Vol.2015-CG-161 No.14
Vol.2015-CVIM-199 No.14
2015/11/6

主成分分析に基づく類似口形状検出による ビデオ翻訳動画の生成

古川 翔一[1]　加藤 卓哉[1]　野澤 直樹[1]　サフキン パーベル[1]　森島 繁生[2,3]

概要：映画やテレビ番組などの映像コンテンツでは，吹替処理を施すことによって多言語化を実現している．一般に吹替処理では映像中の話者の口の動きに合わせて翻訳内容が考案される．しかし，これは翻訳家に大きな労力を強いる上，本来の意味と異なる台詞に翻訳される恐れがある．そこで，従来の吹替処理とは逆に，吹替音声に合わせて映像内の口形状を直接編集する「ビデオ翻訳」が提案されている．ビデオ翻訳を行う多くの既存手法では 3D 顔モデルが用いられるが，これらは首などの形状復元できない部位に適用できない問題がある．また結果映像はレンダリングやテクスチャマッピングの精度に依存するため，顔としての自然さに欠ける恐れがある．そこで本研究では自然な結果が得られる画像処理ベースのビデオ翻訳手法を提案する．入力には俳優と声優の 2 つの発話動画を用いる．俳優の動画フレームのうち声優口形状と最も類似するものを主成分分析を用いて選択し，フレーム間の時間連続性を考慮しつつ並び替える．これにより発話に伴う喉の動きまでを表現したビデオ翻訳映像を得ることができ，従来手法よりも自然な出力が可能になる．

1. はじめに

映画やテレビ番組などの多くの映像コンテンツは吹替処理を施すことによって多言語化を実現している．一般に吹替処理では吹替台詞を決めるために以下の 3 つを考慮する必要がある．

(1) 尺合わせ：映像内の音声と吹替台詞の尺を合わせる．
(2) リップシンク：映像内の口の動きと吹替台詞に伴う口の動きを合わせる．
(3) ブレス合わせ：息継ぎのタイミングを映像と吹替台詞とで合わせる．

これらは吹替処理において翻訳家の表現の自由度を制約する要因である．例えば映像中の言語 A を別の言語 B に翻訳する場合を考える．言語 B の文化圏における自然な言い回しとなる理想の翻訳台詞を翻訳家が思い描いたとしても，上記の制約のために翻訳台詞の修正が必要となる．そのため，翻訳家は上記の制約の下で理想に可能な限り近づけた翻訳台詞を考え出している．つまり，現状多くの制約があり，翻訳家は自由な表現ができていない問題がある．一方で，視聴者の立場に注目すると，吹替映像内の口の動きと音声情報が完全に一致しないことによる違和感を感じる問題がある．実際，口の動きと音声情報の一致は発信された内容の理解度を高める上で非常に重要であり，人間が口の動きと音声情報の不一致に対して違和感を感じやすいことが証明されている [1]，[2]，[3]．

上記の問題を解決する方法として「ビデオ翻訳」がある．ビデオ翻訳は映像に合わせて台詞を決める吹替処理とは逆に，台詞に合わせて映像を直接編集するものであり，その目的は以下の 2 点である．

- 吹替処理の制約を取り除き，翻訳家の自由な表現を可能にする．
- 口の動き情報と音声情報を一致させ，コンテンツの内容の理解度を向上させる．

これまでにビデオ翻訳に関する多くの研究が行われてきたが，翻訳家の自由な表現を十分にサポートできていない．そこで，本研究では原音の長さに捉われない翻訳家の自由な翻訳を支援し，可能な限り自然さを保った結果映像を得る手法を提案する．

2. 関連研究

2.1 画像ベースの処理を用いた研究

Bregler ら [4] は画像ベースの処理を施してリアルなスピーチアニメーションを作成する手法を初めて提案した．Bregler らは連続した 3 つの音素を 1 単位とするトライフォ

[1] 早稲田大学
Waseda University
[2] 早稲田大学理工学術院理工学総合研究所
Waseda Research Institute for Science and Engineering
[3] JST CREST

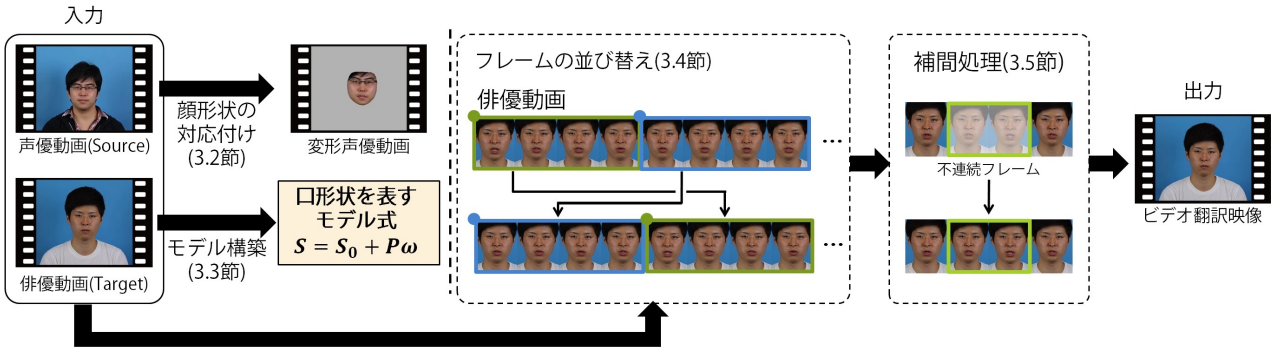

図 1　本手法の概要

ンと口形状画像との対応関係モデルを構築し，新しい音声に合う口の動きをしたスピーチアニメーションを作成した．一方，Ezzat ら [5] は Multidimensional Morphable Model(MMM) を用いて口形状を合成し，頭の動きが自然な映像に張り付けることで，スピーチアニメーションを作成した．しかしながら，これらの手法は口領域を既存の映像に張り付ける際の位置合わせが難しく，高解像度の映像では不自然なぶれが目立つ問題がある．Chang ら [6] は新しい人物の短い発話動画から MMM を修正し，新しい人物に合う MMM を用いてスピーチアニメーションを作成する手法を提案した．しかしながら，短い入力動画で近似したモデルを使うため，長文のスピーチアニメーションの合成の精度が下がる．そのため長文の翻訳には適さず，翻訳家の自由な翻訳を実現できない．

2.2　3D 顔モデルを用いた研究

3D 顔モデルを用いた表情変化転写に関して，多くの研究がなされてきた．Li ら [7] は人物の 3 次元顔形状を取得し，個人特有のブレンドシェイプモデルを構築した上で，任意のキャラクターに表情を転写する手法を提案した．また Cao ら [8] は学習を通して，より正確に個人特有のブレンドシェイプモデルを構築する手法を提案し，高速かつ高精度な表情転写を可能にした．これらと同様の考え方を用いて，Garrido ら [9] は俳優と声優の 2 つの動画に対してブレンドシェイプモデルを適用し，頭の動きに頑健なビデオ翻訳手法を提案した．しかし，この研究は形状復元できない部位に対して適用できず，結果映像において俳優が発話していないときに喉が動く，あるいはその逆が起きる問題がある．そのため，翻訳台詞を原音と同じ長さにする必要があり，翻訳家の自由な翻訳を実現できない．また，最終的な結果映像を得るために 3DCG レンダリングが必要であり，口内には歯のテンプレートを埋め込む必要がある．そのため，結果映像の自然さがレンダリング性能依存であることや，口内に違和感を感じる問題がある．

2.3　本手法のアプローチ

本研究では上記の問題点を踏まえて，翻訳家の自由な表現を実現させるための手法として，フレームの並べ替え処理を施してビデオ翻訳動画を生成する手法を提案する．入力には俳優と声優の 2 つの動画のみを用いる．声優の口形状と類似する口形状を持つものを俳優動画から選択し，動画フレームを並べ替えることによってビデオ翻訳された映像を作成する．入力動画のフレームを局所的な編集を加えずに再利用することで，Ezzat らの合成によるぶれの問題，Chang らの長文翻訳の際に正しい口形状を合成できない問題を解消する．また，顔から首までを 1 枚の画像として扱うことで，Garrido らの問題 (口の動きと喉の動きの不一致，レンダリング性能依存，口内の違和感) を解消し，喉の動きまで表現した結果映像が得られる．本手法の新規性をまとめると以下のようになる．

- 原音の長さ，あるいは翻訳台詞の長さに依存しない自由度の高い翻訳が可能．
- 口と喉の動きの一致，口内といった顔の自然さを保持したビデオ翻訳映像の作成が可能．

これらを達成することで，翻訳家の自由な翻訳を可能にし，視聴者に違和感を与えないビデオ翻訳映像を作成することができる．

3.　提案手法

図 1 に本手法の概要を示す．初めに，声優の顔形状を俳優の顔形状に対応付けることで正規化をする (3.3 節)．次に俳優動画の全フレームの口形状を表す特徴点座標を用いて，主成分分析を施し，口形状を表すモデルを構築する (3.3 節)．モデル構築後，俳優動画と変形済声優動画の 2 つの動画に対して，得られたモデルを用いて口形状のトラッキングを行う．その後，トラッキングした結果を用いて，変形済動画内の声優の口形状と類似するフレーム列を俳優動画から選択して並び替える (3.4 節)．最後に 3.4 節で並び替えたフレーム列の時間連続性を向上させるためにフレーム列同士の繋ぎ目において補間処理を施す (3.5 節)．以下で

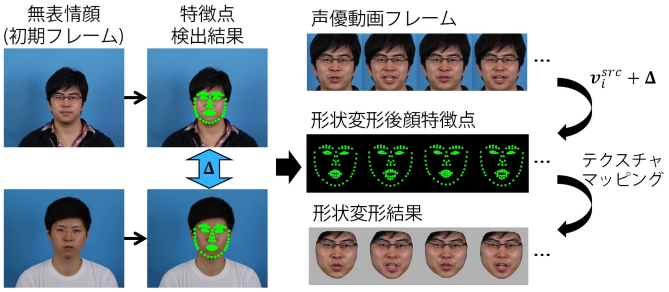

図 2　顔形状の対応付け

は入力に用いる動画の条件と手法の詳細について述べる．

3.1　入力動画

入力には俳優と声優がそれぞれ発話している動画を用いる．俳優動画には，多様な口形状を含むように音素をバランス良く含むセンテンス群を発話させたものを用いる．また，声優動画には俳優に発話させたい任意のセンテンス (翻訳台詞) を発話させたものを用いる．声優，俳優の両者の頭の向きはおよそ正面を向いており，初期フレームは無表情となっているものとする．

3.2　顔形状の対応付け

口形状は個人間で異なるため，異なる口形状同士で類似するものを選ぶことは困難である．そこで，俳優と声優とで類似する口形状を検出する前処理として，声優の顔形状を俳優の顔形状に近づける．図 2 に顔形状の対応付けの概要を示す．初めに，俳優動画と声優動画のそれぞれの初期フレーム画像 (無表情顔画像) に対して顔特徴点を取得する．今回は Irie ら [10] の手法を用いて 86 点の顔特徴点を取得した．続いて，取得した俳優初期フレーム特徴点 $\boldsymbol{v}_0^{tar}$ と声優初期フレーム特徴点 $\boldsymbol{v}_0^{src}$ から顔形状の差として $\boldsymbol{\Delta}=(\boldsymbol{v}_0^{tar}-\boldsymbol{v}_0^{src})$ を定義する．ただし，$\boldsymbol{v}_0$ は顔特徴点座標を列ベクトル表記したものであり，$\boldsymbol{v}_0=(x_1,y_1,\cdots x_N,y_N)^T$ となる．その後，声優動画の全フレームに対して，先ほどと同様にして顔特徴点検出を行い，それぞれのフレームの顔特徴点 $\boldsymbol{v}_i^{src}$(i：フレーム番号) と $\boldsymbol{\Delta}$ との和を取る．$\boldsymbol{\Delta}$ は声優と俳優の顔形状の違いを意味するので，これにより声優の顔形状を俳優の顔形状に変形した特徴点座標 $\boldsymbol{v}_i^{newsrc}$ を得ることができる．

$$\boldsymbol{v}_i^{newsrc}=\boldsymbol{v}_i^{src}+\boldsymbol{\Delta}=\boldsymbol{v}_i^{src}+\boldsymbol{v}_0^{tar}-\boldsymbol{v}_0^{src} \tag{1}$$

最後に，声優動画の各フレーム画像を形状変形後の特徴点座標 $\boldsymbol{v}_i^{newsrc}$ にテクスチャマッピングする．これにより声優の顔形状を俳優の顔形状に近づけた動画 (以下変形声優動画と呼ぶ) を得ることができる．なお，事前に検出した 86 点の顔特徴点座標の情報だけでは十分なテクスチャマッピング結果が得られないため，$\boldsymbol{v}_i^{src}$ と $\boldsymbol{v}_i^{newsrc}$ に対して Radial Basis Function 補間 [11] を施し，661 頂点の平均顔

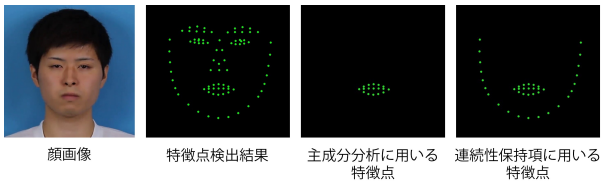

図 3　特徴点座標

モデルを当てはめた上でテクスチャマッピングを施した．

3.3　主成分分析に基づくモデルの構築

俳優動画を用いて口形状を表すモデルを構築する．モデル構築には Cootes ら [12] と同様の考え方を用いる．まず俳優動画の全フレームに対して口形状の特徴点検出を行う．得られた特徴点座標 (図 3 参照) に対して主成分分析を施し，主成分行列を求め，口形状を表すモデル

$$\boldsymbol{S}(\boldsymbol{\omega})=\boldsymbol{S}_0+\boldsymbol{P}\boldsymbol{\omega} \tag{2}$$

を構築する．ただし，それぞれ

$\boldsymbol{S}=(x_1,y_1,\cdots,x_N,y_N)^T$　　：口形状特徴点群
$\boldsymbol{S_0}=(\bar{x}_1^{tar},\bar{y}_1^{tar},\cdots,\bar{x}_N^{tar},\bar{x}_N^{tar})^T$：俳優の平均口形状
$\boldsymbol{P}=(\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_N)$　　：主成分行列
$\boldsymbol{\omega}=(\omega_1,\omega_2,\cdot\omega_N)^T$　　：重みベクトル

である．このモデル構築の段階では平均口形状 $\boldsymbol{S_0}$ と主成分行列 $\boldsymbol{P}$ が求まるため，フレーム画像から特徴点検出により口形状 $\boldsymbol{S}$ を取得することで，線形計算により重みベクトル $\boldsymbol{\omega}$ を求めることができる．上記では次元を削減せず，N 次元のまま主成分行列を用いる．これは以下で重みベクトルの差を用いてコスト関数を定義するため，寄与率の小さい成分の重みの差はコスト関数の値を大きく変化させないためである．

3.4　フレームの並び替え

特徴点検出を行った俳優動画，変形声優動画のそれぞれの全フレームに対して，3.3 節で得たモデルを用いて重みベクトルを計算する．以下では類似する口形状は類似する重みベクトルを持つと仮定し，重みベクトルに基づいた類似口形状検出を行う．図 4 に概要を示す．まず，変形声優動画の t 枚のフレーム列 (図 4 の青色枠のフレーム列) と類似する口形状を俳優動画から選択する．このとき以下のコスト関数 E_j を最小化するフレーム列を，類似する口形状を持つフレーム列として選択する．

$$E_j=\sum_{k=0}^{t}|\omega_k^{src}-\omega_{j+k}^{tar}|^2 \tag{3}$$

ここで j は俳優動画のフレーム列の先頭フレーム番号を表し，$0\leq j\leq$(俳優動画最終フレーム番号 $-t$) の範囲で全探索を行う．すなわち，E_j を最小化する j は最も類似した口形状を持つフレーム列の先頭フレーム番号を意味する．なお，j の最大値を (俳優動画最終フレーム番号 $-t$) と定めた

のは，$j >$(俳優動画最終フレーム番号 $-t$) のときには t 枚のフレーム列を選択することができないためである．以下では，このような一連の類似口形状検索を 1 回の“ステップ”と呼ぶこととする．

続いて，対象フレーム列を時間方向にシフトしたフレーム列 (図 4 の緑色枠のフレーム列) に対して同様の処理を行う．このとき，対象フレーム列は前ステップにおける対象フレーム列と重複するようにシフトさせる．重複を許すことで前ステップで選ばれたフレーム列と滑らかに接続するフレーム列を選択する．ここで，前ステップと同様のコスト関数を用いて，類似口形状を選択すると，フレーム列の接続箇所において不連続性が生じ，結果映像に違和感が生じる．形式的には顔の動きの軌跡が図 5 の左に示すような形を取ってしまう問題が生じる．そこで顔の動きの軌跡を連続的にする (図 5 の右に示す形に近づける) ために式 (3) に連続性保持項を加え，以下のような形にする．

$$E_{i,j} = \alpha \sum_{k=0}^{t} |\omega_{i+k}^{src} - \omega_{j+k}^{tar}|^2 + (1-\alpha)|\boldsymbol{v}_j^{tar} - \boldsymbol{v}_l^{tar}|^2 \quad (4)$$

第 1 項が類似口形状項，第 2 項が連続性保持項に相当し，$0 \leq \alpha \leq 1$ は重みパラメータである．i は変形声優動画における対象フレーム列の先頭フレーム番号である (ここでは i は図 4 における変形声優動画の緑色枠内の先頭フレーム番号に相当する)．l は前ステップにおいて選ばれた俳優動画のフレーム列の最後尾フレーム番号である (ここでは l は図 4 における俳優動画の青色枠内の最後尾フレーム番号に相当する)．また，$\boldsymbol{v}$ は口形状と顔輪郭の特徴点座標 (図 3 参照) を列ベクトル表示させたものである．上記のコスト関数を最小化することで連続性を保持しつつ，類似した口形状を持つフレーム列を選択することができる．

以下同様にして，ステップを繰返すことにより，最終的な並び替え結果を得る．ただし，最終的な並び替え結果は 3.5 において補間処理を施して得られる．なお，コスト関数のそれぞれの項は 0 ～ 1 の値を取るように正規化した上で，各ステップの処理を行う．

3.5 フレーム間の補間

人間は顔などの意味を持つものの動きに対して敏感である．そのため，3.4 節で並び替えによって得られた動画のフレーム列の接続部分においてより連続性を高める必要がある．そこで Saito ら [13] の「Patch Move」を用いて，フレーム列の接続部分を補間する．Patch Move では 2 枚の画像をパッチレベルで対応付けることで，その 2 枚の画像間を補間する．今回はフレーム列の接続部分の 2 枚のフレーム画像をその前後 2 枚のフレーム画像 (図 4 における黄色枠) から補間した画像で置き換えた．

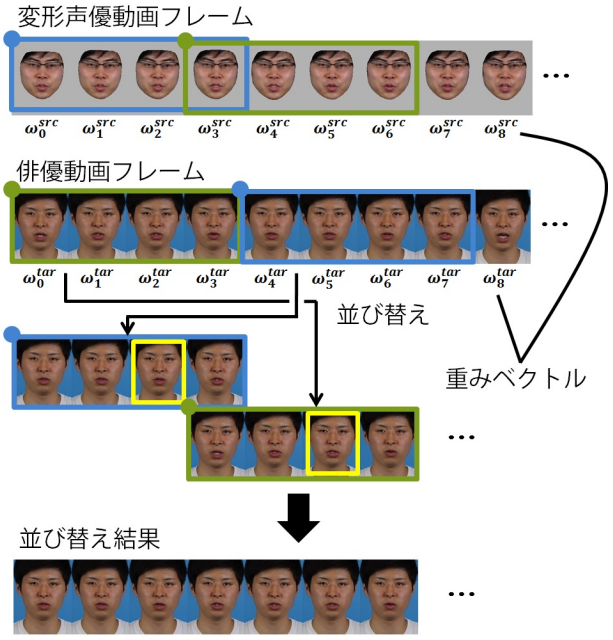

図 4 フレームの並び替え

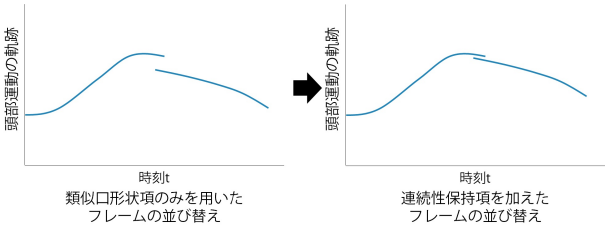

図 5 頭部運動の軌跡

4. 結果と考察

本手法を適用して得られた動画の一部を図 6 に示す．声優動画には俳優に発話させたいセンテンスを発話させたもの (6 秒程度) を，俳優動画には音素をバランス良く含む文として，ATR503 文の 1 セット分を発話させたものを用いた．また，正解動画として，タイミングを合わせて声優と同じセンテンスを俳優に発話させたものを示した．それぞれの動画のビットレートは 30fps，解像度は 1920×1080 のものを用いた．類似口形状検索処理では，一回のステップにおいて対象とするフレーム列の長さを $t = 8$ とし，重複フレーム数 $= 1$ とした．また，コスト関数の重みパラメータ $\alpha = 0.7$ とし，Patch Move による補間処理では 7×7 のパッチサイズを用いた．

図 6 からは声優の口形状と類似する俳優口形状が選択されていることが確認できる．また，正解動画と比べても類似した口形状が選ばれていることが確認できる．本手法ではフレーム間の連続性を向上させるために補間処理を施しているが，動画全体で顔全体あるいは局所部分 (口内等) が不鮮明になることなく，結果動画が作成できていることが確認できた．しかしながら，本手法では二点の問題があ

る．一つ目はフレーム列の接続部分の補間が不十分な点である．これは，人間が人の顔の動きに関して敏感であるため，フレーム単位の補間では違和感が残る場合がある．二つ目は歯の露出具合や舌の見え方といった口内の見た目を考慮できていない点である．そのため，正解動画と結果動画とで口内の見え方 (歯など) が類似していないフレームが選ばれることがある．これは口形状の類似度を量る際に口形状の特徴点情報のみを用いていることが原因である．

5. まとめと今後の課題

本研究では原音の長さに捉われない翻訳家の自由な翻訳を可能とし，かつ視聴者に違和感を与えないビデオ翻訳映像を作成する手法を提案した．目的の達成のために，俳優と声優の 2 者の発話動画を入力として用い，声優の口形状と類似する口形状を直接俳優動画から選択してくることでビデオ翻訳映像を作成した．その結果，鮮明さを保持したまま，音声と口形状が一致した結果映像を得ることができる．今後は口内の類似度を考慮し，フレーム間の連続性向上，背景が動く場合に取り組んでいく予定である．口内の類似度に関しては，コスト関数に見た目の類似度尺度を取り入れる予定である．フレーム間の連続性向上には，結果動画のオプティカルフロー等を取得し，位置合わせをするという対応がある．また，背景が動く場合には，人物と背景部分とで領域分割を行った上で本手法を適用し，インペインティング等を施しながら背景領域に合成するという処置がある．さらに，より翻訳家の自由な表現を支援するために今後は動画編集のシステムを導入する予定である．具体的には，翻訳家が翻訳台詞の一部のみを編集したいときにすぐに適切な口形状が選ばれるといったインタフェースの導入などを考えている．

参考文献

[1] Sumby, W. H. and Pollack, I.: Visual contribution to speech intelligibility in noise, *The journal of the acoustical society of america*, Vol. 26, No. 2, pp. 212–215 (1954).
[2] Owens, E. and Blazek, B.: Visemes observed by hearing-impaired and normal-hearing adult viewers, *Journal of Speech, Language, and Hearing Research*, Vol. 28, No. 3, pp. 381–393 (1985).
[3] Summerfield, Q.: Lipreading and audio-visual speech perception, *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 335, No. 1273, pp. 71–78 (1992).
[4] Bregler, C., Covell, M. and Slaney, M.: Video rewrite: Driving visual speech with audio, *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co., pp. 353–360 (1997).
[5] Ezzat, T., Geiger, G. and Poggio, T.: *Trainable videorealistic speech animation*, Vol. 21, No. 3, ACM (2002).
[6] Chang, Y.-J. and Ezzat, T.: Transferable videorealistic speech animation, *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, ACM, pp. 143–151 (2005).
[7] Weise, T., Li, H., Van Gool, L. and Pauly, M.: Face/off: Live facial puppetry, *Proceedings of the 2009 ACM SIGGRAPH/eurographics symposium on computer animation*, ACM, pp. 7–16 (2009).
[8] Cao, C., Hou, Q. and Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation, *ACM Transactions on Graphics (TOG)*, Vol. 33, No. 4, p. 43 (2014).
[9] Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Perez, P. and Theobalt, C.: VDub: Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track, *Computer Graphics Forum*, Wiley-Blackwell (2015).
[10] Irie, A., Takagiwa, M., Moriyama, K. and Yamashita, T.: Improvements to facial contour detection by hierarchical fitting and regression, *Pattern Recognition (ACPR), 2011 First Asian Conference on*, IEEE, pp. 273–277 (2011).
[11] Buhmann, M. D.: *Radial basis functions: theory and implementations*, Vol. 12, Cambridge university press (2003).
[12] Cootes, T. F., Taylor, C. J., Cooper, D. H. and Graham, J.: Active shape models-their training and application, *Computer vision and image understanding*, Vol. 61, No. 1, pp. 38–59 (1995).
[13] Saito, S., Sakamoto, R. and Morishima, S.: PatchMove: Patch-based Fast Image Interpolation with Greedy Bidirectional Correspondence (2014).

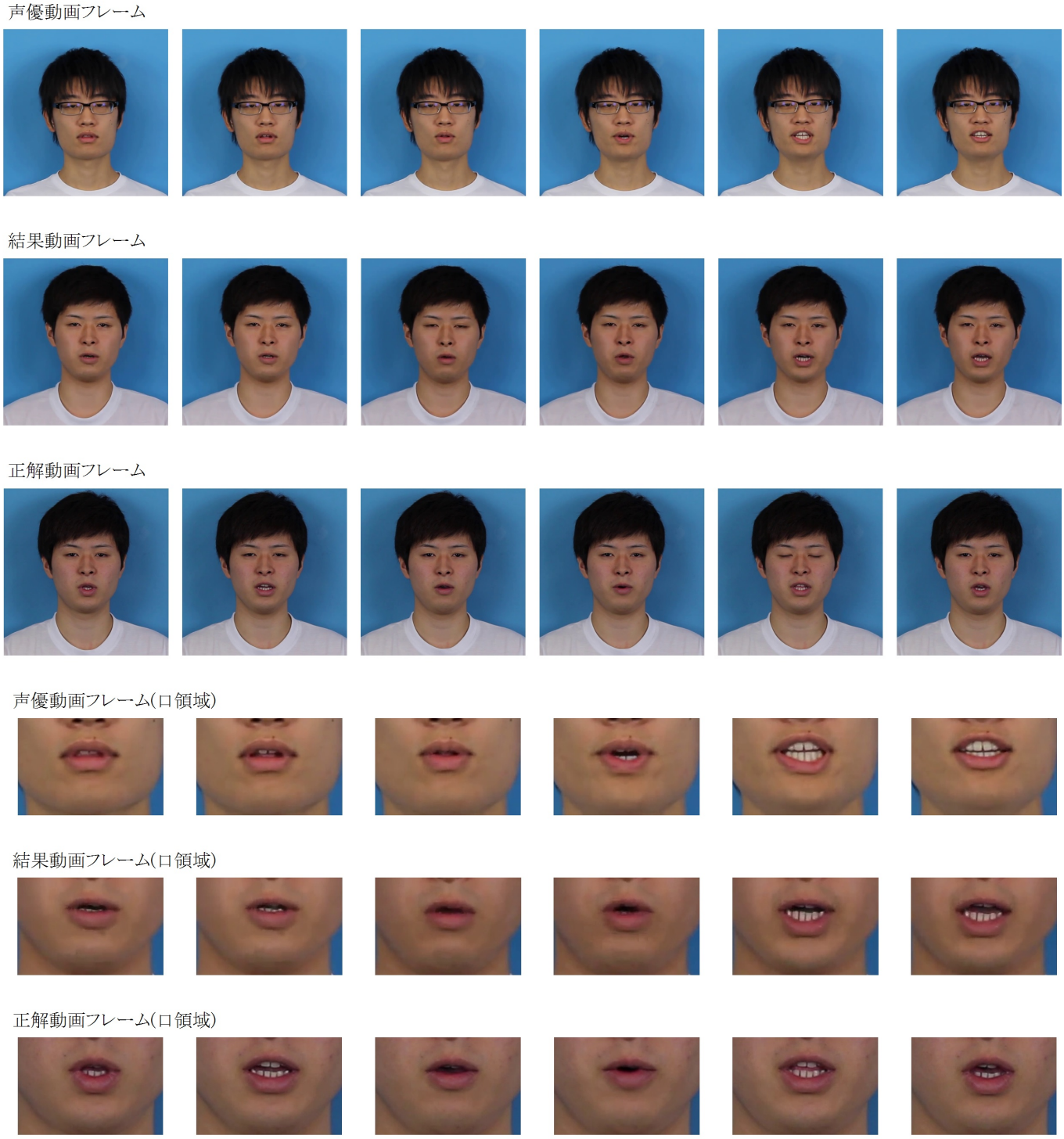

図 6 合成結果