5-2　マルチモーダルコミュニケーション

5-2　Spoken Communication in Multiple Modalities

Eric Vatikiotis-Bateson（株式会社 国際電気通信基礎技術研究所）

要旨

サイエンスとテクノロジーは関心領域が異なるが、それぞれの目標を達成するための道筋をわざわざ分ける必要はない。ATRコミュニケーションダイナミクスプロジェクトでは、CRLと協力して、自然なコミュニケーション（例えば、表情を伴った発話、情動、身振りなど）がどのように生成され、またどのように知覚、処理されているのかを解明する研究を進めている。このようなコミュニケーションは、聴覚や視覚に限らず様々なモダリティが絡み合っており、また複雑な環境下で起こりうる。本プロジェクトの主たる目標は、マルチモーダルでかつ自然な発話行為の解析と生成である。さらに、情動や身振りといった発話を伴わないコミュニケーションの表現や、複雑な環境下での3次元物体の探索と記述など視覚処理までをも研究対象とする。本研究によりコミュニケーションの計算モデルを確立することを通じて、人間と機械のより良いコミュニケーションの発展に役立てる。

Basic science and technology tend to have distinct interests, however they need not follow separate paths in achieving their goals. The research program of the ATR-I Communication Dynamics Project attempts to serve these multiple purposes. The aim of the project is to understand how naturally occurring communicative events such as expressive speech, emotion, and gestures are produced, perceived, and processed neurally. Communication occurs in multiple modalities, principally the auditory and visual modalities, and in complex environments. Therefore, a major challenge is the analysis and synthesis of multi-modal speech behavior as it occurs naturally - integrated with emotional expression in multi-talker environments. Non-speech expressions of emotion and other gestural forms of communication are also being examined, as are basic visual processes such as the detection and representation of three-dimensional objects in interactive and changing environments. As computational models of communicative phenomena emerge, they will be applied to developing communicatively plausible human-machine systems.

[キーワード]
視聴覚発話信号処理，マルチモーダル・コミュニケーション，トーキングヘッド・アニメーション，マルチモーダルな知覚と生成，コミュニケーションの身振りと表現
auditory-visual speech processing, multi-modal communication, talking head animation, multi-modal perception and production, communicative gesture and expression

1　ＡＶ(音響・視覚)発話処理

従来、コミュニケーションの生成と知覚は別々に研究されてきた。しかし、両者は密接に関連し合っているため、両者をまとめて研究することはむしろ必然である。我々は、コミュニケーションの生成と知覚の関係を知るために、次の三つの切り口から研究を進めている。①実験により、マルチモーダルなコミュニケーションがどのように生成及び構造化されるのかについての基礎的な理解を得る。②パラメータを自由に変化させることができるトーキングヘッド・アニメーションシステムを構築する。このシステムでは実際の計測データやその解析結果に基づいてアニメーションを生成することも可能である。③このトーキングヘッド・アニメー

ションシステムの知覚的なリアリズムを評価するとともに、生成結果の妥当性を検証する。図1に、この研究パラダイムの概要を示す。

2　マルチモーダル発話の計測と解析

時間的に変化する発話現象を、発話音響信号、声道、顔、頭部の運動や筋肉のEMG（筋電信号）といった様々な種類のデータで計測し、解析する。しかし、実験環境そのものにより、測定データが大きな影響を受けてしまうことがある。そのため、二人以上の人間のコミュニケーションを、より自然な状況で記録する新しい計測方法を開発した。これにより、コミュニケーションにおいて最も重要な顔と頭部の運動を、ビデオ映像の画像解析によって非侵襲で高精度に測定できるようになった[Kroos et al., in press]。また、時間的に変化する現象を同定し特性を解析するため、性質の異なる、あらゆる角度からの測定データを線形及び非線形な方法を用いて解析する手法も開発した。これにより、声道のデータや筋肉の運動データから顔の動きと発話音響信号の両方を高い信頼度で推定できるようになった。この推定結果を用いて、筋肉の制御系や物理的な構造特性を記述できるマルチモーダル発話生成に関する計算モデルを構築した[Vatikiotis-Bateson et al., 2000a]。また、顔と頭部の運動を音響信号のスペクトルから復元することに成功した[Yehia et al., 1998; 1999; in press]。これは、発話現象が人間の中では視覚と聴覚で冗長に処理され、両者

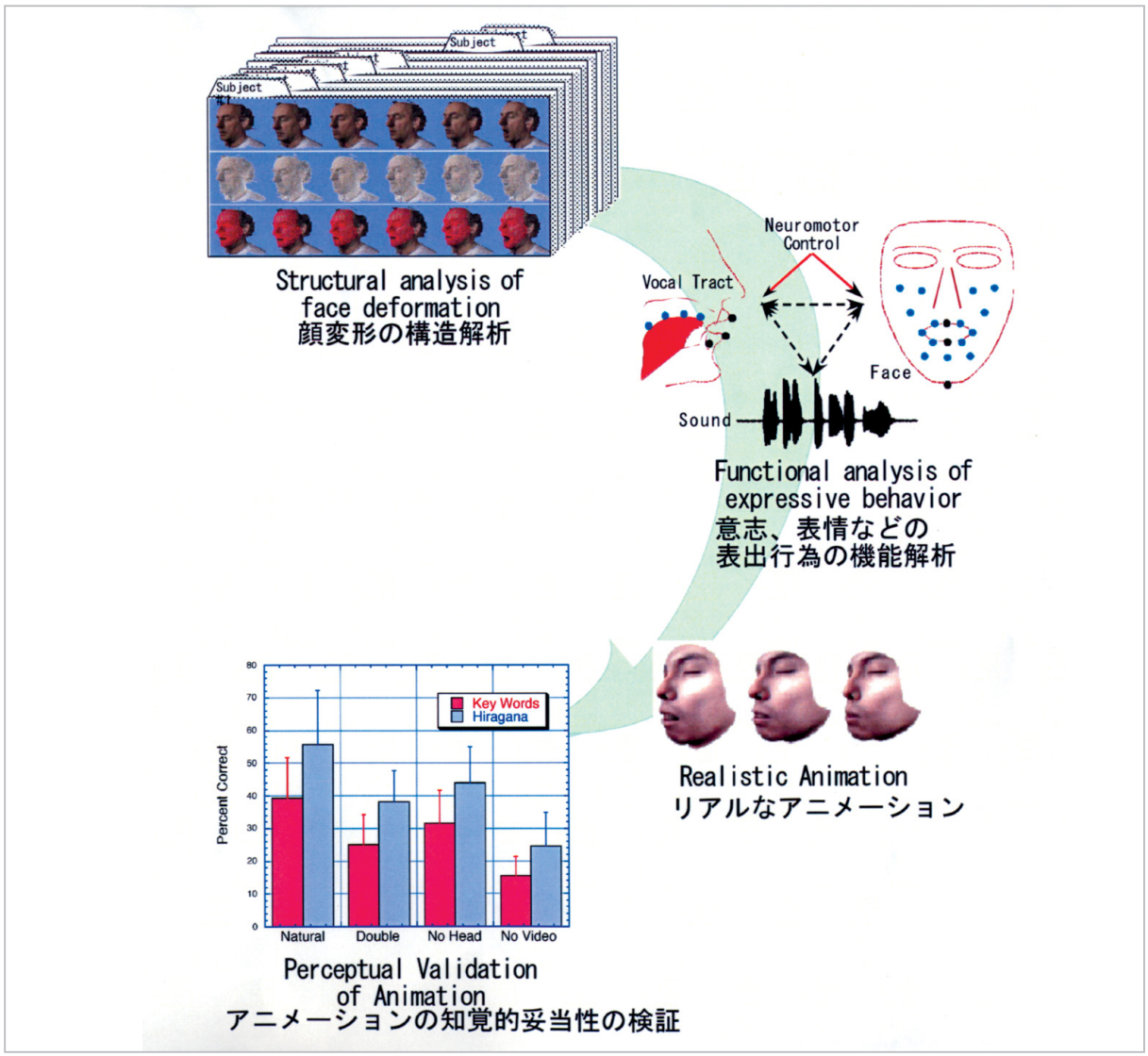

図1

が発話生成に対して情報を共有していることを示唆している。これらの結果は、話者の顔が見えれば話の内容が理解しやすくなるという、昔から知られている精神物理学の知見とよく合致している[e.g., Sumby & Pollack, 1954]。今後、この解析手法を発話現象と非発話現象（情動の表出や談話マーカー、Turn-Taking）との関係の解明にも利用していく。

3　顔と頭部の構造解析

時間的に変化する行動データとともに大規模な構造データベースを構築中である。これは、発話中や表情を表出している時の様々な顔と頭部の形状を円筒状に3次元計測し記録したものである。現在構築済みの300人分のデータベースを用いて、顔のデータを解析し、トーキングヘッド・アニメーションシステムの変形パラメータを得ている。また、顔の構造（基礎的な3次元形態学）と機能（発話、表情等）を比較するための多次元尺度を計算している。このデータベースは、静的な構造と時間的に変化する機能との関係を知るためにも利用できる。例えば、ある被験者のトーキングヘッド・アニメーションと、その被験者とは異なる顔の構造を持つ被験者のアニメーションを比較、あるいは、異なる被験者の構造や機能を別の被験者のアニメーションに合成し、精神物理的な実験によって比較することもできる。さらに、構築したデータベースには、発話現象だけでなく、非発話現象のデータが含まれているので、トーキングヘッド・アニメーションの変形パラメータを使って発話以外の表情（情動など）を付加することも容易である。

4　トーキングヘッド・アニメーションシステム

マルチモーダルな行動を生成するに当たり、二つの方法でトーキングヘッド・アニメーションシステムを開発した。

一つは、運動学に基づくシステムで、被験者の3次元データをもとに顔の運動や顔の変形パラメータを制御するものである。顔の運動データは直接測定するとともに筋電位（EMG）や音響信号から間接的に推定したものを用いる。いったんパラメータ化してしまえば、この方法を用いて、顔のアニメーションを十分な時間・空間分解能で、しかもほぼリアルタイムに生成することが可能になる[Kuratate et al., 1998, submitted]。

もう一方は、カナダの研究グループと共同で開発したもの[Lucero & Munhall, 1999]で、筋肉・骨格構造や皮膚の多層モデル（質点・バネモデル）といった物理モデルに基づくものである。そこでは、Waters[1987]とTerzopoulos[1990]らによって開発された、筋肉の動きを再現する顔のアニメーションモデルが採用されているが、ATRでは更に七つの顔の筋肉のEMG信号から顔のアニメーションを生成できるように拡張した。このような物理モデルに基づく方法は、計算時間を要するため時間分解能を上げることは難しいが、皮膚の変形（特に口の周り）を忠実に描写することが可能という特長がある。

前者の運動学に基づくアニメーションシステムでは、現在、アフィン・メッシュ皮膚モデルと呼ばれる方法[Vatikiotis-Bateson et al., 2000b]を採用しているが、物理皮膚モデルを利用できるように改良中である。

5　マルチモーダル発話の知覚的評価

トーキングヘッド・アニメーションを使ってコミュニケーションの知覚と生成の間の関係を調べるためには、知覚的な見地及び運動学的な見地からこのトーキングヘッド・アニメーション自身の妥当性を評価する必要がある。つまり、トーキングヘッド・アニメーションは実際の顔や頭と同じ運動を表現でき、かつ同じ言語学的情報を伝達することを示さなくてはならない。先に述べたアニメーションシステムは両者ともフィードフォワード型であるため、精度の検証が特に重要になってくる[Vatikiotis-Bateson et al., 2000b]。

知覚的な評価においては、知覚者が、どのような場合にトーキングヘッド・アニメーションが本物でないという疑念を払拭するのか、そしてこれが知覚に及ぼす影響について明らかにした。一つの例として、顔のビデオ映像におけるリアリティの不完全さは、アニメーション化された線画や動物のマンガにおける不完全さより

も大きな影響を与えることが明らかになった[Kuratate et al., submitted]。

さらに、日本語あるいは英語を母国語とする被験者に対して、様々な聴覚的、視覚的条件下でアニメーションを評価する実験を行った。トーキングヘッド・アニメーションと、顔の光点表示、発話中の自然な顔のビデオ映像を比較した。頭部の運動の有無といった異なる条件下で、制御パラメータをシステマティックに変化させて実験を行った結果、トーキングヘッド・アニメーションでは、目、歯、舌などが提示されないにもかかわらず、ビデオ映像を提示したのと同じ応答が被験者から得られることが分かった。さらに、制御パラメータの変化に対する応答も妥当であることが明らかになった。

また、このような人工的に生成された刺激を使用した知覚実験から、頭部の運動と発話の音響信号との間には相関があり[Yehia et al., in press]、ノイズによって音響信号がマスクされた場合でも、トーキングヘッド・アニメーションは言語的な韻律という重要な情報を含んでいることも明らかになった[Munhall et al., in preparation]。

6 さらなる展開

これまで述べてきたように、発話という行動を、解析、生成、評価するための三つの研究手法を確立した。今後は次のように研究を展開していく予定である。①発話を伴わない身振りや情動を織り込む。②マルチモーダルな刺激に対する脳活動を計測し、精神物理における知覚研究と比較する。③複雑な環境下でも我々の研究手法を適用できるか否かを検証する。もちろん、これらは多くの可能性のほんの一例に過ぎないのは言うまでもない。

7 非発話表現

従来は、動きのない静止した顔に注目した研究が圧倒的に多かった。しかし、意識的に表出した表情であれ[e.g., Ekman, 1972]、既知の顔の識別であれ[Bruce & Young, 1986]、動きといった時間変化する情報が重要であることは明らかである。例えば、両極端な表情（例えば、中立と怒り）の間を異なるスピードで変化させたアニメーションの知覚応答から、それぞれの表情には、固有の運動スピードがあることが分かった。例えば、「微笑み」は「しかめ面」よりもスピードが速く、また、適切なスピードで提示することで、より確実に知覚されるのである[Kamachi et al., 2001]。このような現象や、発話とともに起こり得る自然な表情について詳細に解析・検証していく。

8 マルチモーダル刺激に対する脳の反応

機能に関する研究（fMRI, MEG）の問題点は、脳活動の反応が従来の行動学的な研究結果と必ずしも一致しないことである。

例えば、幼児においては、母国語でない音を分類する能力と、その違いを脳の聴覚領域に記憶する能力の間には、発達速度に差がある。特に、生後一年間は認知過程と神経過程の発達速度は異なり、生後一年後にやっと音声そのものの識別と音声カテゴリの識別が一致するようになる。

大人にはこのような違いはないと思われるものの、マルチモーダルなコミュニケーションでは、運動、聴覚、視覚、そして触覚といった様々な処理が同時に行われている。各モダリティに冗長性があることを鑑みると、AV発話は一つあるいは複数のモダリティ情報の欠落に対して、認知過程はあまり影響を受けないと考えられる。我々は、認知学的な研究と脳機能研究を比較することにより、認知過程と神経過程の間の冗長性を検証し、知覚される情報が急激に変化したときに素早く適応するためのメカニズムを解明する作業を行っている。

9 複雑な環境下での適応とナビゲーション

人が、他者やまわりの環境とインタラクションを持とうとする場合、外界の情報を獲得する条件は必ずしも理想的ではない。情報や文脈が全く未知のものであったり、激しく劣化していたりすることもしばしばである。したがって、情報を獲得するには柔軟で適応的である必要が

あり、また、冗長性も兼ね備えていないといけない。AV発話過程の研究では、複数のモダリティ間での情報の冗長性を解明することに焦点を当ててきた。しかし、我々人間が、一般的な環境下でいかに情報を獲得し、認識するのかというのは、より基本的なテーマである。人間は、今得られている情報が対象を認識するのに十分でないなら、視点を移動させることで新たな情報を獲得しようとする。つまり、3次元の対象物体を「能動的」に見ているのである。そこで、このような3次元環境の情報を能動的に獲得するシステムを構築した。復元した3次元情報を評価し、最も精度良く観測できる位置・姿勢にカメラを移動させ、撮影した画像を使って、対象物体の3次元構造の精度を向上させることを繰り返す。このような処理を行うだけで情報の精度が約4倍になることを確認した。

10　まとめ

マルチモーダルコミュニケーションの研究における問題点やアプリケーションを限られた紙面ですべて語り尽くすのは困難である。特に、コンピュータや開発ツールは日進月歩である。例えば、現在我々はテレビ電話のように低ビットレートで情報をやり取りしているが、将来は映像と音声だけでなく、インタラクティブなコミュニケーションロボットをコミュニケーションメディアとして用いることができるようになるであろう。そのようなシステムがマルチモーダルコミュニケーションの神経学的、認知的な理解を深めることに役立つことを、我々は心底望んでいる。

参考文献

1 Bruce, V. & Young, A.W. (1986), "Understanding face recognition", British Journal of Psychology, 77, 305-327.
2 Ekman, P., Friesen, W. V., & Ellsworth, P. (1972), "Emotion in the human face: Guidelines for research and a review of findings", New York: Pergamon Press.
3 Kamachi, M., Bruce, V., Mukaida, S., Gyoba, J., Yoshikawa, S., & S.Akamatsu (2001), "Dynamic properties influence the perception of facial expression", Perception, 30, 875-887.
4 Kroos, C., Kuratate, T., & Vatikiotis-Bateson, E. (in press), "Video-based face motion measurement", Journal of Phonetics.
5 Kuratate, T., Vatikiotis-Bateson, E., & Yehia, H. C, (under revision), "Talking faces synthesized by facial motion mapping", Speech Communication.
6 Kuratate, T., Yehia, H., & Vatikiotis-Bateson, E. (1998), "Kinematics-based synthesis of realistic talking faces. In D. Burnham, J. Robert-Ribes, & E. Vatikiotis-Bateson (Ed.)", International Conference on Auditory-Visual Speech Processing (AVSP'98), (pp.185-190). Terrigal-Sydney, Australia: Causal Productions.
7 Lucero, J. C., & Munhall, K. G. (1999), "A model of facial biomechanics for speech production", Journal of the Acoustical Society of America, 106, 2834-2842.
8 Sumby, W. H., & Pollack, I. (1954), "Visual contribution to speech intelligibility in noise", Journal of the Acoustical Society of America, 26, 212-215.72.
9 Terzopoulos, D., & Waters, K. (1990), "Physically-based facial modeling, analysis, and animation", Visualization and Computer Animation, 1, 73-80.
10 Vatikiotis-Bateson, E., Kuratate, T., Munhall, K. G., & Yehia, H. C. (2000a), "The production and perception of a realistic talking face. In O. Fujimura, B. D. Joseph, & B. Palek (Eds.)", Proceedings of LP'98, Item order in language and speech 2 (pp.439-460). Prague: Charles University (Karolinum Press).
11 Vatikiotis-Bateson, E., Kroos, C., Kuratate, T., Munhall, K. G., & Pitermann, M. (2000b), "Task constraints on robot realism: The case of talking heads. In K. Kamejima (Ed.)", 9th IEEE International Workshop on

Robot and Human Interactive Communication (RO-MAN 2000), (pp.352-357). Osaka, Japan: IEEE.

12 Waters, K. (1987), "A muscle model for animating three-dimensional facial expression", Computer Graphics, 22, 17-24.

13 Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (1999), "Using speech acoustics to drive facial motion. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Ed.)", Proceedings of the 14th International Congress of Phonetic Sciences, 1 (pp.631-634). San Francisco, CA: Linguistics Dept., UC Berkeley.

14 Yehia, H. C., Rubin, P. E., & Vatikiotis-Bateson, E. (1998), "Quantitative association of vocal-tract and facial behavior", Speech Communication, 26, 23-44.

15 Yehia, H., Kuratate, T., & Vatikiotis-Bateson, E. (in press), "Linking facial animation, head motion, and speech acoustics", Journal of Phonetics.

Eric Vatikiotis-Bateson, Ph.D.
㈱国際電気通信基礎技術研究所(ATR)
先端情報科学研究部コミュニケーションダイナミクスプロジェクトリーダー