グラフィクスとCAD 87－6
（1997. 10. 17）

三次元運動学に基づく顔面アニメーション

倉立 尚明, Frédérique Garcia, Hani Yehia, Eric, Vatikiotis-Bateson
{tkurata, garcia, yehia, bateson}@hip.atr.co.jp

(株)ATR 人間情報通信研究所

〒619-02 京都府相楽郡精華町光台 2-2

発話の際に話者から与えられる画像情報としては、唇の動きだけではなく、頬や顎も含めた顔面領域の動きも重要な働きを担っており、特に声道と顔面領域の動きに高い相関があることがこれまでに調べられている。本稿では、レーザースキャナにより得られた代表的な母音の連続発話時の形状を含む基本的な顔形状をもとに、この顔面領域の少数の特徴点の三次元運動から発話時に近い顔形状を合成し、音声と同期した顔面アニメーションを生成する手法について報告する。

Facial Animation based on 3D Kinematics

Takaaki Kuratate, Frédérique Garcia, Hani Yehia, Eric Vatikiotis-Bateson
{tkurata, garcia, yehia, bateson}@hip.atr.co.jp

ATR Human Information Processing Research Laboratories

2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

In this paper we describe a method for animating 3D representation of faces during speech from 3D kinematic data of orofacial points. The description includes the basic components of the model and the methodology used for configuring and animating faces. The motivation for this analysis is based on previous results where we have shown that much of the speech acoustics can be reliably estimated from the position of a limited set of points located around the lips and over the face. This fact indicates that these same points contain the information necessary to characterize the critical features of facial motion.

1 はじめに

コンピュータグラフィックスによるリアルな顔画像生成のためには、形状や質感の静的情報とともに、表情や発話による動作や形状の時間変化などの動的情報が重要となる。

これまで我々は音声研究の分野において、発話時の音響情報が顔面の少数の特徴点の座標の時間変化から精度良く見積ることができることを示している。[1][2] このことから、我々はこの顔面の同じ特徴点が発話時の表情生成に大きく起因しているものと考えている。

しかし、これまでの顔面部分のアニメーションに関しては、感情表現などの表情合成の分野においては顔面領域のモデル化が行われているが、発話同期のものでは唇部分が重要視されている。表情筋と皮膚をモデル化したリアルな発話表情合成モデルも報告されているが [3][5]、非常に複雑な問題となり、発話同期表情生成などのアプリケーションを考えると、できるだけ簡易にかつ制御しやすい形態で顔面を含めた発話表情表現を行なうことが望まれる。

そこで我々は、発話時に表面的に観測される顔面の特徴点の三次元運動学的データを元に三次元顔面アニメーションを生成するシステムを構築している。これら特徴点の動きは声道パラメータとの相関が高く、自然音声のみならず合成音との同期アニメーション生成が期待できるものである。

2 顔面アニメーションシステム

我々が検討を行っている顔面アニメーションシステムでは主に以下の点を目標としている。

- リアルな顔面アニメーション生成
- 自然音声および合成音声との完全同期
- 話者固有の自然な顔面運動の分析および合成
- 顔面三次元運動の少数パラメータでの表現

このため、図1に示すような三次元顔面アニメーションシステムを構築中である。

現在のところ、発話中の顔形状をリアルタイムでアニメーションに必要なレートおよび空間解像度で同時に取り込む事が可能な装置が存在しない。そこで、モーションキャプチャによる顔面アニメーションシステムを基本とし、発話時の顔の動きに関しては顔面上の数点に配置したマーカーをリアルタイムでトラッキングを行うデバイスを用いて計測し、このデータを解析した後、その結果を用いて予め用意された基本的な顔形状を変形することによりアニメーションを生成することとした。

以下にこのシステムの詳細について説明する。

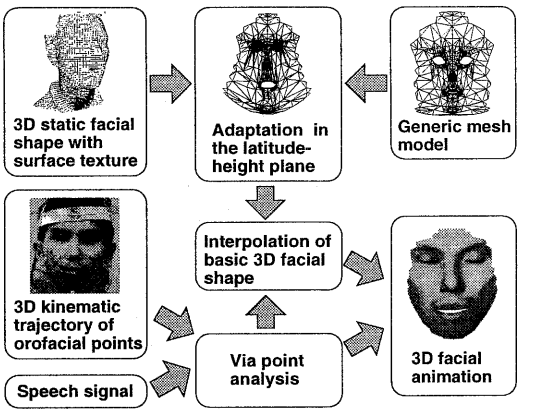


Figure 1: 三次元アニメーションシステムの構成図

3 話者形状データの取得

話者形状データに関しては、代表的な発音時等における静的な状態での顔全体の基本形状と、実際の発話時の顔面上の特定部位の三次元座標時間変化、すなわち運動学的データを計測した。アニメーション生成には、この発話時の運動学的データを再現するために、各フレームにおける運動学的データを表現できる最適な顔形状を得るような基本顔形状の重ね合わせを求め、必要に応じてさらにその形状の部分変形を行う。以下にこれら基本顔形状データと運動学的データの取得方法について解説する。

3.1 基本顔形状データの取得

特定話者の顔面アニメーションを生成するにあたり、まずその代表的な発音時等における基本形状を計測した。計測には Cyberware のレーザースキャナを用い、母音 /a/, /i/, /u/, /e/, /o/ を連続発声時の形状と、口腔を意識的に開いた場合と閉じた場合および自然な状態 (open, close, neutral) の合計 8 つの三次元形状を計測した。このスキャナは鉛直回転軸まわりに 360 度回転し、各回転位置において軸方向に形状と表面テクスチャを走査するもので、解像度は回転方向・縦方向ともに 512 分割で計測を行った。この Cyberware レーザースキャナの特徴としては円柱座標系 (r, θ, z) において、走査方向の θ および z に対して r が一意に決定されてしまうという点である。このためスキャンの際の r 方向でのオクルージョン (例えば顎を引いたときに顎の下側が隠れるような場合) には注意が必要となる。

図 2に母音 /a/ の計測データに関して単純に三角形パッチを施した例を示す。この例ではメッシュ構造が見易いようにオリジナルデータの半分の密度でパッチ

Figure 2: 基本顔形状データ例 (母音 /a/)

Figure 3: 計測テクスチャ (母音 /a/)

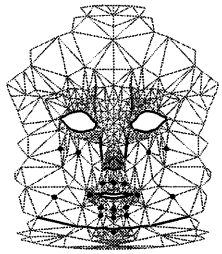


Figure 5: ジェネリックメッシュ例

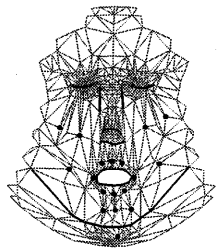


Figure 6: (θ, z) 平面での適合結果

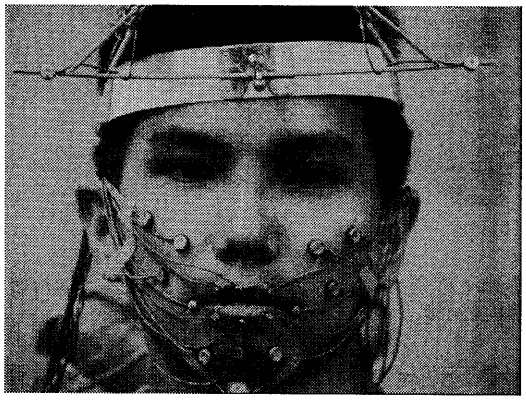


Figure 4: 赤外線ダイオードの配置例

を生成してある。このオリジナルデータの場合、顔面アニメーションに利用できるような有効領域は図3に示すような 310x274 の領域となる。

3.2 発話時の運動学的データの取得

発話時の特定部位の時間変化を正確に取得するため、高速かつ高精度でのトラッキングが可能な OPTOTRAK (Northern Digital, Inc.) を用いた。これは赤外線ダイオードを用いたマーカーを CCD により計測するもので、複数のマーカーに関してビデオレート以上でのサンプリングが可能である。

計測には唇のまわり、頬・顎を主としてマーカーを配置する。この配置例を図4に示す。この例では顔面に 18 個のマーカーを使用し 60Hz でサンプリングを行った。また頭部全体の動きを得るためにこれら 18 個とは別に 5 個のマーカーを頭部に装着している。

実験ではこれらマーカーの軌道以外にも、音声信号 (10kHz サンプリング) と、ビデオにより正面顔画像を同時に記録している。

4 形状データの一般化適合

先に得られた基本顔形状の重ね合わせを行うことにより、発話時の運動学的データを再現する任意の顔形状にほぼ近い顔形状の合成が可能となる。しかし、個々の計測データ間のずれやメッシュ構造が顔の特徴に対して無関係に配置されているため、計測された基本形状データそのものを重ね合わせ、またアニメーションのために制御を行うには困難であり、何らかの構造化が必要となる。そこで、我々はジェネリックメッシュを用い [3]、同じ構造のメッシュを個々の計測された基本顔形状に適合することにより一般化を行なった。

このようなメッシュの適合には多くの方式が提案されているが [6][7]、我々は今回利用する全ての基本顔形状データが円柱座標系 (r, θ, z) において (θ, z) に対して r が一意に決定されるという性質を利用し (θ, z) 空間で特徴部分に基づいた適合を行うこととした。

具体的には図5に示すような、変形を考慮して目・口などにノードを集中させた基本的なメッシュ構造を作り、この目・鼻・口・顎などを特徴部分とする。そして (θ, z) 平面上において、予めジェネリックメッシュ上で定義されたこれらの部分と、基本顔形状データ上でこれらに対応する特徴部分を指定することにより対応関係を明らかにし、この対応関係を元に特徴部分以外のノード点を (θ, z) 平面における顔基本形状データ上の適当な場所に配置し、最後に個々のノードに関し、(θ, z) から一意に決定される r 値を用いて形状への適合を行う。

この特徴部分以外のノード点の配置には、今回は特徴ベースのモーフィングを利用した [8]。モーフィングは本来画像変形のために用いられる手法で、画像中に変形の基準となる特徴ベクトルを指定し、他の画像においてその特徴ベクトルに対応するベクトルを決定することにより、画像間の対応づけを行い、その特徴ベクトルからの相対位置に基づいて画像中の全ての画素の移動や画素値の補間を行うことにより、滑らかな画像変形を行うものである。この特徴ベクトルに基づく画素の移動を利用して、ここでは単純に特徴部分以外

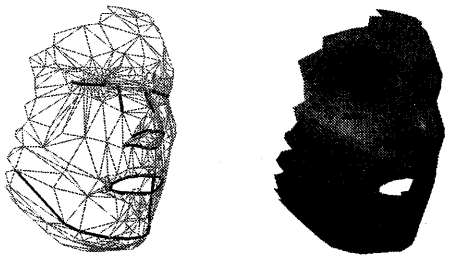


Figure 7: 三次元空間での適合結果

のノード点位置の決定を行った。

図5に今回用いた $(\theta,\ z)$ 平面上におけるジェネリックメッシュを、図6に先の図2,3における母音 /a/ の形状の $(\theta,\ z)$ 平面上における適合結果を示す。図中で太い線で示される目・鼻・口・顎の輪郭線と、黒丸で示すノードを特徴部分として変形を行った。この黒丸は運動学的データを取得したマーカーに対応している。今回用いたジェネリックメッシュはノード数544、942ポリゴンにより構成され、この内約120ノードがこれら特徴部分として定義されている。基本顔形状におけるこれら特徴部分の指定は現在は OpenInventor を利用した環境で手作業で行っている。この適合結果をもとに、$(\theta,\ z)$ から一意に決定される r 値により三次化した結果を図7に示す。特徴部分の指定や適合で得られる $(\theta,\ z)$ は、離散的に与えられる基本形状データ中の近傍の計測点より内挿により求めている。また今回は顔面しか考慮していないため口腔内はモデル化されていない。

以上の一般化適合により全ての基本顔形状データが同じメッシュ構造を持つこととなり、異なる基本顔形状間での内挿や重ね合わせが容易に行える。

5 運動学的データと基本顔形状データの対応づけ

今回のアニメーションの生成は図8に示すような複数の基本顔形状データの重ね合わせと、部分変形により運動学的データの各フレームにおけるより近い形状を獲得することにある。

しかし、マーカーから得られた運動学的データと、静的データとして取得した個々の基本顔形状データとでは観測した座標系が異なるため、これらの座標系間の整合を取る必要がある。　今回はマーカーと対応する点を基本顔形状データ上で手作業により指示し、これらの点を参照点として個々の基本顔形状データ間の整合性を得る。そして、運動学的データの平均的な座

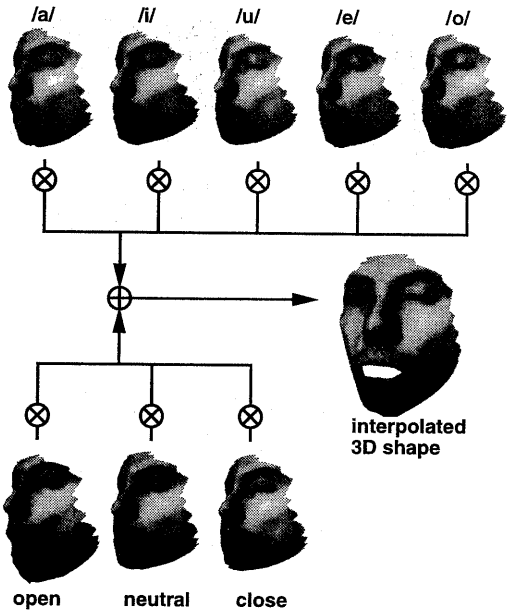


Figure 8: 基本顔形状データの合成による任意時刻の顔形状を近似

標値に対してこれら基本顔形状データから運動学的データへの座標変換を求めた。

この座標変換では、回転および平行移動の6自由度をもった座標変換によりそれぞれの座標系間が関係づけられるものとし、対応する参照点との距離の差の合計を評価関数とする単純な最小値問題としてこれらの座標変換を求めた。個々の基本顔形状データにおいて、発話形状の違いによる参照点の位置のづれが生じるが今回は顎先のように極端に大きなずれが生じる部分は適宜除外して座標変換を求めた。

これらの座標変換によりほぼ整合性が取られた基本顔形状データの集合から、それらの成分の合成により各時間フレームにおける運動学的データを表現するべく、各基本顔形状の合成成分を求める。基本顔形状 j での i 番目のマーカーに対応する座標ベクトルを $\vec{v}_{s(i,j)}$、基本顔形状 j の合成係数を k_j とすると、基本顔形状より合成される i 番目のマーカに対応する座標ベクトル $\vec{v}_{d(i)}$ は次式で表すことができる。

$$\vec{v}_{d(i)} = \sum_j k_j \vec{v}_{s(i,j)}$$

これと実際の運動学的データとして得られる i 番目のマーカの座標ベクトル $\vec{v}_{d(i)}$ との距離の差の総和 $\sum_i |\vec{v}_{d(i)} - \vec{v}_d b(i)|$ を評価関数として最小値問題として解を求めた。

6 アニメーション生成例

今回は日本人男性被験者に対して、「桃太郎」で良く使われる文章から、通常の発話で5秒程度のものを5種類各々4回づつの発話を記録した。

この中から、「おばあさんは川へ洗濯にでかけました」という文章に対する入力音声信号を図9に、そしてこの時の運動学的データをもとに前述の方法により求めた各基本形状の合成比率を図10に示す。この例では自然な状態の基本顔形状を主として、他の母音等の形状成分を加算(時には削減)することにより入力データに近い出力を得る結果が得られた。

この例におけるアニメーションの生成から、時刻 $t = 0.00, 1.17, 2.43, 3.20(s)$ における画像と、原画像との比較を図11に示す。これよりおおむね原画像の形状を再現できていることを確認した。

しかし、合成比率の際の計算精度の影響により場合によっては予想通りの結果が得られないフレームもある。このような場合の例として、本来ならば口が完全に閉じているべき状態でアニメーションではそれが実現できていなかったりする場合が視覚的に特に目立つものとしてあげられる。

また基本顔形状側でのマーカー対応点の指定誤差により、全般的に上唇の領域での合成の誤差が大きく、動きの表現上不完全のものとなっている。

さらに運動学的データは via point analysis [9] により5次のスプライン補間により表現することができ、より少ないパラメータで同等のアニメーション生成が実現できた。

7 まとめと今後の課題

顔面の唇・頬・顎などの特徴点の運動学的データをもととして発話同期三次元顔面アニメーション生成基礎システムを構築し、記録音声と同期したアニメーション生成を行った。

生成アニメーションには部分的に視覚的に違和感の生じる部分が残されており、これらは基本顔形状上における特徴点指定の誤差および、これらの誤差から生じる合成係数の算出精度によるものと考えられる。

今後の課題として、これら問題点の解決と運動学的データの via point analysis によるデータ圧縮率と生成アニメーションの評価などがあげられる。

さらにシステムの改良として、唇や口腔内の3Dモデル化などにより形状表現の向上などが考えられる。特に、データ計測に関しては赤外線ダイオードによるトラッキングではなく、Optical Flow などを用いた非接触による顔面運動の検出や、Active Shape Modeling (ASM) による唇の内側および外側輪郭の抽出とトラッキング等も検討している。この ASM に関しては、今回と同等の発話条件において顔面部のマーカー除去して記録した発話画像では、唇の内側および外側輪郭の追跡が比較的精度良く行えるので、特に唇領域に関してはさらに形状を正確に記述できることが期待できる。

このように、将来的には基本顔形状の入力を含め、カメラ画像ベースでの統合処理化を行ってゆくつもりである。また現在は実験中記録した自然音声との同期再生のみを行っているが、さらに声道と顔面運動の相関を利用して、合成音声からの画像の同期生成も行う予定である。

References

[1] E. Vatikiotis-Bateson and H. Yehia, *"Physiological modeling of facial motion during speech"*, Trans. Tech. Com. Psycho. Physio. Acoust., Vol.H-96-65, pp.1-8, 1996

[2] E. Vatikiotis-Bateson, I.-M. Eigsti, S. Yano and K.Munhall, *"Eye movement of perceivers during audiovisual speech perception"*, Perception & Psychophysics, in press.

[3] Y. Lee, D. Terzopoulos, K. Waters, *"Realistic modeling for facial animation"*, Computer Graphics, Vol.29, No.4, pp.55-62, 1995.

[4] K. Waters, *" A muscle model for animating three-dimensional facial expression"*, Computer Graphics, Vol.22, pp.17-24, 1987.

[5] 森島 繁生, 世良 元, D. Terzopoulos, "物理法則に基づいた筋肉モデルによる口唇形状の制御", 第12回 NICOGRAPH 論文コンテスト論文集 pp.219-229, 1996.

[6] D. Terzopoulos and Manuela Vasilescu *"Sampling and Reconstruction with Adaptive Meshes"*, CVPR91, pp.70-75, 1991

[7] F.I.Parke and K. Waters *"Computer facial animation"*, A K Peters, Ltd., 1996.

[8] T.Beier, S.Neely, *"Feature-Based Image Metamorphosis"*, Computer Graphics, Vol.26, No.2, pp.35-42, 1992.

[9] Y. Wada and M. Kawato, *"A theory for cursive handwriting based on the minimization principle."*, Biological Cybernetics, Vol.73, pp.3-13, 1995.

[10] K. Waters and D. Terzopoulos, *"Modeling and animating faces using scanned data"*, Visualization and Computer Animation, Vol.2, 1991

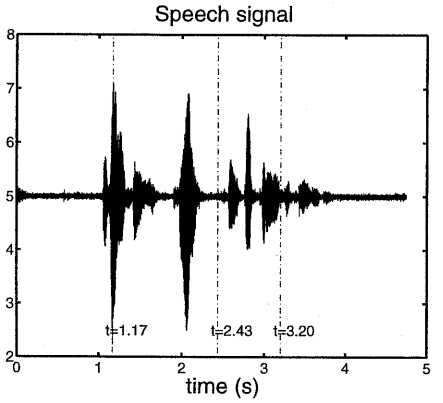


Figure 9: 音声信号例 (日本語「おばあさんは川へ洗濯に出かけました」).

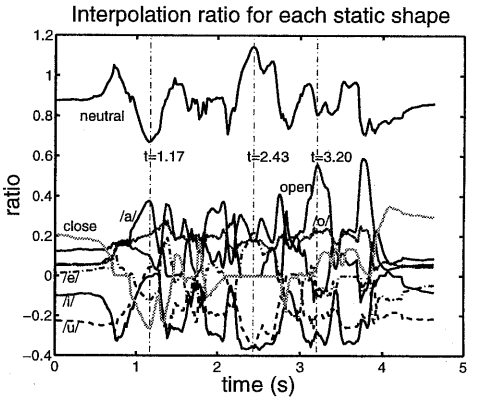
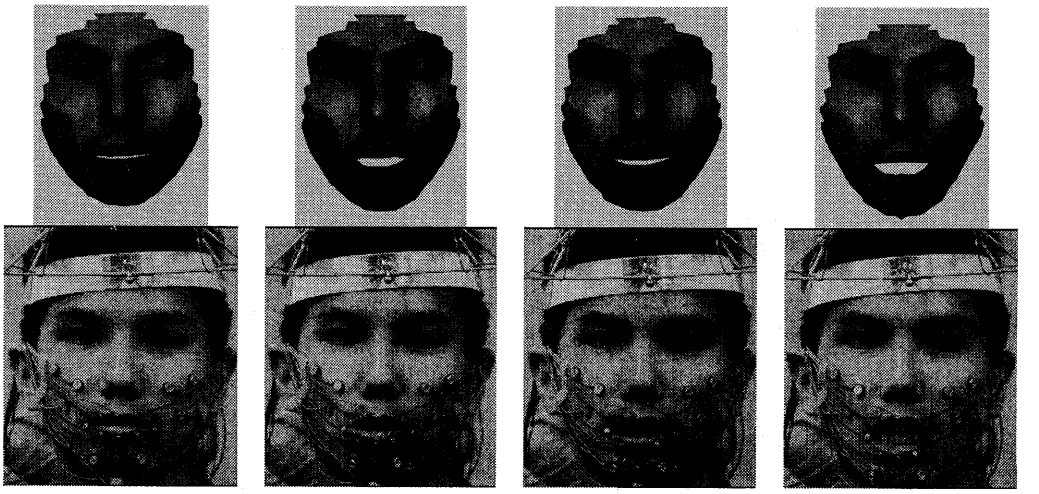


Figure 10: 基本顔形状データ (5 母音他 3 形状) の合成比率の時間変化



t = 0.00　　t = 1.17　　t = 2.43　　t = 3.20

Figure 11: アニメーション生成結果と原画像との比較