” *Speech Commun.*, **50**, 810–828 (2008).
[18] 上田和夫, “音色の表現語に階層構造は存在するか,” 音響学会誌, **44**, 102–107 (1988).
[19] M. Akagi, “Analysis of production and perception characteristics of non-linguistic information in speech and its application to inter-language communications,” *Proc. APSIPA 2009*, Sapporo, pp. 513–519 (2009).