感情音声知覚モデルの提案とその応用*

○赤木正人（北陸先端大）

1　まえがき：研究のねらい

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．（図 1）

音声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，たとえば個人性（性別，年齢），感情・健康状態，声質などの非言語情報が多数送受される．非言語情報を多分に含む音声は，Expressive Speech と呼ばれている[1][2]．非言語情報の送受が音声コミュニケーションにおいて重要な要素であるならば，言語・民族・文化を越えたユニバーサルコミュニケーションのために，なおさら非言語情報の送受を深く考えるべきである．すなわち，Expressive Speech について基礎的に探求し非言語情報についての音声コミュニケーションを解明することが，言語を越えたグローバルでユニバーサルな音声コミュニケーション環境構築の一助となる．

ところが，言語・民族・文化が異なる人々の間で，音声中に含まれるどのような情報が共有されるのか，また，どのような非言語情報がこれらの人々のコミュニケーションにとって重要であるのかは定かではない．筆者らは，この疑問にこたえるために，次の二つの問題を中心に据えて研究を行っている．

問題１：人ー人の音声コミュニケーションにおいて，音声知覚・生成はその根幹を成す．また，人ー機械コミュニケーションにおいても，ヒトの音声生成・知覚機構を工学的に実現した音声合成・認識システムが重要な役割を果たそうとしている．このため，Expressive Speech の研究においても，Expressive Speech の知覚・生成の総合的な解明，さらには工学的応用に貢献でき得る知見の獲得が必要となってくる．しかし，ヒトの音声生成・知覚機構は未だ解明途上であり，ユニバーサルな音声コミュニケーション環境実現へ貢献し得る知見はまだまだ少ない．

問題２：言語・民族・文化を超えた非言語情報でのユニバーサルコミュニケーションが可能となるためには，Expressive Speech の生成・知覚において言語・民族・文化によらないヒトの生物学的「共通要素」，すなわち，生成のための万国共通の構音運動，共通の構音運動から作り出される共通の音声特徴，音声特徴を呈示することにより生起される共通の知覚特徴・脳活動，そして，この上に立つ人間の共通の行動が存在しなければならない．しかし，未だこの点についての有用な議論はなされていない．

音声の生成と知覚は不可分であり，しかも図 2 に示すように環構造となっている．これらの問題を解くためには，環構造の中でのそれぞれの関係を考慮しながら研究を進める必要がある．このため筆者らは，脳と音声生成の相互作用，脳と音声知覚の相互作用，生成から音声特徴を経て知覚への経路，それぞれの中で，Expressive Speech の生成・知覚機構の解明を目指している．さらに，音声コミュニケーションの言語・民族・文化を超えたグローバル化を指向して，言語・民族・文化によらない Expressive Speech の生成・知覚機構の共通要素とは何かについて検討を行っている．そして，この共通要素を核として，非言語情報の合成・認識を試みている．

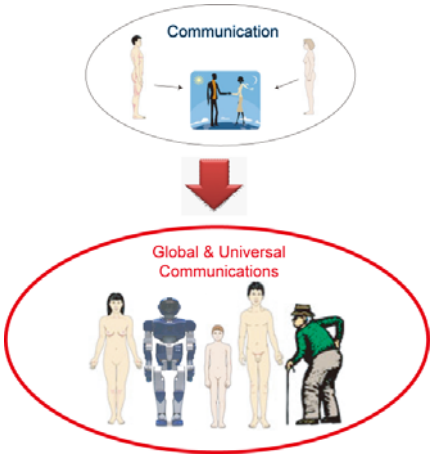


図 1　グローバルコミュニケーションに向けて

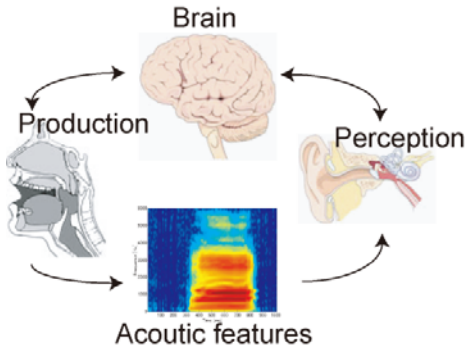


図 2　音声生成・知覚の環構造

本スペシャルセッションでは，筆者らが取り組んでいる課題の一つである Expressive Speech（特に感情音声）の知覚機構について，そのモデルである感情知覚の多層構造モデルを紹介し，その応用として，感情の程度を制御できる感情音声合成法を説明する．

2　感情知覚の多層構造モデル

感情音声知覚モデルを工学的に使用できるモデルとするためには，知覚機構の動作の説明のためだけ

* Introduction of a model for emotion perception in speech and its applications, by AKAGI, Masato (Japan Advanced Institute of Science and Technology).

の記述用モデルではなく，シミュレーションも可能な，アルゴリズムとしてインプリメントできるモデルの構築が必要である．筆者らはこの考えにもとづいて，次のような感情音声知覚モデルを構築した[3].

2.1 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，聴覚印象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．おそらくは，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，聴覚印象はことば（形容詞）で表現されることが多いため，「怒った声」などの非言語情報と物理量を扱う信号処理との間は，ことばを介して結びつけるのが自然であろう（図 3）．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．心理印象にそった形容詞を選び出し，このことばと「怒った声」の関係，および，この単語と音響特徴の関係を考える必要がある．さら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．

2.2 聴覚印象の多層モデル

筆者らは，2-1 節で述べた仮定をもとに，次のような聴覚印象の多層モデルを提案した．概念図を図 4 に示す．モデルは，(1) 上位の心理的特徴（感情(Natural, Sad, Joy etc.)）を基本的な心理特徴(semantic primitives)で説明するとともに，(2) 基本的な心理特徴と物理的音響特徴の関係を説明し，(3) 音声の音質についての聴覚印象と物理量を関連付ける，というコンセプトで構成されている．

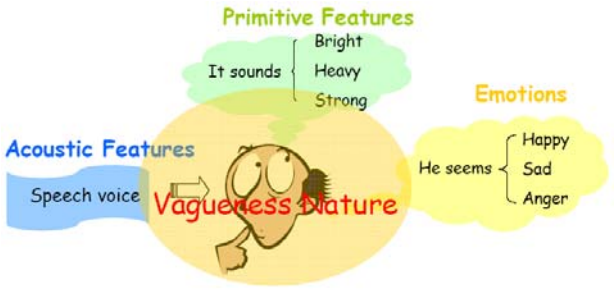


図 3　感情音声知覚の概念図

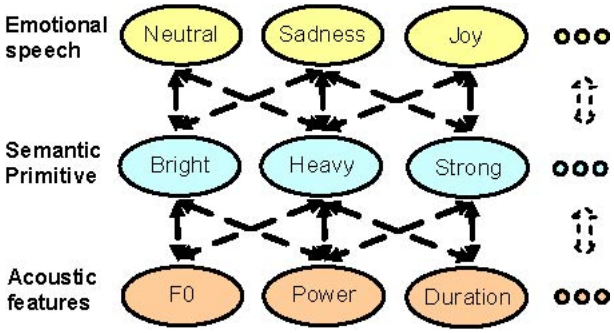
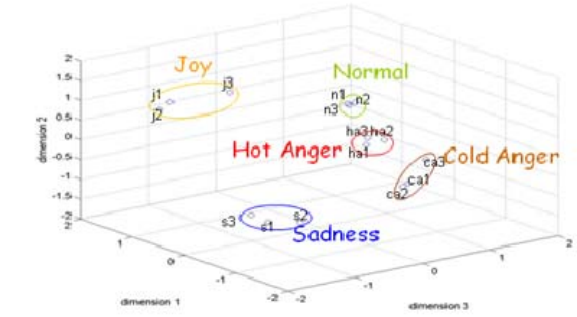
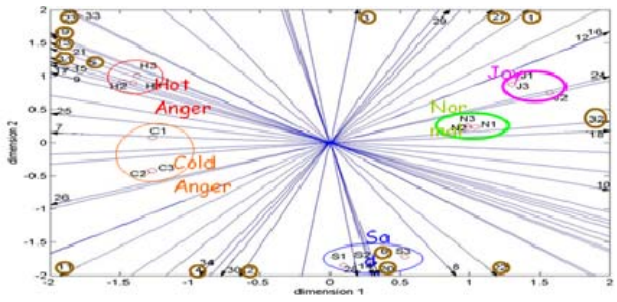


図 4　感情音声知覚の多層構造モデル



(a)



(b)

図 5　(a) 5 感情の知覚的距離空間上の布置，および，(b) 心理特徴候補の多重回帰直線

2.3 三層モデルの構築

目的としている聴覚印象を基本的な心理特徴（形容詞）に分解し，これらの関係を記述する手法の例を紹介する．ここでは多次元尺度構成法と多重回帰分析を用いた方法，および，Fuzzy Logic を用いた関連性の記述例を示す．

2-3-1 モデルの構築：Layer-1 から Layer-2 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おこなってもらい，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．

基本的な心理特徴を選択するために，過去の音質表現語の研究結果[4]から 34 個の心理特徴候補を用意し，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 個を基本的な心理特徴（英語表記：bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast and slow）として採用した．図 5(a)に 5 感情の知覚的距離空間上の布置，図 5(b)に知覚的距離空間上に 34 個の心理特徴候補を重ね合わせた結果を示す．

2-3-2 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，モデルを計算機上に構築して信号処理システムとして働かせるためには，関連性を数学的に記述する必

表 1　各感情と関係の強い基本的な心理特徴. PF : 基本的な心理特徴. S : FIS で予測される関係の強さ.

Neutral		Joy		Cold Anger		Sadness		Hot Anger	
PF	S	PF	S	PF	S	PF	S	PF	S
heavy	-0.329	quiet	-0.039	slow	-0.231	sharp	-0.079	calm	-0.063
weak	-0.181	weak	-0.036	monotonous	-0.073	strong	-0.049	quiet	-0.047
clear	0.127	unstable	0.063	fast	0.153	weak	0.065	unstable	0.120
monotonous	0.270	bright	0.101	heavy	0.197	heavy	0.074	well-modulated	0.124
calm	0.103	clear	0.034	well-modulated	0.091	quiet	0.057	sharp	0.103

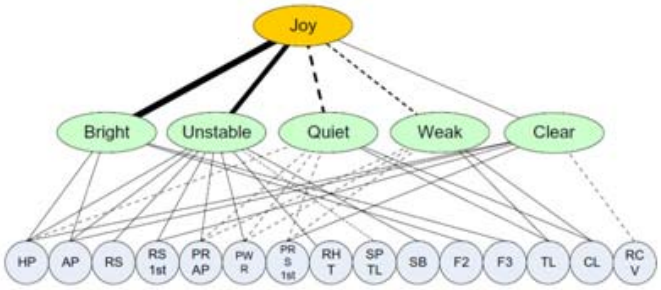


図 6　感情音声 Joy のモデル構築結果．実線が正の関係，破線が負の関係を表している．また線幅は関係の強さを表している．

要がある．そこで，自然言語によって基本的な心理特徴と感情の関係を構築可能な Fuzzy Logic を用いることとする．実際には Fuzzy Interface System (FIS) を用いて，基本的な心理特徴と聴覚印象の関係を記述する．FIS を用いれば，ある基本的な心理特徴が強まったときに，出力である感情の聴覚印象がどのように変化するかが予測可能となり，結果として，どの基本的な心理特徴が感情の聴覚印象と強い関係（正および負の関係を含む）を持つのかが推定できる．表 1 に，各感情と関係の強い基本的な心理特徴を上位 5 位まで示す．表中の数字が正の時は正の相関（負の時は反対）を持つことを表している．感情 Joy を例にとれば，Joy 音声は bright, unstable, clear, であり，quiet および weak ではない音声となる．

2-3-3 モデルの構築 : Layer-2 から Layer-3 へ

基本的な心理特徴と音響特徴の関係を記述する．音響特徴候補を選択するために，STRAIGHT[5]を用いて F0 包絡，パワー包絡，パワースペクトラムおよび発話長を分析し，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. 各音響特徴と基本的心理特徴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，その音響特徴が基本的心理特徴の印象に関係していると判断した. 結果として，16 個の音響特徴を採用した．F0 包絡（F0 の最大値:HP, F0 の平均値:AP, F0 上昇の傾きの平均値:RS, 第 1 句での F0 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:PRS1st，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（第一ホルマント周波数:F1，第二ホルマント周波数:F2, 第三ホルマント周波数:F3, 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音声（Joy）に適用した例を図 6 に示す．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．

3　モデルの評価

3.1 モデルの検証

多層モデルの検証を行うために, Bottom-up 的にモデルから音響特徴の変形ルールを作成し，これに基づいて音響特徴を変化させた合成音について，(1) 基本的な心理特徴は制御可能か，そして，(2) 基本的な心理特徴の変化が感情音声の知覚を生起させられるか，を調査することにより，構築した多層モデルの検証を行う．

3.2 音声変換

モデルから合成音を作成するためには，音響特徴を表す 16 個のパラメータをモデルから生成したルールにより独立に変形・制御できる手法が必要となる．筆者らは，音声波形を STRAIGHT[5]により分析した後，F0 包絡および時間 - 周波数スペクトルを Temporal Decomposition で分解し，ターゲットとなるスペクトルをさらにガウス関数で分解することにより，16 個のパラメータに分解した[6][7]. 16 個のパラメータはルールにもとづいて制御され，逆過程を通って音声波形として合成される．

3.3 評価実験および結果

評価実験は二段階で行う．第一段階として Neutral 音声の 16 個のパラメータを制御することにより, 17 種類の基本的心理特徴それぞれの印象を生起させることができるかどうかを調査する．個々の基本的心理特徴に合わせて 16 個のパラメータ値を設定し, 合成した波形を聴取者に呈示することにより聴衆実験

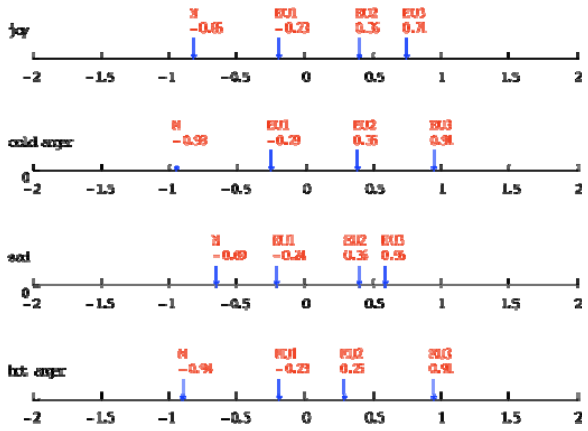


図 7　シェッフェの対比較法による聴取実験結果.

を行った．その結果，聴取者は合成された音声から個々の基本的心理特徴を知覚できることが明らかとなった[3].

第二段階として，17 種類の基本的心理特徴の適切な組み合わせにより，Neutral を除く 4 種類の感情の印象を生起させられるかどうかを調査する．Neutral 音声の 16 個のパラメータを制御することにより，17 種類の基本的心理特徴それぞれの印象の強さを制御し，4 種類の感情音声を合成する．図 7 に聴取実験結果を示す．図では，1 つの Neutral 音声から 4 種類の感情が合成され，その印象が N < EU1 < EU2 < EU3 と強くなっていることがわかる．この実験から，基本的心理特徴の適切な組み合わせにより異なる感情の印象を生起させることができること，また，その強さも制御できることが示された.

4　モデルの応用

4.1 感情音声知覚の共通要素発見への貢献

筆者らは，三層モデルによる感情知覚機構の記述法を応用し，言語が異なる聴取者における感情知覚の共通性の発見を試みている[8].

日本語を理解しない中国語話者に，2-3-1 および 2-3-1 節で説明した実験を課し，日本人聴取者の結果との比較を行った．この結果，基本的心理特徴の約 7 割が一致し，これらの基本的心理特徴に関係する音響特徴パラメータ値もほぼ同じ値を持つことがわかった.

この研究は，「１．まえがき」で触れた Expressive Speech の生成・知覚において言語・民族・文化によらないヒトの生物学的「共通要素」の発見に貢献し，非言語情報によるユニバーサルコミュニケーションの可能性を示すものである.

4.2 感情音声合成への貢献

本稿で示したモデルの評価方法は，そのまま感情音声合成へ応用可能である．本合成手法は，GMM などを用いたマッピング手法ではなく，ルールベースの合成手法なので，基本的に誰の声でも変換可能であり，しかも印象の強さまでも制御可能である.

三層構造モデルは，感情音声に限らず他の表現（たとえば歌声[9]）に対しても適用できる.

4.3 音声中の感情自動認識への貢献

人間の音声は多様な感情を同時に含む事があり，かつ感情の強さも様々である．そして，我々は一つの発話音声から複数の感情をその強さも含めて同時に感じ取ることが出来る．ところが，現在の感情音声認識に関する研究では，音響特徴量と感情カテゴリの間の直接的な関係に着目しており，同時に複数の感情を強さも含めて認識することは困難である．筆者らは，三層構造モデルを感情認識器として構成し，感情認識を試みている[10]．実験は初歩的段階であるが，FIS を用いた三層構造モデルにより，複数の感情の印象の強さを同時に推定可能であることがわかっている．現在，他言語への適用も検討中である.

5　まとめ

筆者らは，言語・民族・文化を越えた音声によるユニバーサルコミュニケーションを指向して，非言語情報の送受について研究を行っている．本スペシャルセッションでは，これらの研究の中で，非言語情報（特に感情音声）に焦点をあてて，その知覚機構のモデル化およびモデルの応用について説明した.

謝辞

本研究は総務省戦略的情報通信研究開発推進制度 SCOPE（071705001）の援助を受けて行われた.

参考文献

[1] Erickson, D. (2005). “Expressive speech: Production, perception and application to speech synthesis,” Acoust. Sci. & Tech., 26, 4, 317-325.

[2] 赤木正人(2005). “表現豊かな音声　―その生成・知覚と音声合成への応用―”, 日本音響学会誌, 61, 6, 346-351.

[3] Huang, C-F. and Akagi, M. (2008) "A three-layered model for expressive speech perception," Speech Communication 50, 810-828.

[4] 上田和夫(1988). “音色の表現語に階層構造は存在するか”, 日本音響学会誌, 44, 2, 102-107.

[5] Kawahara, H., et al. (1999). “Restructuring Speech Representations Using a Pitch Adaptive Time-Frequency Smoothing and an In-stantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds,” Speech Communication, 27, 187-207.

[6] Nguyen, B. P. and Akagi, M. (2009) "A flexible spectral modification method based on temporal decomposition and Gaussian mixture model," Acoust. Sci. & Tech., 30, 3, 170-179.

[7] Nguyen, B. P., Shibata, T., and Akagi, M. (2008). "High-quality analysis/synthesis method based on Temporal decomposition for speech modification," Proc. InterSpeech2008, Brisbane, 662-665.

[8] Huang, C. F., Erickson, D., and Akagi, M. (2008). "Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners," Acoustics2008, Paris, 2317-2322.

[9] 齋藤，辻，鵜木，赤木(2008). ”歌声らしさの知覚モデルに基づいた歌声特有の音響特徴量の分析”, 日本音響学会誌, 64, 5, 267-277.

[10] 青木，黄，赤木(2009). ”音声からの感情認識による感情知覚多層モデルの評価”, 日本音響学会平成 21 年春季研究発表会, 2-P-18.