データドリブンなフォトリアル口内アニメーションの自動生成

川井 正英[1,a)]　岩尾 知頼[1]　前島 謙宣[1]　森島 繁生[1,b)]

概要：CG 発話アニメーションの自動生成手法は既に数多く提案されている．しかし，口内 CG アニメーションの自動生成手法は未だ提案されておらず，特に舌の複雑な表現を行うことは大きな課題である．そこで本研究では，既存手法によって生成された発話アニメーションの口内領域に対して実際に撮影した口内画像を挿入し，その口唇部に複数人の口唇画像を用いて Visio-lization 法を施すことで，複雑な口内表現を可能にした．

Automatic Generation of Photorealistic Inner Mouth Animation Driven By Data Stream

Kawai Masahide[1,a)]　Iwao Tomoyori[1]　Maejima Akinobu[1]　Morishima Shigeo[1,b)]

Abstract: There are a lot of CG speech animations proposed by researchers. However, automatic generation of CG inner mouth animations is still challenging, especially, complex tongue movements are not represented. Therefore, this paper describes a novel method that inserts inner mouth images and applies Visio-lization to the speech animation proposed by previous works. As a result, we can represent complex inner mouth animations.

1. はじめに

CG 映画やゲーム等の制作現場において，写実的かつ説得力のある発話アニメーションを制作することは未だに重要なトピックとして挙げられる．しかしながら現状では，この制作にアーティストの経験に基づく精巧な手作業が必要であり，多大な労力や時間がかかる問題がある．この問題を解決するために，自動的に発話アニメーションを生成する手法が，多くの研究者によって提案されてきた．キーシェイプモデルを用いたブレンドシェイプ手法や，モーションキャプチャを用いたリターゲッティング手法などでキャラクタの発話を表現する，3 次元モデルを用いた手法 [1,2,3,4] や，事前に撮影したビデオのデータベースを用いて任意の発話シーンを合成する，2 次元のイメージベースな手法 [5,6] などがある．両手法とも，簡易に発話アニメーションを生成する代表的な手法であり，リップシンクの精度は高い．しかし，口内の動きの表現には乏しく，ましてや口内の詳細表現は未だ適切に表現できずにいる．つまり，従来手法では高精度に口形の動きを表現可能だが，とりわけ口内の動きの表現には致命的な課題がある．そのため，アニメーション生成の後処理として，口内の複雑な動き（詳細表現を含む）を自動付加することで，アニメーションのクオリティを向上できると言える．それゆえ本稿では，既存手法により生成された写実的な口形表現を持つアニメーション（以後，既存アニメーション）に対して，口内表現を付加することで，写実性豊かな口内表現を持つ発話アニメーションを自動再生成する手法を提案する．本手法は，実在人物や CG キャラクタを含む幅広い発話アニメーションに適用でき（図 1），歯で舌を噛むような音節/θ e/の動きや舌の裏側の表現といった複雑な詳細表現も可能とした．システムの入力は，既存アニメーションの動画像，歯の見える正面顔画像（以後，正面歯画像と呼ぶ）1 枚，及びセンテンステキストの最低限の 3 点に抑えた．ま

[1] 早稲田大学
Okubo, Shinjyuku-ku, Tokyo 169–8555, Japan
[a)] doara-waseda@toki.waseda.jp
[b)] shigeo@waseda.jp

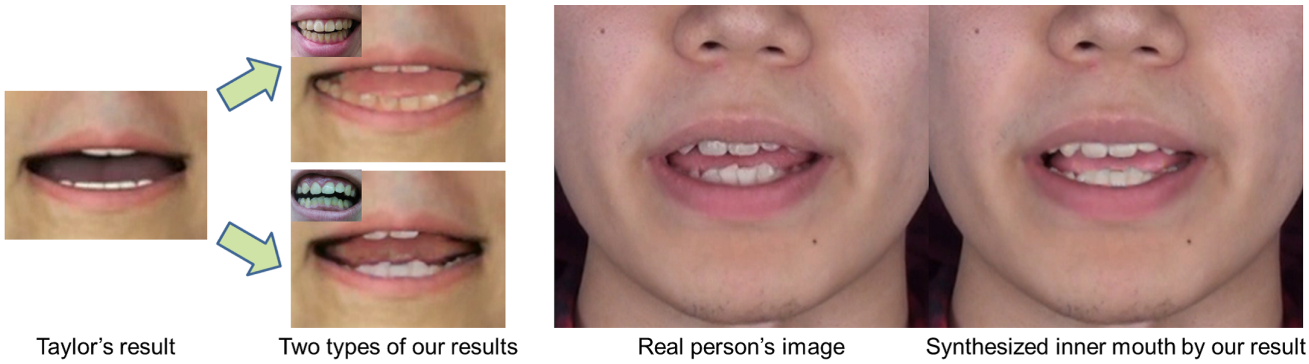

図 1 発話アニメーションのワンシーン比較
左：“Dynamic Units of Visual Speech” proposed by [Taylor et al. 2012]. との比較
右：実写画像との比較

た，事前に任意の 1 人物の舌の動き，多人数の通常発話を動画撮影し，それぞれ連番舌画像データベース，口唇画像データベースを構築しておく．従来手法により生成されてきた口内表現が乏しい既存アニメーションに対して，口内を写実的に改善した新たなアニメーションを生成することができる”口内自動付加フィルター”としての応用を期待している．

2. 関連研究

過去には，発話アニメーションを生成するために，数多くの方法が提案されてきた．Chang ら [6] は，Multidimensional Morphable Model を用いて，個人の発話スタイルを反映した発話アニメーションの生成するイメージベースの手法を提案した．この手法では，特定人物の動画に対して特徴点をとることで，口形の動きを取得し，そのデータを新しい人物に付加させることで，任意人物の個人性を反映した発話アニメーションを生成できる．しかしながら，口形の動きに合わせて口内画像がモーフィングされるため，歯や舌が伸縮して見えるという問題があった．また，Taylor ら [7] は，”Viseme”で定義される口形を接続することで，リアルにリップシンクした発話シーンを生成する手法を提案している．この手法では，発話動画から取得した Active Appearance Model パラメータの時間推移をセグメンテーションし，各セグメント（”Viseme”）と音素情報との対応付けを行う．この対応関係を用いることで，任意の入力音声に対し最適な”Viseme”の組み合わせが選択され，リアルなアニメーションが生成可能となる．しかしながら，口内は厳密に定義されておらず，口内画像の時間的連続性を保つために，予め作られたモデルの舌は全く動かないモデルを使用せざるを得なかった．さらに Li ら [8] は，事前知識を用いて，1 つのターゲットシェイプのみからターゲットらしい様々な表情を作成できるブレンドシェイプのための手法を提案した．しかしながら，彼らの手法で合成されたモデルの口内は全く空の状態である．因みに，モーションキャプチャリングシステムを用い，顔表情の動きデータを取得し転写する手法なども存在するが，そもそも口内動きデータはそのシステムでは取得できない．以上，上記に示した Chang ら，Taylor ら，Li らの手法，及びモーションキャプチャリングシステムを用いた手法のように，口形の動きを写実的に再現する研究は数多く行われているが，口内表現に注力した発話アニメーション合成手法に関する研究は行われていない．しかし，従来手法による結果と本手法結果を比較する必要はある．

3. 前準備

人間は，発話の際，歯と舌をある程度独立に動かすことができる．例えば，一定の歯の開き具合を保ったまま，舌を自由に動かすことが可能である．つまり，歯と舌を一色単にして考え，そのデータベースを構築することは不可能ではないが現実的ではない．そこで本研究では，複雑な口内の動きを少量のデータベースから表現するために，歯の動きと舌の動きを別々に分類し，それぞれの実写画像データベースを構築する．また，別々に挿入された歯と舌に生じる違和感を解消するために，Visio-lization 法（後述）の応用を施す．そのために，口唇画像データベースを構築する．なお，歯に関するデータベースは，事前に構築する必要がないため，4 章で示す．さらに入力として，ベースとなる既存アニメーションの動画像，歯画像データベースを自動構築するために必要となる正面歯画像 1 枚，及び舌画像挿入の際に必要なセンテンステキストの 3 点を準備する．

3.1 連番舌画像データベースの事前構築

事前に，一被験者の舌の動きの様子を動画撮影した．撮影にはソニー社製の HD ビデオカメラを 29.997[fps] で使用し，図 2 に示すように照明を 3 台並べて撮影した．なお，その際に，唇を大きく広げることができる TC マート社製のアングルワイダー [9] という実験機器を用いた．これを

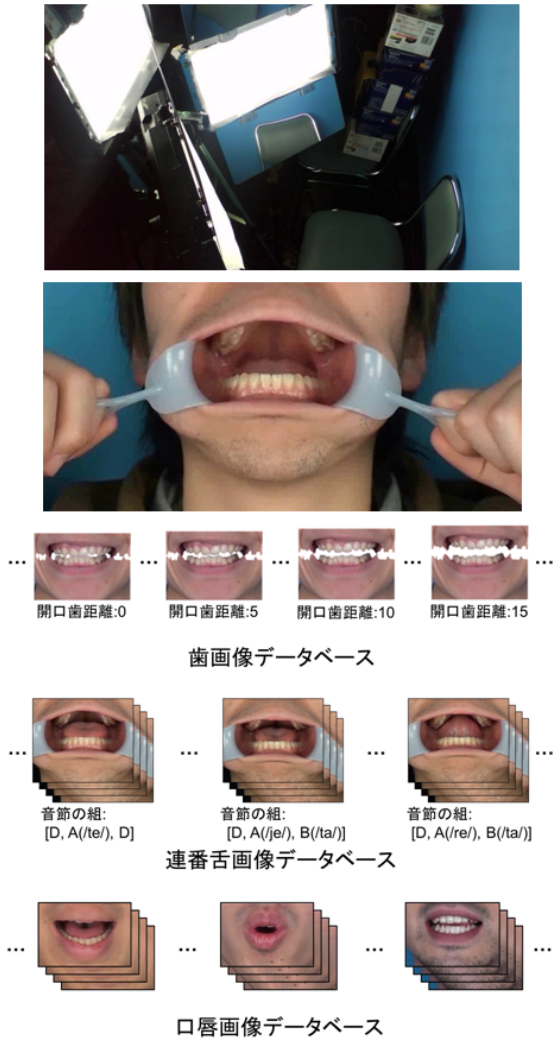

図 2 アングルワイダー使用時の口内の様子，撮影環境，及び構築されるデータベース例

名称	舌の位置	状態	例
前母音	前方	1	/e/, /æ/
	前方	0	/i/
後母音	後方	0	/a/, /u/

表 1 母音による舌の位置の分類

用いることで，唇により遮断されることなく舌の大部分を撮影することができる．アングルワイダー使用時の口内の様子を図 2 に示す．以下に示すように音節によって舌の動きを分類し，データベースを構築する [10]．音節とは，母音や子音により表現されるため，母音と子音による舌の位置の違いを考える必要がある．舌が見える状態を 1，舌が見えない状態を 0 とする．まず母音は，その調音を発声した際の舌の先端部の位置によって分類される．舌の前後関係から，母音を舌が前方にある場合（前母音）と舌が後方にある場合（後母音）の 2 種類に分類できる．母音による舌の位置の分類を表 1 に示す．

前母音である音素/i/は閉母音とも呼ばれ，歯が閉じているため舌が見えず，状態 0 と分類される．次に，調音の位置（以後，調音部位と呼ぶ）による子音の分類をする．子音は発音する際に，呼気の通路がどこで遮断されたり狭められたりするかによって 8 種類に分類される．子音による舌の位置の分類を表 2 に示す．

名称	舌の位置	状態	例
両唇音	上下両唇の間	0	/p/, /b/
唇歯音	上歯と下唇の間	0	/f/, /v/
歯音	上歯と舌尖との間	1	/θ/
歯茎音	上歯茎と舌尖との間	1	/t/, /d/
硬口蓋歯茎音	歯茎から硬口蓋にわたる部分と舌端との間	1	/r/
硬口蓋音	硬口蓋と前舌面との間	1	/j/
軟口蓋音	軟口蓋と前舌面との間	0	/k/, /g/
声門音	両声帯の間	0	/h/

表 2 子音による舌の位置の分類

名称	状態	例
前母音	A(1)	/e/
歯音＋前母音	A(1 → 1)	/θ//e/
歯茎音＋前母音	A(1 → 1)	/t//e/
硬口蓋歯茎音＋前母音	A(1 → 1)	/r//e/
硬口蓋音＋前母音	A(1 → 1)	/j//e/
歯音＋後母音	B(1 → 0)	/θ//a/
歯茎音＋後母音	B(1 → 0)	/t//a/
硬口蓋歯茎音＋後母音	B(1 → 0)	/r//a/
硬口蓋音＋後母音	B(1 → 0)	/j//a/
両唇音＋前母音	C(0 → 1)	/p//e/
唇歯音＋前母音	C(0 → 1)	/f//e/
軟口蓋音＋前母音	C(0 → 1)	/k//e/
声門音＋前母音	C(0 → 1)	/h//e/
後母音	D(0)	/a/
両唇音＋後母音	D(0 → 0)	/p//a/
唇歯音＋後母音	D(0 → 0)	/f//a/
軟口蓋音＋後母音	D(0 → 0)	/k//a/
声門音＋後母音	D(0 → 0)	/h//a/

表 3 音節による舌の動きの分類

音素間の時間的な連続性を考慮してデータベースを構築するには，母音と子音を独立に表さずに，母音や，子音と母音の組である音節として表現する必要がある．音節を構成する要素である母音と子音の音素による舌の位置の分類は上述したため，それを考慮に入れて，音節による舌の位置の分類を考える．そのため，音節毎の舌の位置の推移の違いに注目し，下記の状態 A, B, C, D のように舌の位置の推移を分類した．

A.「舌が見える → 舌が見える」 状態
B.「舌が見える → 舌が見えない」 状態
C.「舌が見えない → 舌が見える」 状態
D.「舌が見えない → 舌が見えない」 状態

音節による舌の動きの分類を表 3 に示す．

D に関しては，終始舌が見えない状態であるため，全てのパターンをまとめると，A は 5 パターン，B は 4 パター

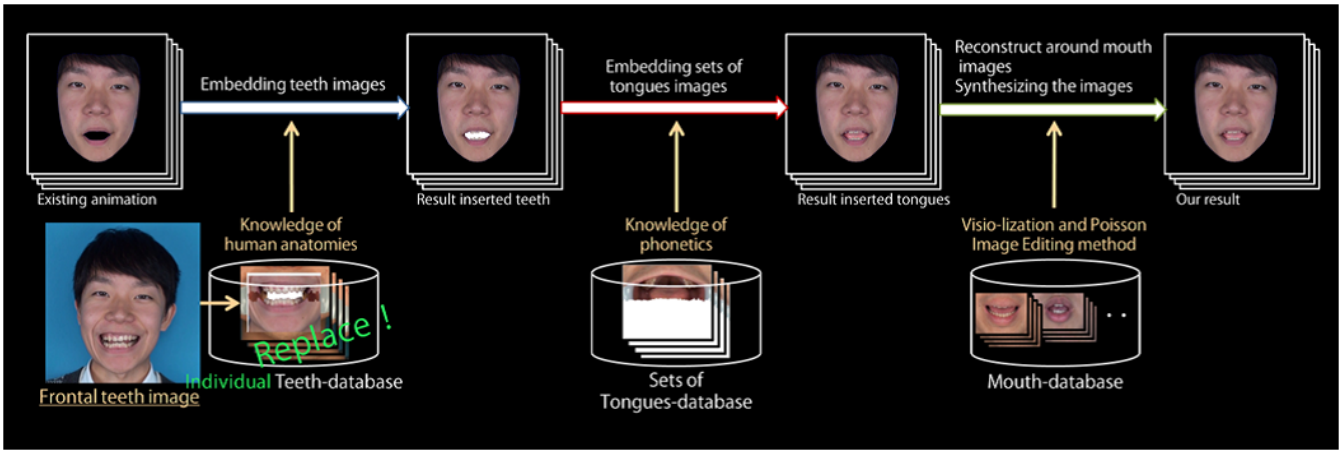

図 3 本研究の概要

ン，C は 4 パターン，D は 1 パターンとなる．このパターンの本質的な意味は，舌の見え方には様々な種類あるということである．単に状態 A であると言っても，舌を巻いて舌の裏側が見える様子（/r//e/）や舌を歯に接触させる様子（/t//e/）など数パターンあるため，表 3 のようにパターンを分類した．表 3 に示したパターンを組み合わせることによって英語の全ての発話を網羅できる舌のバリエーションが表現できる．今回，時間的連続なデータベースを構築するために，この 4 状態をさらに組み合わせて，連続的な舌の動きを持った「音節の組」を作った．実際には，「D, A(/te/), D」（舌が見えない，見える，見えない）の組のように「舌が見えない」で始まり，途中で「舌が見える」になり，最後に「舌が見えない」で終わるように組み合わせた文章を発話した，1 人分の動画を撮影した．この組は「C, B or D」，「C, A, B or D」，「D, B」，「D, A, B or D」の組み合わせのみある．合計で，4 × 5+4 × 5 × 5+1 × 4+1 × 5 × 5=149 組あり，149 組の発話を動画撮影（1 組あたり約 2 秒）し，動画から歯を取り除いた画像を取得し，連番舌画像データベースとした．データベースの画像サイズは，入力の既存アニメーション中の人物の両目間距離を基準にリサイズされ，その後口周辺部のみの 241 × 201[pixel] 取得した．連番舌画像データベースの例を図 2 に示す．

3.2 口唇画像データベースの事前構築

事前に，通常発話の様子を動画撮影した．撮影環境は 3.1 章に示した通りである．具体的には，基本 5 母音（/a/, /i/, /u/, /e/, /o/）と，舌の様々な動きを取得するために，代表的な音節（/θ//e/, /t//e/, /r//e/, /j//e/, /p//a/, /f//a/）を発話した 7 人分の動画を撮影した．発話時間に関して，個人差はあるものの 1 人当たり約 10 秒となり，それらの動画から口周辺部を取得し，7 人分をまとめて 2213 枚の口唇画像データベースとした．データベースの画像サイズは 3.1 節と全く同様にして，241 × 201[pixel] とした．口唇画像データベースの例を図 2 に示す．

3.3 入力の準備

まずは，ベースとなる既存アニメーションを作成する必要がある．2 章で述べた通り，発話アニメーションの生成法は種々存在するが，本稿では，前島ら [11] と三間ら [12] の手法を組み合わせた簡易的な発話アニメーションの生成手法により既存アニメーションを作成した．なお本研究では，生成される動画像のサイズを，512 × 512[pixel] とした．二つ目の入力である正面歯画像は，歯画像データベースを作製するために必要となる．これは実際に歯を剥き出しにした画像を撮影したもので，後にリサイズするため，画像サイズは任意とする．生成された既存アニメーションの動画像と，撮影された正面歯画像は図 3 の左部に示す通りのものである．最後に，センテンステキストの用意をする．このテキストには，既存アニメーション中どのフレームでどの音節を発しているかが記入されている．例えば，36 フレーム目に，音節/t//e/を発していた場合，te:36 と記入される．

4. 口内自動付加フィルター適用

図 3 に本研究の概要を示す．本手法は，既存アニメーションに対して口内情報を付加することで，より写実性のある発話アニメーションを新たに作り出すものである．本章では，歯画像データベースを作成し，それと予め構築されたデータベース中の画像を用いて，既存アニメーションに口内画像を挿入する方法について述べる．その後，別々に挿入された歯画像・舌画像の間に生じる違和感の解消方法や，より写実的に口内を表現する方法について述べる．

4.1 歯画像データベースの作製

正面歯画像 1 枚を使用して，自動的に歯画像データベースを作成する．具体的にはまず，正面歯画像から特徴点を 40 点検出 [13] することで，顔器官の位置情報を取得する．その上で正面歯画像全体を，検出された特徴点の位置情報から計算される両目間距離を基に，3.3 節で生成された発話アニメーションの両目間距離に合うようにリサイズする．また，ここで，歯の中心位置を，唇の左右両端に対応する特徴点の位置座標から計算しておく．そして，上歯下部と下歯上部領域をミーンシフト法 [14] により自動抽出し，それと前述した歯の中心位置との，2 つの情報から上歯下部中心座標と下歯上部中心座標を取得した．取得した上歯下部中心座標から下歯上部中心座標までの距離（開口歯距離）を 0〜50[pixel] の間を 1[pixel] 毎に変化させ 51 枚の歯の開閉画像を生成し，歯画像データベースを作成できる．なお，データベースの画像サイズは，241 × 201[pixel] とした．歯画像データベースの例を図 2 に示す．

4.2 頭蓋骨の構造を考慮した歯画像挿入

入力画像の口内部に歯画像データベースから適切な歯画像を選択し，挿入する．そのため，口内が全く空の入力画像から，歯の位置を推定する必要がある．本稿では，頭蓋骨の構造を考慮の上，「鼻頂点から上歯の距離（鼻上歯距離），また顎から下歯の距離（顎下歯距離）は常に一定である」という生体構造の知見から，鼻頂点と顎の位置から歯の位置を推定する．頭蓋骨の構造上，開口時も閉口時も骨の長さは変化しないため，鼻上歯距離と顎下歯距離は常に一定である（図 4）．歯画像選択の概要を図 5 に示す．この知見に基づいて，実際に歯画像を挿入する．上記の知見は，距離一定性を示すもので，上歯と下歯の初期位置さえ取得できれば，全フレームの歯の位置を鼻頂点と顎から推定できるものである．なお，鼻頂点と顎の頂点座標は特徴点検出によって算出でき，歯の初期位置は閉口した時のフレームを基準とし，唇の境界の位置を歯の境界の位置と仮定の上，取得する．知見により，全フレームに対して，鼻頂点と顎の位置から上歯下部中心座標と下歯上部中心座標を推定し，開口歯距離を決定する．あるフレーム f に対して，入力動画中の開口歯距離を d_{I_f} とし，歯画像データベース中の開口歯距離を d_{D_i} とする．そのとき，開口歯距離差が最小となるデータベース中の任意の歯画像 i

$$\arg\min_i |d_{I\_f} - d_{D\_i}| \ (0 \leq i \leq N) \tag{1}$$

を選択し，入力画像に歯画像を挿入する．本研究では，データベースの画像数 N=51 である．

4.3 センテンステキストに対応した舌画像挿入

歯画像挿入後，音節毎にフレーム番号が記述された入力

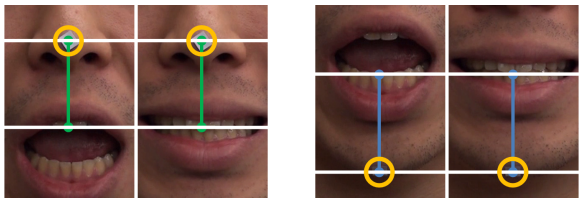

図 4 開口時・閉口時の鼻上歯距離と顎下歯距離

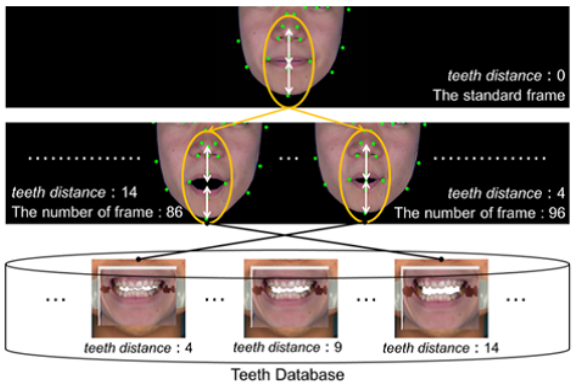

図 5 歯画像選択の概要

のセンテンステキストに合わせて，連番舌画像データベースから適切な連番舌画像を選択し，挿入する．連番舌画像選択の概要を図 6 に示す．例として，I take a yellow book and [ai teik a jelou buk end] というセンテンスが与えられた場合を考える．これを音節の組毎に [ai, te, ik][a, je, lou][buk, e, nd] のように分離することで，データベース中の [D, A(/te/), D][D, A(/je/), B(/ta/)][D, A(/e/), D] の組と各々対応付けることが可能となり，連番舌画像を選択できる．しかし，ある音節の組 [D, A(/te/), D] のフレーム区間，つまり入力画像中で [ai, te, ik] と発している区間のフレーム画像数（I_N）とデータベース中の [D, A(/te/), D] の連番画像数（D_N）が合わない場合があることは明白である．連番舌画像中の画像数よりも入力画像中のフレーム画像数が多い場合は，連番舌画像の一部を重複させて選択し，逆の場合は，連番舌画像の一部を間引いて選択する．ある [ai, te, ik] と発している区間において，n フレーム目に選択されるデータベース画像中の画像番号を N_n とすると，舌画像の選択は式 (2) に従う．

$$N_n = \sum_{n=1}^{I_N} \lceil n \times \frac{D_N}{I_N} \rceil \tag{2}$$

また，連番舌画像データベースを，「舌が見えない」で始まり，「舌が見えない」で終わるように構築していた．この理由としては，連番舌画像の接続部，つまりはデータベース同士の接続に関しては，「D と D」や「B(/ta/) と D」のように，両者とも舌が見えない場合同士の接続となるため，画像間の時間的な連続性を保ちつつ挿入できる．

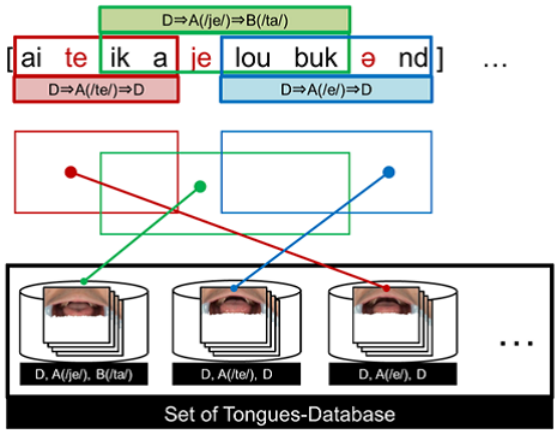

図 6 連番舌画像選択の概要

4.4 写実的な画像の再生成

4.2,4.3 節に示した方法で，別々に歯画像と舌画像を挿入したため，その境界には生じる不自然な輝度変化が生じる．それが原因で動画として再生した際に，唇・舌・歯が独立に動くように見えてしまう．また，歯で舌を噛むような歯と舌を関連させた複雑な口内表現は困難である．歯と舌を独立として考えることは，データベース構築の際の負担を激減できたが，上記に示したような致命的な弊害も生じてしまう．その問題点を解消するために入力画像の口周辺部に Visio-lization 法 [15] を適用する．Visio-lization 法の概要を図 7 に示す．具体的には，入力画像と口唇画像データベース中の画像をパッチという矩形に複数区切り，それぞれ同じ位置でのパッチ間の RGB 距離を計算する．次に，あるフレーム f，あるパッチ領域 Ω に対して，入力画像のある位置 (x, y) における RGB 値を $\boldsymbol{C}_{I\_f\_xy} = R_{I\_f\_xy}, G_{I\_f\_xy}, B_{I\_f\_xy}$ とし，データベース画像のある位置 (x, y) における RGB 値を $\boldsymbol{C}_{D\_i\_xy} = R_{D\_i\_xy}, G_{D\_i\_xy}, B_{D\_i\_xy}$ とする．パッチ毎に RGB 距離が最小となるデータベース中の任意の口唇画像 i

$$\arg\min_{i} \sum_{(x,y)\in\Omega} ||\boldsymbol{C}_{I\_f\_xy} - \boldsymbol{C}_{D\_i\_xy}||^2 \ (0 \le i \le N) \quad (3)$$

となるパッチを選択し，画像の貼り換えを行う．データベース中の自然な画像を用いて入力画像を貼り換えることで，境界の不自然さを解消しつつ，入力画像に似た画像を生成できる．一般的に顔画像に適用する際は，パッチサイズを 20 × 20[pixel]，重複部分を 3[pixel] で十分自然に合成できる．しかしながら，この Visio-lization 法は 1 枚の静止画を対象に適用するため，時系列を考慮していない．あくまでもその静止画内の最適なパッチを選択する手法なため，連番画像に適用すると，パッチごとに時間的不連続が生じてしまう．さらに，歯の一本一本を表現するような細かなものは再現できない．上記の 2 点の問題を解決するために，本稿では Visio-lization 法の拡張を行う．実際，パッチサイズを 6 × 6[pixel] という非常に小さいパッチを使用し，その重複部分を 3[pixel] という近接パッチと 50% の重複を行われるように設定した．6[pixel] のサイズは，一本の歯の大きさの約 3 倍小さい値であり，歯の一本一本まで細かな再現を可能とした．さらに，通常の Visio-lization 法では，入力画像とデータベース中の画像をパッチ同士の同じ位置のパッチで式 (3) の計算を行い，パッチを選択していたが，今回データベースのパッチを左右上下に参照範囲を大きく広げたことによって，擬似的にパッチ数を増加させた．このようにパッチの位置を固定しないことによって，データベースの擬似的増加ができ忠実に入力画像を表現する精度を向上できる他，広範囲からパッチを選択できるため，データベース中にはない歯と凹凸表現も可能となった．実際に従来 Visio-lization 法とその拡張を行った本手法との結果を比較したものを図 8 に示す．二段目より三段目，三段目より四段目の方がより入力画像の画像を再現できていることがわかる．なお，それぞれ使用した口唇画像データベースは，全く同じ画像を用いていることを言及しておく．

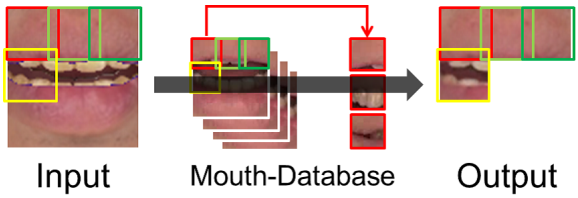

図 7 Visio-lization 法の概要

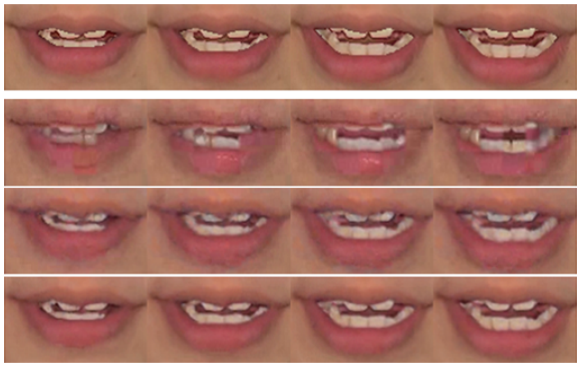

図 8 Visio-lization 法と本手法との比較
一段目：歯画像と舌画像挿入後の入力画像
二段目：パッチサイズ 20 × 20[pixel]，重複部分 3[pixel] で入力画像に Visio-lization 法適用後の画像（従来法）
三段目：パッチサイズ 6 × 6[pixel] 変更後の結果画像
四段目：パッチサイズ 6 × 6[pixel] 変更し，パッチの参照範囲を広げた後の結果画像

4.5 画像の転写

4.4 節で生成された口周辺画像を，入力画像につなぎ目なく挿入するために，Poisson Image Editing 法 [16] を導入する．Poisson Image Editing 法は，ソース画像の一部を

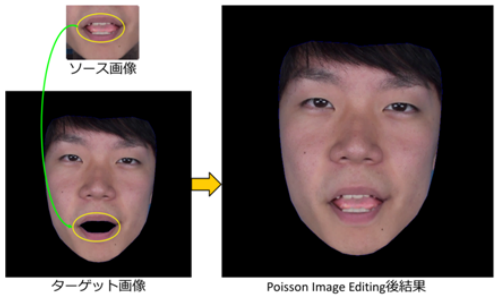

図 9 本手法結果の全体像

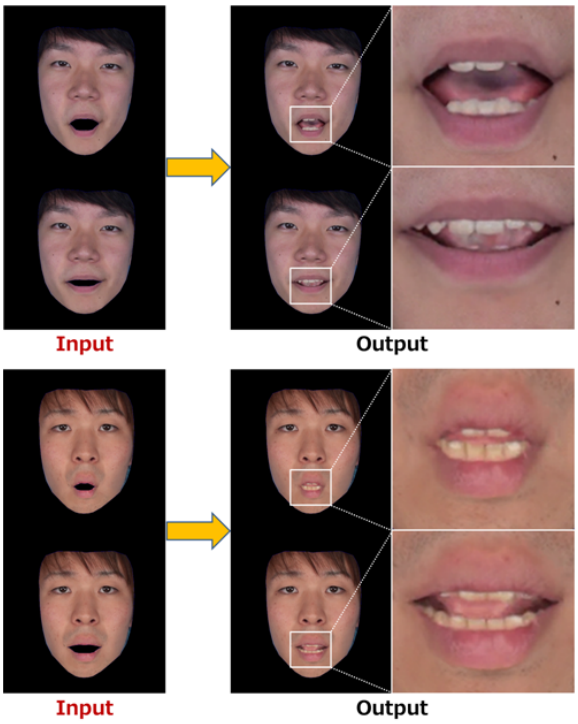

図 10 入力画像及びその画像に口内自動付加フィルターを適用した結果

ターゲット画像につなぎ目なく転写するものであり，ソース画像の色をそのまま移すのではなく，勾配情報を保存するように転写することでターゲット画像の色味に合った違和感のない合成画像を生成することができる．本研究では，ターゲット画像を入力画像，ソース画像を 4.4 節で生成された口周辺画像，図 9 の楕円内を転写領域とした．これにより入力画像中に存在するホクロ等の細部の情報を残しつつ，口周辺画像を挿入することができる．結果を同じく図 9 に示す．以上 4.1〜4.5 節に示した工程によって口内自動付加フィルターを適用できる．入力画像及びその画像に口内自動付加フィルターを適用した結果を図 10 に示す．

5. 実験

関連研究との比較を行うことにより，本手法の汎用性を証明する．今回，Taylor ら [7] が作成したデモムービー中のリップシンクアニメーションに対して，口内領域を手動で抽出し，4 章で示した工程によって口内自動付加フィルターを適用した．図 11 に示したのは，3 種類の異なる結果である．上段は Taylor らの結果の 5 枚の連番画像，中段及び下段は本手法適用結果の 5 枚の連番画像である．Taylor らの結果では，全く舌が動いていないのに対し，本手法結果では舌が前方に出てきている様子が見られる．また本手法は，口内に使用する歯画像を変更し，舌の輝度値を変化させることで，容易に様々な口内の印象を表現できる．このように口内の印象の入れ替えを行えるシステムは，映画やゲーム等の製作に携わるアーティストにとって，非常に利便性の高いシステムであると言える．

6. 評価

本手法の有効性を確かめるために，実写画像に本手法を適用した画像と，元の実写画像との比較を行った（図 12）．本手法では正面歯画像を入力として要するが，それ以外では実写画像の口内情報は一切用いていないことにも言及しておく．元の実写画像と比較して，本手法適用後の画像は，実写画像と見間違えるほど写実的に口内情報を再現できることが分かる．合成による不自然な境界もなく，歯や舌の位置（動き）も正確に表現可能であることが言える．

7. まとめと今後の課題

発話アニメーション生成の研究において，従来までは口形表現のみに注力する研究が多く，口内表現に大きな課題があった．そこで本研究は，口内表現に着目し，歯の位置の推定と舌の動きの分類を行い，Visio-lization 法を用いて口内情報の自然な埋め込みを実現した．これにより，従来課題であった口内表現を実写と同程度のクオリティで自動生成可能とした．今後の課題には，下唇を噛む音節/v//a/の表現を可能にすること，様々な照明環境への対応が挙げられる．

参考文献

[1] Joshi, P., Tien, W. C., Desbrun, M., AND Pighin, F.: Learning Contrals for Blend Shape Based Realistic Facial Animation, Proc. the 2003 ACM SIGGRAPH/ Eurographics symposium on Computer Animation, pp.187-192 (2003).

[2] Tena, J. R., Torre F. D., AND Matthews, I.: Interactive Region-Based Linear 3D Face Models, Proc. ACM SIGGRAPH 2011, No.76 (2011).

[3] Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., Sumner, R. W., AND Gross, M.: High-Quality Passive Facial Performance Capture using Anchor Frames, Proc. ACM SIGGRAPH 2011, No.75 (2011).

[4] Seol, Y., Lewis, J. P., Seo, J., Choi, B., Anjyo, K., AND Noh, J.: Spacetime Expression Cloning for Blendshapes, ACM Transactions on Graphics (TOG), Vol.32 (Issue 2), No.14 (2012).

[5] Ezzat, T., Geiger, G., AND POGGIO, T.: Train-

図 11 Taylor らの結果（上段）と本手法結果 2 種類（中段・下段）の連番画像比較

図 12 本手法適用前（実写画像）（下段）と適用後の結果（上段）

able Videorealistic Speech Animation, Proc. ACM SIGGRAPH 2002, pp.388-398 (2002).
[6] Chang, Y., AND Ezzat, T.: Transferable Videorealistic Speech Animation, Proc. the 2005 ACM SIGGRAPH/ Eurographics symposium on Computer Animation, pp.143-151 (2005).
[7] Taylor, S. L., Mahler, M., Theobald, B.-J., AND Matthews, I.: Dynamic Units of Visual Speech, Proc. the 2012 ACM SIGGRAPH/ Eurographics symposium on Computer Animation, pp.275-284 (2012).
[8] Li, H., Weise, T., AND Pauly, M.: Example-Based Facial Rigging, Proc. ACM SIGGRAPH 2010, No.32 (2010).
[9] 口腔ケア用品唇開口器ワイダー・チ・ビ，口腔ケア用品唇開口器ワイダー・チ・ビ 口腔ケア 介護用品・車いすの TC マート， 入手先 http://www.tcmart.jp/fs/tcmart/0000000171/widerchibit, (参照 2013-2-1).
[10] 鳥居次好，金子尚道：英語の発音，pp.62-63，92-135，大修館書店（1990）.
[11] 前島謙宣，森島繁生：顔変形モデルと顔形状分布制約に基づく単一顔画像からの 3 次元顔モデル高速自動生成，画像の認識・理解シンポジウム（MIRU2010），IS2-41.pdf（2010）.
[12] 三間大輔，小坂昴大，久保尋之，森島繁生：人の発話特性を考慮したリップシンクアニメーションの生成，”Visual Computing / グラフィクスと CAD 合同シンポジウム 2012, 44.pdf，（2012）.
[13] Irea, A., Takagiwa, M., Moriyama, K., AND Yamashita, T.: Improvements to Facial Contour Detection by Hierarchical Fitting and Regression, The First Asian Conference on Pattern Recognition, pp.273-277 (2011).
[14] 岡田和典：ミーンシフトの原理と応用，情報処理学会研究報告 CVIM， Vol.2008-CVIM-27，pp.401-414 (2008).
[15] Mohammed, U., Prince, S J. D., AND Kautz, J.: Visiolization: generating novel facial images, Proc. ACM SIGGRAPH 2009, No.57 (2009).
[16] Perez, P., Gangnet, M., AND Blake, A.: Poisson Image Editing, Proc. ACM SIGGRAPH 2003, pp.313-318 (2003).