7W-03

フレームリシャッフリングに基づく事前知識を用いない吹替映像の生成

古川　翔一　　加藤　卓哉　　野澤　直樹　　サフキン　パーベル　　森島　繁生 †

早稲田大学　　早稲田大学理工学術院理工総合研究所/JST CREST†

1　はじめに

近年，コンテンツの多言語化に伴い，吹替処理の需要が高まっている．しかし，多くの吹替映像において口の動きと吹替音声は一致していない．口の動きと音声の不一致は視聴者に違和感を与えるだけでなく，内容の理解度を低下させてしまう．そのため，口の動きと吹替音声が一致した吹替映像が求められる．

Ezzat ら [1] は音素情報と合う口画像を生成し，それを既存の映像に合成することで口の動きと音声が一致する映像を生成した．しかし，Ezzat ら [1] の手法では 3 点問題がある．1 つ目は，口画像を複数画像の線形和で生成するため，結果映像が不鮮明になることである．2 つ目は，調音結合という，ある音を発する際の口の形が次の音に依存する現象を考慮していないことである．そのため，一つの音に対して単一の口形状しか生成できず，不自然な結果になってしまう．3 つ目は，言語ごとに異なる音素情報を必要とすることである．

そこで本研究では，声優と同じタイミングで同じ口形状が現れるように吹替対象の動画のフレーム列を並び替えること (フレームリシャッフリング) によって吹替映像を作成する．これにより，鮮明でかつ調音結合を考慮した結果を生成する．本手法は音素情報を用いないため，あらゆる言語に適用可能である．また，声優の様子を撮影した動画 (以下，声優動画) と吹替対象となる動画 (以下，俳優動画) の 2 つのみを入力として用い，データベースなどの事前知識を必要としない．ただし，本手法はニュース，講義映像などを対象とし，声優，俳優共に頭の向きはおよそ正面であり，初期フレームは無表情であるものとする．

2　口形状の正規化

口形状は個人ごとで異なるため，異なる人物間で口形状の類似度を計算することは困難である．そこで，事前に俳優と声優の口形状を近づける操作を行う．まず，俳優動画と声優動画のそれぞれに対して Irie ら [2] の手法を用いて点数 $N = 22$ の口特徴点を検出し，その座標 $\boldsymbol{S}_i^{tar}$ と $\boldsymbol{S}_i^{src}$ を得る (i：フレーム番号，tar：俳優動画，src：声優動画)．その後，式 (1) のように，初期フレームの特徴点座標の差分 $\boldsymbol{\Delta} = (\boldsymbol{S}_0^{tar} - \boldsymbol{S}_0^{src})$ を $\boldsymbol{S}_i^{src}$ と足し合わせることで，声優の口形状を俳優の口形状に近づけた特徴点座標 $\boldsymbol{S}_i^{src\prime}$ を得る．

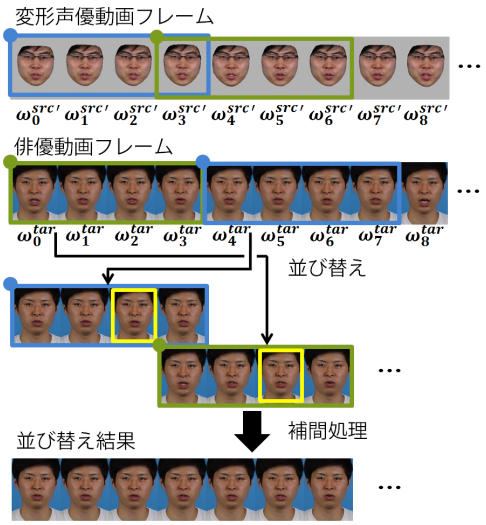

図 1: フレームリシャッフリングと補間処理

$$\boldsymbol{S}_i^{src\prime} = \boldsymbol{S}_i^{src} + \boldsymbol{\Delta} \tag{1}$$

3　フレームリシャッフリング

まず，口形状を表すモデル式を構築する．俳優動画の口形状特徴点座標に対して主成分分析を施し，口形状に寄与する成分から成る主成分行列を求める．求めた主成分行列を用いて以下の式を得る．

$$\boldsymbol{S}(\boldsymbol{\omega}) = \boldsymbol{S}_{\boldsymbol{0}}^{\boldsymbol{tar}} + \boldsymbol{P}\boldsymbol{\omega} \tag{2}$$

ここで，それぞれ

$\boldsymbol{S} = (x_1, y_1, \cdots, x_N, y_N)^T$　　：口形状特徴点群

$\boldsymbol{S}_{\boldsymbol{0}}^{\boldsymbol{tar}} = (\bar{x}_1^{tar}, \bar{y}_1^{tar}, \cdots, \bar{x}_N^{tar}, \bar{x}_N^{tar})^T$：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

$\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_N)^T$：主成分に対応する重みベクトル

である．

次に，式 (2) を用いて，$\boldsymbol{S}_i^{src\prime}$，$S_i^{tar}$ からそれぞれ対応する重みベクトル $\boldsymbol{\omega}_i^{src\prime}$，$\boldsymbol{\omega}_i^{tar}$ を計算する．ここで，類似する口形状は重みベクトルも類似すると仮定する．この仮定に基づけば，声優動画と同じタイミングで同じ

”Dubbing Video Generation based on Frame-Reshuffling without Prior Knowledge”

Shoichi Furukawa　Takuya Kato　Naoki Nozawa　Savkin Pavel　Shigeo Morishima†

Waseda University

Waseda Research Institute for Science and Engineering/JST CREST†

図 2: 結果動画 (上段) と正解動画 (下段) との比較

重みベクトルが現れるように俳優動画のフレーム列を並び替えることで，声優と同じ口の動きを実現できる．図 1 に概要を示す．フレーム列の並び替えの際には式 (3) を最小化する．

$$E_{i,j} = \begin{cases} \sum_{k=0}^{t} |\boldsymbol{\omega}_k^{src\prime} - \boldsymbol{\omega}_{j+k}^{tar}|^2 & (i = 0) \\ \alpha \sum_{k=0}^{t} |\boldsymbol{\omega}_{i+k}^{src\prime} - \boldsymbol{\omega}_{j+k}^{tar}|^2 + (1-\alpha)|\boldsymbol{v}_j^{tar} - \boldsymbol{v}_l^{tar}|^2 & (i > 0) \end{cases} \quad (3)$$

ここで，i は声優動画注目フレーム列の先頭フレーム番号であり，図 1 に示すように数フレーム重なるように更新していく．j は俳優動画フレーム列の先頭フレーム番号，t は一度に並べ替えるフレーム列の長さである．$i > 0$ のときの式 (3) 第 2 項は頭部位置の連続性が高いフレーム列を選ぶためのものである．$\boldsymbol{v}_j^{tar}$ は Irie ら [2] の手法を用いて取得した，俳優動画 j 番目のフレームの顔輪郭特徴点 (21 点) 座標をベクトル表示したものである．l は前ステップで選ばれたフレーム列の最後尾フレーム番号である．$0 \leq \alpha \leq 1$ は重みパラメータである．なお式 (3) の各項は最大値 1，最小値 0 となるように正規化した上で処理を行う．

最後に並べ替えたフレーム列を接続する．このときフレーム列の接続部分で頭部位置の連続性をさらに高めるために補間処理を行う．本手法では Saito ら [3] の手法を用い，フレーム列の接続部分の前後 2 枚 (図 1 における黄色枠) からその間の 2 枚を補間する．全ての接続部分で同様の処理を行い，最終的な生成結果を得る．

4 結果と考察

図 2 に結果動画と正解動画とを比較したものを示す．俳優動画には音素バランス文 1 セット分を発話させたもの (29.97fps，6 分程度) を，声優動画には任意の文を発話させたもの (29.97fps，6 秒程度) を用いた．また，式 (3) において $\alpha = 0.7$，$t = 8$ とし，i の更新は 1 フレーム重なるよう行った．正解動画は声優と同じ文を俳優にタイミングを合わせて発話させたものである．図 2 から，正解動画と同じ口形状が同じタイミングで現れていることが確認できる．また，結果映像の鮮明さも保持されていることが確認できる．

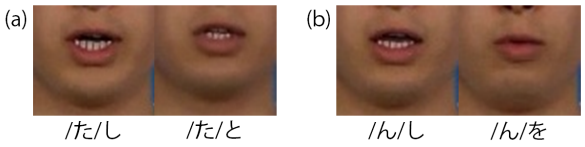

図 3: 調音結合の検証結果

図 3 に調音結合の検証結果を示す．図 3 の (a) は結果映像の中で，「たし」，「たと」と発話している部分の「た」のときの口画像，(b) は「んし」，「んを」と発話している部分の「ん」のときの口画像である．図 3 からは口形状が次の音によって異なることが確認でき，調音結合が考慮できていると言える．

5 まとめと今後の課題

本研究では音素情報を用いずに，声優動画と俳優動画から口の動きと音声とが一致する吹替映像を生成する手法を提案した．本手法により鮮明でかつ調音結合を考慮した結果が生成できた．しかし，結果映像がフレーム列長に依存するという問題がある．今後は，自動で最適なフレーム列長を選択できるよう改善する予定である．

謝辞

本研究の一部は，JST CREST「OngaCREST プロジェクト」の支援を受けた．

参考文献

[1] Ezzat, T., Geiger, G., and Poggio, T.: *Trainable videorealistic speech animation*, ACM TOG, Vol. 21, No. 3, 2002.

[2] Irie, A., Takagiwa, M., Moriyama, K. and Yamashita, T.: *Improvements to Facial Contour Detection by Hierarchical Fitting and Regression*, IEEE, pp.273-277, 2011.

[3] Saito, S., Sakamoto, R. and Morishima, S.: *Patchmove: Patch-based fast image interpolation with greedy bidirectional correspondence*, Pacific Graphics, 2014.