6ZB-7

リアルな口内表現を実現する発話アニメーションの自動生成

川井　正英　　岩尾　知頼　　三間　大輔　　前島　謙宣　　森島　繁生

早稲田大学

1. あらまし

実写ベースの人間の発話シーンを目にする機会が増加している．リアルな発話シーンの作成には，技術力のあるアーティストによる繊細な作り込みが必要となり，多大な労力を要するという問題がある．

この問題を解決するために，Chang らは，Multidimensional Morphable Model を用いて，個人性を反映した発話シーンを生成する手法を提案している[1]．しかしながら，Chang らの手法では口形の動きに合わせて口内の画像をモーフィングしているため，口内の歯や舌の伸縮による破綻が生ずるという問題があった． また，Taylor らは”Viseme”で定義される口形をつなぎ合わせることで，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[2]．しかしながら，”Viseme”は口内に特化した分類ではないため，つなぎ合わせの際，口内は時間的な不連続が生じて，破綻が生ずるという問題があった．さらにこれらの手法には，舌の複雑な動き・歯の細部構造（以下，リアルな口内）の表現に課題があり，改善の余地があると考えられる．

そこで本稿では，口形と口内を分離して考えることで，既作成の発話アニメーションに対して，リアルな口内を自動付加する手法を提案する．入力は，既作成のアニメーション，そのターゲットとなる個人の歯の見える正面顔画像（以下，個人歯画像）1 枚とセンテンス情報である．また，予め任意の人物の舌の動きと通常発話を動画撮影し，連番舌画像データベースと口唇画像データベースを構築しておく．

生成方法は，初めに入力動画中の個人歯画像 1 枚から歯画像データベースを構築し，既作成のアニメーションに対して，開口情報に基づき歯画像を挿入する．次に，センテンス情報に基づきデータベースにある舌画像を挿入する．最後に，歯と舌のつなぎ目の不自然さを解消するために，口唇画像データベースを利用した Visio-lization 法を施す．これにより，リアルな見た目の口内を持つ発話シーンの自動生成が可能となる．

Automatic Generation of Speech Animation Focusing on Realistic Inside of the Mouth
Masahide KAWAI Tomoyori IWAO Daisuke MIMA Akinobu MAEJIMA Sigeo MORISIMA
Waseda University

2. データベースの構築

人間は，歯と舌をある程度独立に動かすことができる．そこで本研究では，開口情報に依存した歯の動きと，音素情報に依存した舌の動きとを別々に分類することとした．

初めに，入力された個人歯画像から，上歯と下歯領域を自動抽出する．抽出された上歯から下歯までの距離（以下，開口歯距離）を単純に広げていき，歯の開閉画像を生成し，それらをまとめて歯画像データベースとする(図 1)． 次に，音素毎の，舌の動きの違いに注目し，A.「舌が前に出る音素 5 種類（/θe/, /te/, /re/, /je/, /e/）」と，B.「舌が前に出た後，舌が見えなくなる音素 5 種類（/θa/, /ta/, /ra/, /ja/, /a/）」と，C.「舌が見えない音素 1 種類（other）」に分類した[3]．そしてそれらの音素を「C, A./te/, C」のセットの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．この組み合わせの 1 セットと同様に，他のセットも撮影し，連番舌画像データベースとした(図 2)．最後に，母音（/a/, /i/, /u/, /e/, /o/）と調音部位の代表的な音素（/θe/, /te/, /re/, /je/, /pa/, /va/）を発話した，7 人分（1 人当たり約 10 秒の発話データ）の動画を撮影し，口唇画像データベースとした．

図 1　歯画像データベース

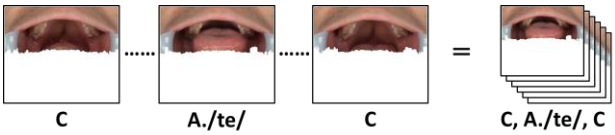

図 2　連番舌画像データベースの一例

3. 開口情報に基づく歯画像挿入

「頭蓋骨の構造上，鼻頂点から上歯の距離，また顎から下歯の距離は一定である」という仮定を立て，開口情報から歯の位置を推定する．

図 3 に示すように，鼻から上歯までの距離は，開口時も閉口時も一定であることが見て取れる．また，顎から下歯までの距離は以下のように考えることができる．閉口時に正面から見たときの顎から下歯までの距離を a，最大開口時に実際の顎から下歯までの距離を a'，最大開口時の上歯と下歯のなす角度を約10°であるとし，a を約

100 [pixel]である（使用した入力動画サイズ：512×512 [pixel]）とすると，2 つの長さの差は式(1)になる．

$$|a-a'|=\left|a-\frac{a}{\cos(10°)}\right|=1[\text{pixel}] \quad (1)$$

入力動画像のサイズと比較してその差は小さい．さらに会話時の開口角度が概ね10°よりも小さいことから，見た目に与える影響は少ない．ゆえに，顎から下歯までの距離を一定とみなすこととした．

この仮定に基づいて，実際に歯画像の挿入を行う．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．開口歯距離が最小となるデータベース中の任意の歯画像 i

$$\arg\min_i \left|d_{I\_f}-d_{D\_i}\right| (0\le i\le N) \quad (2)$$

を選択し，入力画像に歯画像を挿入する．本研究では，データベースの歯画像数を $N=51$ とした．

4. センテンス情報に基づく舌画像挿入

連番舌画像データベースからセンテンス情報に応じて適切な連番画像を選択する．例えば，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，[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[C, A./te/, C][C, A./je/, B./ta/][C, A./e/, C]のセット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．その後，下歯の位置を基準として，入力画像に舌画像を挿入する．なお，連番舌画像の接続部である「C と C」や「B./ta/と C」は，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接続できる．

5. Visio-lization 法による口内の貼り換え

歯画像と舌画像を別々に挿入しているため，歯が舌より手前となり，舌が歯より手前に見えるような様子や，歯と舌の境界にアーティファクトが生じてしまう．この問題を解消するために，Visio-lization 法[4]を導入する．Visio-lization 法は，入力顔画像とデータベース中の顔画像をパッチ単位で比較し，入力の顔をデータベース中の類似するパッチで再構成するというものである．この手法を用いることで，歯と舌の前後関係の変化や不自然なアーティファクトを実写のように自然に表現し直すことが可能となる．本研究では，口周辺部についてのみパッチによる再構成を行う．具体的には，入力画像とデータベース画像をパッチという矩形に区切り，それぞれの RGB 距離を計算し，その距離が最小となるパッチを選択し，タイリングを行う．

あるフレーム f に対して，入力画像のある位置の RBG 値を $\mathbf{C}_{\mathbf{I}\_f}=\{R_{I\_f},G_{I\_f},B_{I\_f}\}$ とし，データベース画像のある位置の RBG 値をそれぞれ $\mathbf{C}_{\mathbf{D\_i}}=\{R_{D\_i},G_{D\_i},B_{D\_i}\}$ 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

$$\arg\min_i \left\|\mathbf{C}_{\mathbf{I\_f}}-\mathbf{C}_{\mathbf{D\_i}}\right\|^2 (0\le i\le N) \quad (3)$$

を選択する．本研究では，データベースの口唇画像数を $N=2213$ とした．なお，パッチサイズは6×6[pixel]，重複部分を3[pixel]とした．また，パッチの重複部分の輝度値を 2 つのパッチの線形ブレンディングにより決定している．

6. まとめと今後の課題

本稿では，口内表現が不十分な既作成のアニメーションに対して開口歯距離とセンテンス情報を用いて，歯と舌の動きの自動付加を行った．その後，Visio-lization 法を施すことで，不自然な部分や違和感を解消し，上歯と下歯で舌を噛むような複雑な表現も可能とした(図 4)．

今後の課題は，下唇を噛むような音素 /va/ などの唇まで考慮した歯の動き表現と，様々な照明環境に対応したパッチタイリング用のデータベースの拡充である．

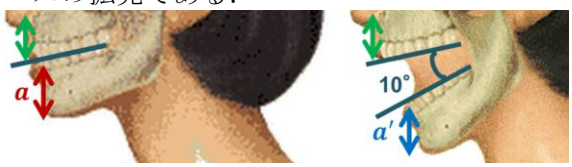

図 3　頭蓋骨の構造


図 4　本手法の結果の 2 フレーム（左：入力，中央：歯・舌挿入結果，右：最終結果）

参考文献

[1]Y.Chang et al, “Transferable Videorealistic Speech Animation”, In Proc. SCA’05, pp.143-151, 2005.
[2]S.Taylor et al, “Dynamic Units of Visual Speech”, In Proc. SCA’12, pp.275-284, 2012.
[3]鳥居次好，金子尚道，“英語の発音”，大修館書店，pp. 62-63,92-135，1990.
[4]U.Mohammed et al, "Visio-lization: Generating Novel Facial Images", SIGGRAPH’09, Article57, 2009.