音声駆動型リップシンクの動作原理と３Ｄキャラクターアニメーションへの応用

J-39 Speech-driven Lip Sync techniques for 3D computer character animation

河野 哲也† 澤田 佳之† 荒木 真一† 栗原 芳己†
KONO Tetsuya SAWADA Yoshiyuki ARAKI Shin'ichi KURIHARA Yoshimi

１．はじめに

コンピュータゲームの映像は年々高精細になり，それにつれて表現力が向上している．そのため，リアルな表現をゲーム機上で再現することが必要になってきている．かつて，ゲームキャラクタのアニメーションは経験や勘を頼りに手作業でつけていたという背景がある．ここ数年，比較的簡単に人体の動きを記録できるモーションキャプチャ技術が発達したため，リアルなアニメーションを効率よく作成できるようになっている．一方，ゲーム機の表現力が向上するに伴い，身体の動きだけではなく，ゲームキャラクタの顔の微細な表現が可能になってきている．そのため，自然な表情アニメーションの需要が高まりつつあり，台詞に合わせて口を動かすシーンも多くなっている．

従来は，口の動きを表現する手法として，作業者が手作業で口やその周辺の形状を決定する手法が用いられてきた．作業者は，ある場面におけるキャラクタの音声を聞きながら，あるいはこの場面用に用意された台詞を読みながら，キャラクタの口の動きを想像して口の形状を作っており，自然な口の形状を再現しようとすると，ある程度の経験や勘が必要になる．ところで，このような手作業によってキャラクタの口やその周辺の形状を決定する手法では，作業者の経験や勘など各個人の熟練度や能力に左右されるため作業効率が悪く，しかも作業者毎に異なる結果が得られるため再現性が低く，自然な表情を生成することが難しいという問題があった．また，音声のタイミングに同期させて口形状を決定するには，熟練と膨大な作業時間を必要としていた．

以上のような課題を解決するために，本稿では音声分析手段，口形状再現手段，座標修正手段を備えており，発声する音声を口形状に反映させる表情アニメーション作成手法を考案する．その目的は自然な表情が得られ，音声に対応して口の動き等が変化する表情アニメーションを生成することである．

２．音声データの解析

音声データから口形状を決定するには，まず音声解析を行い，言語的特徴を抽出する必要がある．音声の言語的特徴はスペクトルとその時間変化によるので、それらを調べ，音声から口の形状の決定を試みた．

(1) フォルマント周波数抽出による音声解析

音声の言語的特長として、フォルマント周波数に注目し以下の様に求めた．

a) アニメーションのフレームレートに合わせて音声データの解析長と解析間隔を設定する．
b) 音声信号から１フレームを取り出し，線形予測係数を算出する．
c) 線形予測係数から入力信号のスペクトル包絡を求める．
d) スペクトル包絡からフォルマント周波数を求める．

この結果から，フォルマント周波数の時間変動が安定した母音位置の抽出が可能になり，自動リップシンクが可能になった．更にフォルマント周波数もある程度の話者依存性があるため，事前に“あいうえお”の５母音を上記の手順で解析し基準用データとして用いる事により，母音抽出の精度を上げることができた．

しかし，この方法では発声中にも口を閉じる音である「マ行」「パ行」「バ行」などの子音の位置とその種類の特定をする事はできず，やはり作業者による修正が必要であった．そのため，子音抽出を行うため幾つかの方法を試みた後，次の手法を採用した．

(2)「大語彙連続音声認識デコーダ Julius」による音声解析

「大語彙連続音声認識デコーダ Julius」‡(以下 Julius)と単語・音素セグメンテーションキット(以下セグメンテーションキット)を用いることにより，子音を含めた音声の言語的特徴を取り出せるようになった．さらに，表情アニメーションに応用するために，以下の変更を加えた．

a) Julius の最大解析時間長を変更できるようにした．
b) 音声データ中に‘0’があった場合に Julius が停止してしまうのを回避した．
c) 母音子音抽出結果をフレームレートに合わせて出力するように，スクリプトを追加した．
d) １フレーム内の平均音量を出力するようにスクリプトを追加した．

さらに，セグメンテーションキットで必要とされている「単語単位の書き下しファイル」(以下台本)を用いた場合と付属辞書を用いた場合の両方を試した．

台本を用いた場合には，音声データと解析結果が異なることなく母音子音の抽出が成功していた．さらに Julius の音素モデルを日本語のままで入力音声を英語音声とし台本に聞こえたようにひらがなで記述した場合にも，日本語として認識され母音子音の抽出が可能であった．しかし，付属辞書を用いた場合，入力された音声に含まれる単語が辞書に無い場合が多く，ほとんどの場合解析結果は入力と異なるものになった．

これらの結果から「Julius と台本付きセグメンテーションキット」を用いることによって，口形状を決定するのに必要な音声解析結果が得られるようになった．

３．表情アニメーションへの応用

口形状は母音，子音の音素ごとに用意した．(図１) 音素 i に対応するある時刻 t における音声解析結果を $\alpha_i(t)$，顔オブジェクトの任意の頂点座標について，音素 i を発音しているときに対応する座標と発音していないときの標準的な表情との差分を $\Delta\vec{P}_i$，音素の数を n とすると，時刻 t における音声に対する頂点座標 $\Delta\vec{P}(t)$ は

$$\Delta\vec{P}(t)=\sum^{n}\alpha_i(t)\Delta\vec{P}_i$$

---

† (株)ナムコＣＴカンパニーＣＴ技術環境グループ, NAMCO LIMITED

となる．ところで，表情アニメーションはキャラクタが発音する音声に対応する口形状を決定するだけでは不十分で，キャラクタの感情を映像の中で再現する必要がある．ある時刻tにおける感情 j の度合いを$\beta_j(t)$，顔オブジェクトの任意の頂点座標について，感情 j に対応する座標と感情がないときの標準的な表情との差分を$\Delta\vec{E}_j$，再現するべき感情の数をmとすると，時刻tにおける感情に対する頂点座標$\Delta\vec{E}(t)$は，

$$\Delta\vec{E}(t)=\sum_{j}^{n}\beta_j(t)\Delta\vec{E}_j$$

となる．

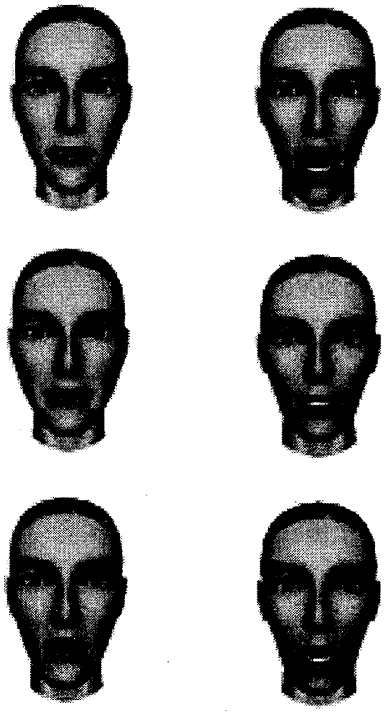

図1　基本表情例

ゲーム制作では多くの場合，画面上で見栄えの良い表情にするため，作業者が表情の微調整を行うため，音素や感情パラメータから得られた頂点座標を作業者が修正する手段を用意する．つまり，最終的な頂点座標は$\Delta\vec{P}(t)$および$\Delta\vec{E}(t)$で決定された頂点座標に作業者が微調整を加える余地を残しておく．また，音声解析から得られた口形状に対しても，場面に応じて口の開け方などを決定するパラメータ$\gamma(t)$を用意し，作業者が場面に応じてより自然な表情アニメーションを作成できる仕組みを用意する．頂点座標の微調整を$\Delta\vec{M}(t)$とすると，最終的な顔オブジェクトの任意の頂点座標と標準的な表情との差分$\Delta\vec{C}(t)$は

$$\Delta\vec{C}(t)=\gamma(t)\Delta\vec{P}(t)+\Delta\vec{E}(t)+\Delta\vec{M}(t) \quad \cdots (1)$$

となる．

顔オブジェクトを変形させる方法は，大きく分けて2種類ある．一つはオブジェクト頂点の移動量をコントロールポイントに関連付けて，そのコントロールポイントを動かしてオブジェクトを変形させる方法と，もう一つは頂点座標が異なる幾つかのターゲットと呼ばれるオブジェクトを作成しておき，ターゲットの割合を変化させて頂点座標を変化させる方法である．前者の場合，頂点座標をコントロールポイントに置き換えればそのまま応用可能であり，後者の場合，（1）式は$\Delta\vec{M}(t)=0$として頂点座標に対してそのまま適用できる．実際に上記二つの方法について（1）式を適用したところ，表情アニメーションを得ることができた．音声解析結果を表情アニメーションに適用するには，母音と子音によって口形状に影響する時間を変えるとともに，音声と口形状を表示するタイミングを調整する必要があるが，これらの調整を行うことにより，より自然な表情アニメーションを得ることができた．

音声解析結果$\alpha_1(t)$は図2のようになっているが，この結果を平滑化すると，音声解析結果のノイズの影響を受けにくくなり，より自然な表情アニメーションが得られると考える．平滑化とキーフレームの削減は，宮沢らの研究による「3Dアニメーション定義における拡張されたLODについて」で述べられている最適キーフレームの自動決定技術（ATKINS）を音声解析結果$\alpha_1(t)$に対して適用することにより可能になると考える．これにより，より自然な表情アニメーションが得られると同時に，作業者の修正作業を容易にすることを検討している．

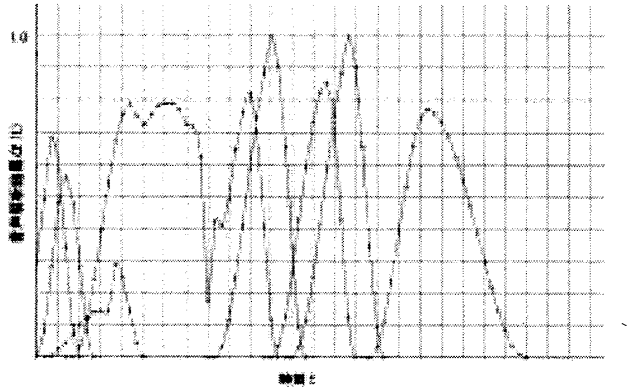

図2　音声解析結果$\alpha_1(t)$

4．おわりに

本稿では音声分析手段，口形状再現手段，座標修正手段を備えており，発声する音声を口形状に反映させる表情アニメーション作成手法を考案した．また，この手法を用いることにより，自然な表情が得られ，音声に対応して口の動き等が変化する表情アニメーションを生成することができた．

参考文献

1）鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄：「音声認識システム」（オーム社，2001）
2）今井聖：「音声信号処理」（森北出版株式会社，1996）
3）中尾喜保，宮永美知代：「美術解剖学アトラス」（南山堂，1995）