ヒューマンインタフェース 80−10
音声言語情報処理 23−10
（1998. 10. 17）

個人性を考慮した顔動画像と音声合成

森山 高明* 蜂谷 雅弘* 小川 均* 天白 成一**
*立命館大学理工学部情報学科 **(株) アルカディア
〒525-8577 滋賀県草津市野路東 1-1-1 〒562-0013 大阪府箕面市坊島 1-3-40

近年, 音声合成システムとコンピュータグラフィックスによる表情や発話動作の顔の動画像を同期したインタフェースの研究が行なわれてきている. このようなシステムでは音声と同期して顔の表情や口の形状が変化する. しかし, 現在開発されている多くのシステムは特定の人物を合成の対象としている. このようなシステムが不特定の人物を対象として個人の顔動画像の合成や音声の合成ができればシステムの応用範囲は広がる. 例えば発話内容によって人物を変更することもできる. また使用者の趣味や趣向をシステムに反映することができ, 結果として自由度の高いインタフェースを提供できる. 本研究では, 不特定人物の個人性を反映するために必要となる基本周波数制御システムを提案し, 不特定人物の個人性を考慮した音声合成と顔動画像の同期システムを構築する.

Facial animation and speech synthesis system for unspecific person

Takaaki MORIYAMA* Masahiro HACHIYA* Hitoshi OGAWA* Seiichi TENPAKU**
*Ritsumeikan University **Arcadia, Inc.
1-1-1, Nojihigashi, Kusatsu, Shiga, 525-77 JAPAN 1-3-40, Bohnoshima, Minoh, Osaka, 562 JAPAN

Recently, speech processing system and facial animation of expression and action to produce a speech sound is studying. But most of these system synthesis specific person's face. To apply these system in many field they must synthesis face expression of unspecific persons. For example, user's taste is reflected in these system. In this research, I propose controlling system of fundamental frequency for unspecific person. And speech synthesis and facial animation system for unspecific person is proposed.p

1 はじめに

人間が他の人間と意思や意図の疎通を行う場合, 様々な伝達手段を用いている. 例えば, 意思の伝達手段として文章表記, 音声, 身振り, 表情変化などが考えられる. 人間はこれらの伝達手段を個々に独立して使用するのではなく, 統合的に利用して自分の意思を他人に伝えていると考えられる.

一方, 人間と計算機の間のコミュニケーションは人間同士の場合とは大きく異なる. 従来, 計算機への入力は固有のデバイスを通じて行われてきたため, 人間の表情を読み取ったり, 身振りなどを理解することは難しい. また, 計算機からの出力は画面への表示や警告音などが一般的である. 人間がストレスを感じることなく計算機を利用するためには, 人間と計算機の間のコミュニケーションを人間同士の意思伝達と同様な手段を用いて行うための技術開発が必要であると考えられる.

そのために, 音声研究の分野では, 計算機に人間の話す音声を合成, 認識させる技術が研究されてきている. また顔画像研究の分野でも, 人間の顔動画像を計算機で合成したり, 顔認識などの研究も盛んである. さらに近年ではこのような音声合成システムとコンピュータグラフィックスによる表情や発話動作の顔の動画像を同期したインタフェースの研究が行なわれてきている [1]. このようなシステムでは音声と同期して顔の表情や口の形状が変化する. これによって人と計算機が音声だけを用いてコミュニケーションを行なうよりも, 円滑なコミュニケーションを図ることができると考えられる.

現在開発されている多くのシステムは特定の人物を合成の対象としている. しかし, このようなシステムが不特定の人物を対象として個人の顔動画像の合成や音声の合成ができればシステムの応用範囲は広がる. 例えば発話内容によって人物を変更することもできる. また使用者の趣味や趣向をシステムに反映することができ, 結果として自由度の高いインタフェースを提供できる.

本稿では, 不特定人物の個人性を反映するために重要な基本周波数制御システムに必要な機能について述べる. さらに, 不特定人物の個人性を考慮した音声合成と顔動画像の同期システムを構築する.

2 発話音声における不特定人物の個人性

人間の発話音声中で話者個人の特徴を表現する情報として以下の 2 つが考えられる.

- 話者の声質に関する情報 (分節的特徴).
- 話者の話し方に関する情報 (韻律的特徴).

話者の声質に関する情報は, 不特定人物の個人性を表現するために必要な情報である. しかし, 単純に 1 単語だけを発話するのではなく, 長い文章や対話などの発話音声を合成する場合, 話者の声質に関する情報以外に, 話し方に関する情報を利用することが重要になる. 話者の話し方に関する情報としては音声の韻律的特徴 (基本周波数, 時間長, パワー) が考えられる. 本稿ではこれらの韻律的特徴のうち特に基本周波数によって表現される話し方を制御することを目的とする. したがって基本周波数の制御を自由に行うことのできる言語情報と非言語情報を利用した基本周波数制御手法を利用する.

3 言語情報と非言語情報を利用した基本周波数生成手法

3.1 言語情報と非言語情報

基本周波数によって表現される情報は言語的な側面から以下の 2 種類に分類できる.

- **言語情報:** 1 モーラ毎の基本周波数の相対的な高さの変化を表す. 東京方言の単語のアクセントなどに代表される情報であり, 発話文によって一意に決定できる.
- **非言語情報:** 言語情報以外の基本周波数変化に関わる情報を表す. 発話の個人性, 話者の感情状態を反映した発話表現, 方言, プロミネンスなどの強調表現などに代表される情報である. 通常の発話では複数の非言語情報が混在し, お互いに影響しあっていると考えられる.

言語情報と非言語情報を用いることで, 発話文における単語のアクセントと, アクセント以外の発話

表現を分けて基本周波数を生成することが可能になる. 非言語情報を変更することで, 話者毎に話し方の異なる基本周波数を生成可能になると考えられる.

3.2 言語情報と非言語情報を利用した基本周波数制御手法

言語情報と非言語情報を利用した基本周波数制御手法での発話の基本周波数生成過程を図1に示す.

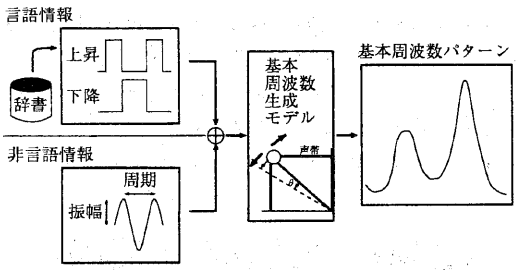

図 1: 基本周波数生成過程

言語情報と非言語情報から基本周波数を生成するために, 基本周波数の上昇と下降を扱うことのできる基本周波数生成モデルを使用する [9].

このモデルは, 喉頭の生理機構のうち, 甲状軟骨, 輪状軟骨, 輪状甲状筋, 声帯に注目し [10], 声帯を収縮させることによって基本周波数を上げる方向に働く力 (τ_1) と, 声帯を伸ばすことによって基本周波数を下げる方向に働く力 (τ_2) から基本周波数の変化を計算する運動力学モデルである.

言語情報と非言語情報を利用して発話の基本周波数を分析するためには, 言語情報, 非言語情報と基本周波数生成モデルへの入力 (τ_1, τ_2) との関係を定義する必要がある.

言語情報は 1 モーラ毎の相対的な音の高さの変化を表している. つまり, あるモーラが「高い (または低い)」音であるということはτ_1(またはτ_2) に一定の大きさの力がそのモーラの時間長だけ代入されることを示す. したがって, 発話文全体では各モーラ毎の τ_1(または τ_2) へ代入された値が時間順に並べられる. これを「言語情報入力パラメータ」と呼ぶ.

非言語情報は言語情報以外の基本周波数変化に関する情報であるから, 発話文の音の高さの変化を表している. したがって, 発話文の音の高さの変化に対応する ($\tau_1 - \tau_2$) の値が発話文の時間長分並んだものが「非言語情報入力パラメータ」である.

3.3 非言語情報の決定方法

言語情報は単語のアクセントであるため発話文から一意に決定することが可能である. 一方, 非言語情報は発話音声の様々な特徴 (感情, 個人性, プロミネンス, 方言) を表現するため, 発話に応じて適切に設定する必要がある.

不特定な人物固有の話し方を非言語情報で表現するためには, 事前に対象人物の発話の基本周波数を解析し, 対象となる人物固有の非言語情報パラメータを求めておく必要がある.

3.4 不特定人物の非言語情報分析手法

不特定人物の非言語情報パラメータを求めるために, 図 1で示した基本周波数生成過程を利用して A-b-S 法による基本周波数分析を行う.

3.4.1 言語情報の決定方法

言語情報は単語のアクセントなので発話文から一意に決定することが可能である. 具体的には次の手順で言語情報を決定する.

1. 発話開始時刻および発話文の 1 モーラ毎の時間長を計測する. 本研究ではサウンドスペクトログラムから目視で 1 モーラ毎の時間長計測を行う.

2. 発話文の各モーラ毎に相対的な音の高さについてラベル付けする. ラベル付けでは, 該当モーラの音の高さが 1 モーラ手前のモーラの音よりも「高い」, 「低い」のどちらかで表される. ただし, 第 1 モーラは次モーラとの相対的な音の高さを比較する.

3.4.2 非言語情報の決定方法

一方, 非言語情報は発話音声の様々な特徴 (感情, 個人性, プロミネンス, 方言) を表現するため, 発話に応じて適切に設定する必要がある. しかし, 一般的な発話では複数の非言語情報が相互に影響しあっていると考えられるため, 非言語情報を決定することは

非常に困難である．したがって，本研究では合成による分析法 (A-b-S 法) を利用し，実際の発話音声を分析することによって非言語情報を解析する．なお，非言語情報を解析する前に言語情報が決定されていることが必要である．具体的には以下の手順に従う．

1. 言語情報としてアクセント発生時刻と持続時間を入力する．
2. 発話文に対応した非言語情報に適当な初期値を与える．
3. 入力された言語情報と非言語情報から基本周波数を生成し，発話から抽出した基本周波数とモデルが生成した基本周波数とのずれを計算する．
4. 非言語情報の値を変更して 3. の比較を行う．非言語情報の取りうる値の範囲は非言語情報毎に決定される．
5. 3. ～4. の結果，最もずれが少ない場合の非言語情報の値を分析した発話の基本周波数に対する非言語情報とする．

3. の比較を行うために，発話から抽出された基本周波数と生成した基本周波数の有声母音部でのずれを用いる．ずれの計算方法を式 (1) に示す．

$$diff = \frac{\sum_{N} \frac{F_0(n) - F_{0real}(n)}{F_{0real}(n)}}{N} \times 100 \qquad (1)$$

$diff$: ずれの平均 (%)
F_{0real}: 発話から抽出した基本周波数 (Hz)
F_0: 生成した基本周波数 (Hz)
N: 有声母音部のサンプル数

4 顔画像における不特定人物の個人性

人間の顔のつくりは個人によって多少異なるため，顔画像も当然個人個人で異なった特徴を持っている．したがって，不特定人物の顔動画像生成を行う場合，人間の顔の個人差をうまく表現する必要がある．

一般に顔動画像生成を行う場合，ワイヤーフレームを変形させることで口や顔表情の変化を実現している．そのために，顔動画像システムはジェネリックモデルと呼ばれる一般的な顔形状を表現したワイヤーフレームモデルを持っている．本研究では不特定人物を対象とするため，このジェネリックモデルを不特定人物の顔に合ったワイヤーフレームモデルに変形させることが必要となる．

4.1 ジェネリックモデル

顔動画像生成部では，顔をワイヤーフレームモデルで表現しそれを発話内容に応じて適切に変形することにより顔動画を実現する．顔のワイヤーフレームモデルはあらかじめ基本となるモデルを作成する．基本となるモデルは特定人物を対象に顔の形状データをサンプリングすることによって作成しておく．これをジェネリックモデルと呼ぶ (図 2)．

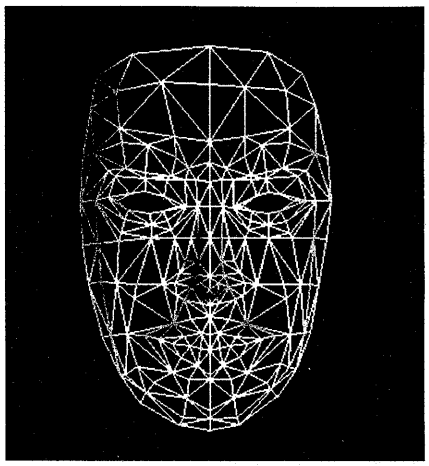

図 2: ジェネリックモデル

4.2 ジェネリックモデルの変形

不特定人物の顔動画合成を行なう場合，合成を行なう人物のワイヤーフレームモデルが必要である．しかし，別の人物の顔動画像合成を行う前にまずその人物のためのワイヤーフレームモデルを一から作り出すためには大変な労力が必要となる．そこで，その人物の顔を正面と側面方向から撮影した写真 2 枚を用いて，ジェネリックモデルを変形し，その人物に適合するワイヤーフレームモデルを生成する．この際にジェネリックモデルを構成する頂点の中から代表的

な点を選択してそれをジェネリックモデルの特徴点とする．またジェネリックモデルを構成する特徴点以外の頂点を非特徴点とする．具体的な作業手順は以下の通りである．

1. ジェネリックモデルの特徴点を正面画像では xy 平面，側面の画像では yz 平面上にそれぞれ配置する．配置したジェネリックモデルの特徴点を画像上の特徴点の位置に手動で移動することによって，ジェネリックモデルの特徴点に対応する対象人物の顔の画像上の特徴点の座標を得る (図 3)．

図 3: 画像上の特徴点の設定 (左:正面, 右:側面)

2. 1 の結果より特徴点の移動ベクトル,$U^i_{fea}, i = 1,2,...,m$ を正面，側面の画像についてそれぞれ算出する．

3. 正面，側面の非特徴点 jの移動ベクトル $U^j_{non}, j = 1,2,...,n$ はそれぞれ式 (2) によって求られる．

$$U^j_{non} = \sum_{i=1}^{m} f(r_{ij}) u^i_{fea} \tag{2}$$

式 (2) に用いられている $u^i_{fea}, i = 1,2,...,m$ は式 (3) を解いて得られたものである．

$$\begin{cases} U^1_{fea} = \sum_{k=1}^{m} f(r_{k1}) u^k_{fea} \\ U^2_{fea} = \sum_{k=1}^{m} f(r_{k2}) u^k_{fea} \\ \vdots \\ U^m_{fea} = \sum_{k=1}^{m} f(r_{km}) u^k_{fea} \end{cases} \tag{3}$$

ここで $f(r)$ は 2 点間の距離 rに対する重み付けの関数で式 (4) を用いる．式 (4) 中の Kの値はモデルの大きさに基づく係数である．

$$f(r) = exp(-Kr) \tag{4}$$

4. 正面画像から得られた特徴点，非特徴点の移動ベクトルをジェネリックモデルを xy平面に投射した座標に適用して，正面からみた対象人物の顔のワイヤーフレームモデルの各頂点座標 (x_{front}, y_{front}) を得る．また側面画像から得られた特徴点，非特徴点の移動ベクトルをジェネリックモデルを yz平面に投射した座標に適用して，側面から見た対象人物の顔ワイヤーフレームモデル各頂点座標 (y_{side}, z_{side}) を得る．ジェネリックモデルに対応する対象人物の対象人物のワイヤーフレームモデルを獲得するため，これらを式 (5) によって結びつけることによってモデルの 3 次元座標を得る．

$$\begin{cases} x = x_{front} \\ y = (y_{front} + y_{side})/2 \\ z = z_{side} \end{cases} \tag{5}$$

この手法によって得られたワイヤーフレームと，それを用いて合成した顔画像を図 4に示す．

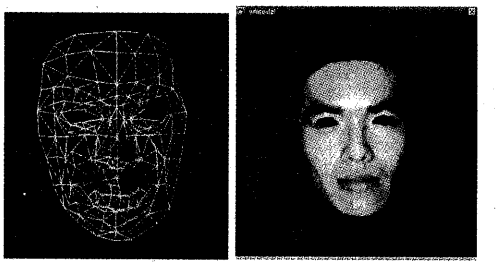

図 4: 生成されたモデルと顔画像合成例

4.3 発話動作の動画の生成

顔動画像生成部では，発話内容に応じて顔のワイヤーフレームを適切に変形することにより顔動画を実現する．人間の発話時の口形は幾種類かに分類できるので，ジェネリックモデルを作成する際にこれらの口形のワイヤーフレームモデルも同時に作成する．

発話時の顔動画像合成の際には，まず口形のワイヤーフレームモデルをジェネリックモデルから顔動

画像合成を行う人物のワイヤーフレームモデルを作成したのと同様に, ジェネリックな口形モデルを顔画像合成を行う人物の口形モデルに変形する.

この口形モデルを利用し, ある口形から別の口形へワイヤーフレームモデルを変形させることによって顔動画像を合成することが可能である.

4.4 顔動画像合成部と音声合成部の同期

音声合成部は発話内容のテキストを解析し, 発話の前に一文ごとに各音素の発声時間を出力する. 顔動画像合成部はこの出力結果を受けとり, 口形の表示プランを生成する. プラン生成が終了した時点で発声開始を音声合成部に告げ, 動画合成を開始する.

5 現状と今後

本稿では, 不特定人物の音声合成と顔動画像合成を行い, それらを同期させて動作させるために必要となる基本周波数生成手法および顔動画像合成システムの提案を行った. 現在, 音声合成部では東京と大阪の2方言の基本周波数を分析し, それぞれの非言語情報パラメータの決定法を求めている. 顔動画像合成システムでは, 図4に示したように, ジェネリックモデルを正則2枚の顔写真から変形し, 不特定人物の顔のワイヤーフレームモデルを作成することが可能である.

しかし, 現在の提案システムでは不特定人物の非言語情報パラメータは現在方言のようなおおまかな話し方の特徴のみを扱っている. しかし, 同じ方言を話す話者での個人の話し方の違いを表現するためには非言語情報パラメータをさらに精緻に分析する必要がある.

今後は上述のような問題点を克服し, 不特定人物の音声合成と顔動画像合成システムを作成する.

参考文献

[1] 倉立, :"三次元運動学に基づく顔面アニメーション", グラフィックスと CAD, pp.31-36(1997.10)

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

[3] 阿部匡伸: "音声合成技術がもたらすコミュニケーション革命", InterCommunication, No.20 Spring, pp.166-175(1997).

[4] 森山高明, 小川均, 天白成一: "大阪方言に見られる特徴的な基本周波数変化の分析" 音講論集1-Q-20, pp.337-338(1997-3)

[5] 森山高明, 小川均, 天白成一, 橋本雅行: "言語情報と非言語情報を利用した基本周波数制御の新手法", 情報処理学会研究報告, 97-SLP-17, pp.97-102(1997-7).

[6] 藤崎博也, 須藤寛: " 日本語単語アクセントの基本周波数パタンとその生成機構のモデル", 日本音響学会誌, 27巻, 9号, pp.445-453(1971).

[7] 森山高明, 小川均, 天白成一, 橋本雅行: "非言語情報を利用した大阪方言発話の基本周波数生成手法", 音講論集, 2-2-4, pp.231-232(1997-9).

[8] 森山高明, 小川均, 天白成一, 橋本雅行: "大阪方言発話の基本周波数分析", 音講論集, 1-7-21, pp.221-222(1998-3).

[9] Takaaki MORIYAMA, Hitoshi OGAWA, Seiichi TENPAKU: "A new control model based on rising and falling fundamental frequency", Proc. of ASA/ASJ Third Joint Meeting, pp.1171-1176(1996-12).

[10] H.Fujisaki, M.Tatsumi, and N.Higuchi: "ANALYSIS OF PITCH CONTROL IN SINGING", "*VOCAL FOLD PHYSIOLOGY*", UNIVERSITY OF TOKYO PRESS, pp.347-363 (1981).

[11] 武田 昌一: "日本語音声合成における音素とプロミネンスの影響を考慮した韻律制御に関する研究", 東京大学工学部博士論文 (1991).

[12] 杉藤美代子: "日本語アクセントの研究", 三省堂 (1982).