Vol.2019-CG-173 No.8
2019/3/5

一枚画像と音情報を用いた動画生成

土屋 志高[1,a)]　板摺 貴大[1]　夏目 亮太[1]　加藤 卓哉[1]　山本 晋太郎[1]　森島 繁生[2,b)]

概要：近年，人の話し声や楽器演奏のように，音と連動した動作を音情報から再現する研究が行われている．従来手法では，顔の特徴点や体のボーンのような対象に特化した特徴量を用いることで，口や体の動きを生成しているが，音と動きが連動している任意の現象に対しては適用できない．本稿では，一枚画像と数秒の音を入力とすることで，対象に依存しない画像の見た目を保持したまま，音に連動した動画を生成する深層学習を用いた手法を提案する．実験において，口や体の動きだけでなく，花火や海の波などの様々な動画において提案手法が有効であるかの検証を行い，対象ごとに特徴量を定めることなく動画生成が可能であることを確認した．

1. はじめに

人間は，画像とそれに対応する音を聞いたときに，その画像が音に合わせてどのように変化をするのかを想像することができる．画像と音が対応している例として，人の話し声や楽器演奏のような人の動作，海の波や雷といった自然現象などが挙げられる．これらのように，視覚と聴覚の情報が対応している現象が日常的に多く存在している．

近年，画像と音から動画を再現する人間の能力をコンピュータで実現する研究が行われている．そのような研究の例として，音を入力とすることで，その音に合わせて対象の状態を予測する研究が挙げられる．特に，顔の特徴点や体のボーンといった特徴量を用いることで，口や体の動きを生成する研究がある [11], [13]．しかし，これらの手法は対象に特化した特徴量を用いているため，適用する対象によって特徴量を定める必要があり，同じネットワークを任意の対象に対して適用できない問題点がある．

本稿では，動かしたい対象の画像と音から，音と動きの対応が取れた動画の生成を行う手法を提案する．本稿では，音と動きの対応として，「空間対応」「時間対応」「種類対応」を考える．それぞれ「空間対応」は音に対応する領域が動いているか，「時間対応」は音と動きのタイミングが合っているか，「種類対応」は音に対する動きの種類が合っているかである．提案手法では，それぞれの対象に特化した特徴量を用いず画像と音を入力とすることで，任意の対象に対して同じネットワークを適用し，動画の生成が可能となる．実験では，人の手を叩く動作，簡単な発音をする人の口の動き，花火が開く様子，海の波，これら 4 つの対象を別々に学習を行うことで，異なる対象に同じネットワークを適用できるかの検証を行った．実験の結果，同じネットワークで，人の手を叩く動作，人の口の動き，花火に関して，音に対応した動画を生成できることが確認できた．従来手法では，自然現象に対しての動画生成はできないのに対して，提案手法では，見た目を保持した動画を生成できることが確認できた．

2. 関連研究

近年，画像と音を複合的に利用した研究が行われている．楽器演奏のデータセットを用いて，画像から音，音から画像を生成する手法 [4] や唇の画像と音声から人が話している動画を生成する手法 [3] などが提案されている．これらの手法では，楽器演奏や音声などの特定の対象で実験を行なっている．本稿では，音から動画を生成する手法を提案するが，動画に音声を付与する研究として，動画に合う音の波形を生成する手法 [15] や無音動画に対して音のデータベースから最も近い効果音をつける手法 [9] などが提案されている．Tulyakov ら [14] は画像の見た目の情報と動きの情報を分離する手法を提案した．画像に対して動きの条件を加えることで，与えた動きに合う動画を生成できる．本稿では，音によって動きを制御する手法を提案する．

従来研究では，特徴点を利用した動画生成の手法が提案されている．Suwajanakorn ら [13] は，RNN を用いて音声から口形を予測し，各フレームの口形に合う顔のテクスチャ画像を合成することで，音声に合ったリアルな顔の動画を生成する手法を提案した．Shlizerman ら [11] は，ピア

1 早稲田大学
2 早稲田大学理工学術院総合研究所
a) y-tsuchiya0920@asagi.waseda.jp
b) shigeo@waseda.jp

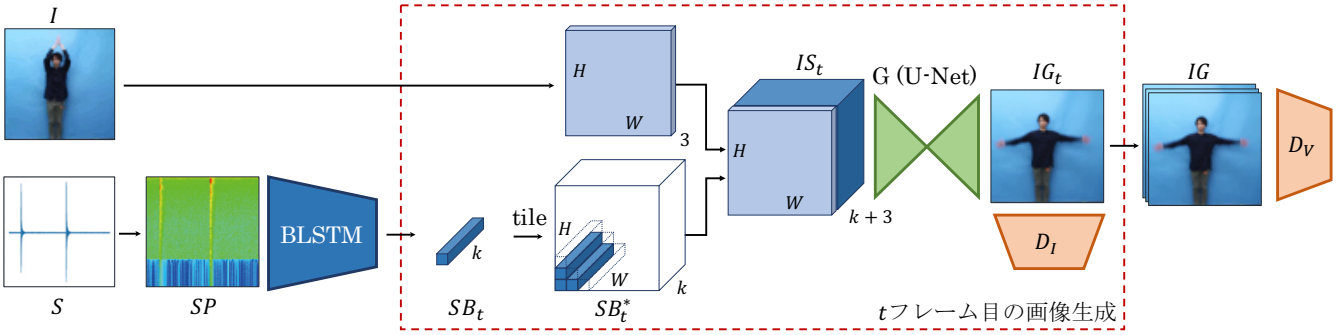

図 1 提案ネットワーク

ノやヴァイオリンなどの楽器演奏から LSTM により人間の腕や指などのボーンを予測する手法を提案した．これらの手法では，それぞれの対象で口の特徴点や人間のボーンの情報が必要となるため，特徴点が存在しない音と動きが連動した現象に対しては適用できない．本稿では特徴点が存在しない対象に対して適用可能な手法を提案する．

3. 提案手法

数秒の音から得られる音の時間的変化の情報と，一枚の画像から得られる見た目の情報を入力として，GAN [5] を用いて動画を生成する．提案手法のネットワークを図 1 に示す．まず，Bidirectional LSTM (BLSTM) [6] により音の時間的な変化の特徴を抽出する．次に，得られた特徴量と画像から GAN の Generator (G) により画像を生成する．生成された画像を GAN の Discriminator（D_I）に入力し，自然な画像であるかを判別する．最後に，生成された連続画像をまとめて GAN の Discriminator (D_V) に入力し，時間的に自然な画像であるかを判別する．

3.1 スペクトログラムの作成

動画は，複数枚の画像と音の波形から構成されている．それぞれのサンプリングレートは異なり，音のタイミングに合う動画を生成するためには，画像と音の情報を対応付けることが必要である．画像のサンプリングレートの方が音よりも小さいことから，画像のサンプリングレートに合わせて音の波形をフーリエ変換することで，画像ごとに音の情報を対応付ける．音のサンプリングレートが F_s [Hz]，画像のサンプリングレートが F_i [fps] で構成されている動画に対しては，F_i/F_s 秒ごとにフーリエ変換を行うことで，画像ごとにスペクトログラム SP を得ることができる．

3.2 音の時間的変化の特徴抽出

得られたスペクトログラムは無音の区間が含まれており，無音の区間におけるスペクトログラムは動きを予測するための必要な情報が含まれていないため，そのまま音の特徴量として用いるのは適切ではない．したがって，スペクトログラムの時間的な変化を抽出することで，無音の時間から音が鳴る時間までの動きの情報を取得することを考える．

ここでは，BLSTM を用いて，得られたスペクトログラムから音の時間的変化の特徴を抽出する．BLSTM は過去から未来の方向と，未来から過去の方向の二つの時間方向で学習を行う．本稿においては，入力音全体があらかじめ与えられているため，BLSTM を適用できる．3.1 節で得られた各フレーム 735 次元のスペクトログラムを BLSTM に入力することで各フレーム $t \in [1, T]$ (T は生成動画のフレーム数) に対して k 次元の特徴量 SB_t を得る．

3.3 画像と音からの画像生成

図 1 に示したように，入力画像 I と音の時間的変化の特徴量 SB_t より，t フレーム目に対応する画像を生成し，合計 T フレームの画像を生成する．3.2 節で得られた t フレーム目の SB_t を入力画像 I のサイズ $H \times W$ に合わせて複製し，$k \times H \times W$ のテンソル SB^*_t に変換する．次に，入力画像 I と音の時間的変化の特徴 SB^*_t を結合することで，$(3+k) \times H \times W$ のテンソル IS_t を得る．IS_t を G に入力することで，t フレーム目の画像 IG_t を生成する．本稿では，入力画像の見た目を考慮した画像を生成するために，G として U-Net [10] の構造を用いた．

3.4 学習

自然な動画を生成するためには，各フレームごとに画像としての自然さと，連続したフレームの動画としての自然さが重要になる．そこで，本手法では次の三つの損失関数を定義する．一つ目は，生成画像と正解画像での再構成誤差 $\mathcal{L}_{L1}$ である．二つ目は，生成画像がより正解画像のような自然な画像になるような敵対的損失 $\mathcal{L}_{G_I}$ である．三つ目は，生成画像と正解画像の連続性が自然であるかを学習するための時系列的な敵対的損失 $\mathcal{L}_{G_V}$ である．これは，生成された連続した画像を τ フレームごとにまとめて $3\tau \times H \times W$ と変形した後に敵対的損失をとる．以上の三つの損失関数の線形和を G における損失関数 $\mathcal{L}$ とする．

$$\mathcal{L} = \alpha\mathcal{L}_{L1} + \beta\mathcal{L}_{G_I} + \gamma\mathcal{L}_{G_V} \quad (1)$$

α，β，γ はそれぞれ損失関数の重みのパラメータを表す．

4. 実験

本稿では，人の手，人の口，海，花火の 4 種類の動画を用いて実験を行なった．人の手は，手を上下運動させる間に頭上で手を叩く動作である．これは，手を叩いた時のパルス音のような音に対して，本手法が適用できるかを確認する．また，異なる複数の音に対して動きの変化を再現できるかの確認を行うため，人の口の動画を用いる．無作為に ”あ”, ”い”, ”う”, ”え”, ”お” を発音した動画である．花火は，打ち上げ花火の動画を用いることで，音の鳴る瞬間と音が鳴った後の様子を生成できるかの確認する．海は，波打ち際の波が打ち寄せる様子の動画である．これは，入力画像に含まれていない白波が出たり消えたりする様子を生成できるかを確認する．今回の実験では，人の手，人の口，花火，海の動画は別々に学習を行った．

4.1 データベース

既存の動画のデータセット [2], [7], [12] では音に関係のない動作が含まれているため，本手法の有用性を確認するのには適していない.，また，顔の表情のデータセット [1] では，顔の表情の変化に対応する音声データが含まれていないため，提案手法を検証するのには適していない．本稿では提案ネットワークの有用性を確認するために，音と動きの対応が取れている動画のデータベースを構築する．データの条件としては以下の三つが挙げられる．

- 音と画像が対応している
- 画像に関係のない音が入っていない
- カメラが固定されている

カメラが固定されていることで，音には関係のない動きを排除することができる．人の手，人の口の動画は独自に撮影したもの，花火，海は YouTube に公開されている動画を使用した．学習に用いたそれぞれの動画の本数と総秒数を表 1 に示す．

4.2 実験設定

学習には，音のサンプリングレートが $F_s = 44100$ [Hz]，画像のサンプリングレートが $F_i = 30$ [fps] である長さが 4 秒の動画を用いた．学習時の各パラメータは $k = 32$，$T = 120$，$\tau = 5$ とした．損失関数の最適化には，Adam [8] を用いて，$lr = 0.0002$，$\beta_1 = 0.5$，$\beta_2 = 0.999$ とした．その際に，損失関数の重みはそれぞれ，$\alpha = 100$，$\beta = 1$，$\gamma = 1$ とした．入力画像は 64×64 の画像を用いて，64×64 の画像を 120 枚生成し，30 [fps] の動画を生成した．

表 1 各対象の動画の本数と総秒数

	人の手	人の口	花火	海
動画の本数	30	6	3	6
総秒数 [s]	944	408	2944	10560

5. 結果

学習時には含まれていない画像と音を入力とした時の人の手，人の口，花火，海の動画生成の結果を図 2 から図 5 に示す．各対象に対して，入力音を固定して入力画像を変化させた場合と，入力画像を固定して入力音を変化させた時の比較を行う．各図において入力音の横軸はフレーム数，縦軸は振幅を表す．本稿では，振幅値を $[-1, +1]$ に正規化した．緑枠，赤枠，橙枠はそれぞれ入力画像，元の動画，生成動画を表している．

5.1 人の手と口

図 2 より，入力音 hand-X に対して，入力画像を hand-A，hand-B とした時に，入力画像の服装や手の位置に依存せず，hand-X の振幅が大きく変化している 1，60，120 フレーム付近で手を叩いている画像が生成されていることが分かる．音が鳴っている手を叩く瞬間だけでなく，打撃音が鳴っていない区間であっても，手の上下運動の様子を生成できていることが分かる．また，入力音 hand-Y とした時に，60 から 90 フレームにかけて連続で音が鳴っている場合では，手を下に下げることなく，連続して手を叩いている動画が生成されている．

図 3 より，入力音 mouth-X に対して，入力画像を mouth-A，mouth-B とした時に，入力画像の口の状態に関わらず，mouth-X の元の動画と同じ口の形を生成できていることが分かる．また，入力音 mouth-Y とした時に，mouth-X の時とは異なるタイミングで口を動かす動画を生成できていることが分かる．

5.2 花火と海

図 4 より，入力音 fireworks-X に対して，入力画像を fireworks-A，fireworks-B とした時に，入力画像の花火の位置を反映して動画を生成できていることが分かる．また，40 フレームまでは明るい花火の動画が生成され，それ以降は元の動画のように花火が暗くなる様子が再現できている．入力音 fireworks-Y の時は，fireworks-X の時と比較して振幅の変化が大きく，120 フレームまで明るい花火の動画が生成されていることが分かる．

図 5 より，入力音 sea-X に対して入力画像を sea-A，sea-B とした時に，それぞれの入力画像の波際に対して，平行になるように白波が生成されていることが分かる．入力音 sea-Y とした時は 1，60 から 90，120 フレームで類似した場所に白波が生成されていることが分かる．

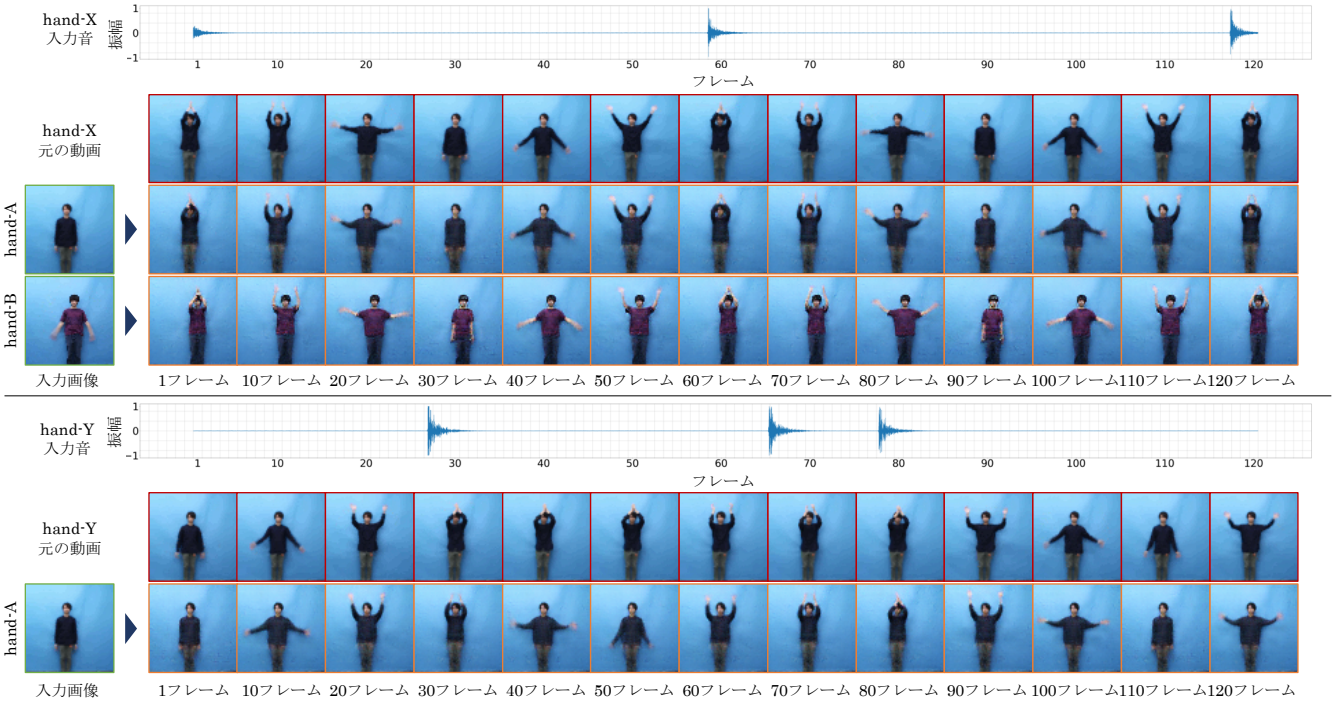

図 2 入力画像と入力音を変化させた時の元の動画と生成動画の比較（人の手）

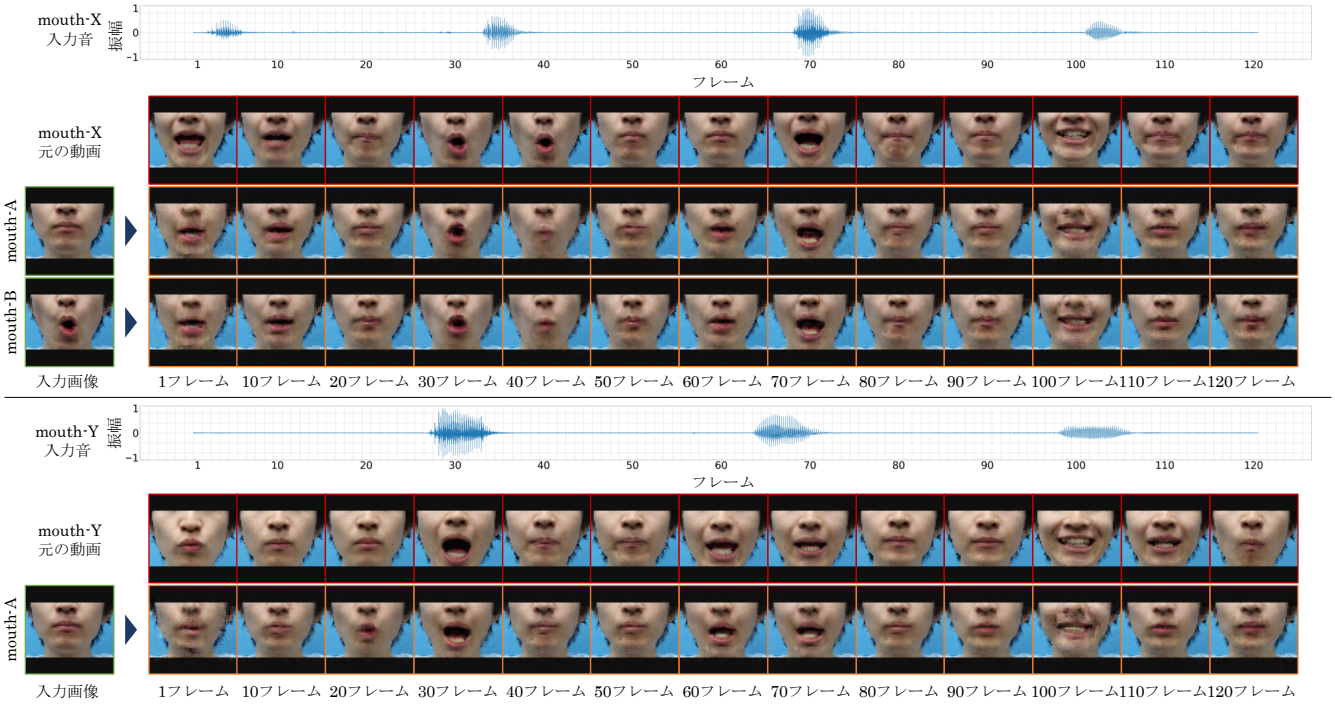

図 3 入力画像と入力音を変化させた時の元の動画と生成動画の比較（人の口）

6. 考察

6.1 人の手と口

撮影した動画を用いて学習を行った人の手，人の口について考察する．まずは，人の手の結果について図 2 より，手を叩く瞬間だけでなく，打撃音が鳴っていない区間でも手の上下運動を生成できていることが分かる．これは，BLSTM により入力音から時間的な変化の特徴量を抽出できたからである．また，入力音 hand-Y のように連続して音が鳴る場合では，手の上下運動をするのではなく，手を上に挙げたまま連続して手を叩く様子が再現できた．このことから，BLSTM は単純な周期的な動きを学習するのではなく，入力音から予測される物理的に自然な動きの時間的変化を学習していると考えられる．異なる入力画像の時

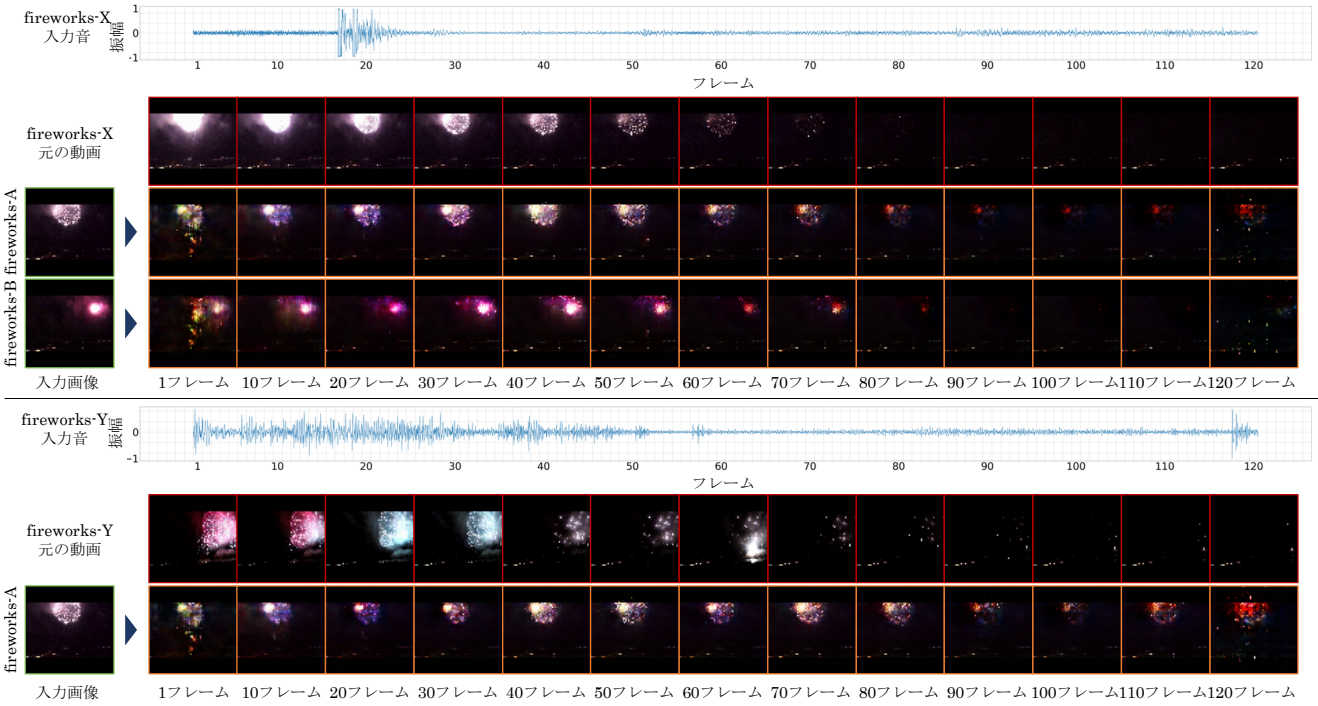

図 4 入力画像と入力音を変化させた時の元の動画と生成動画の比較（花火）

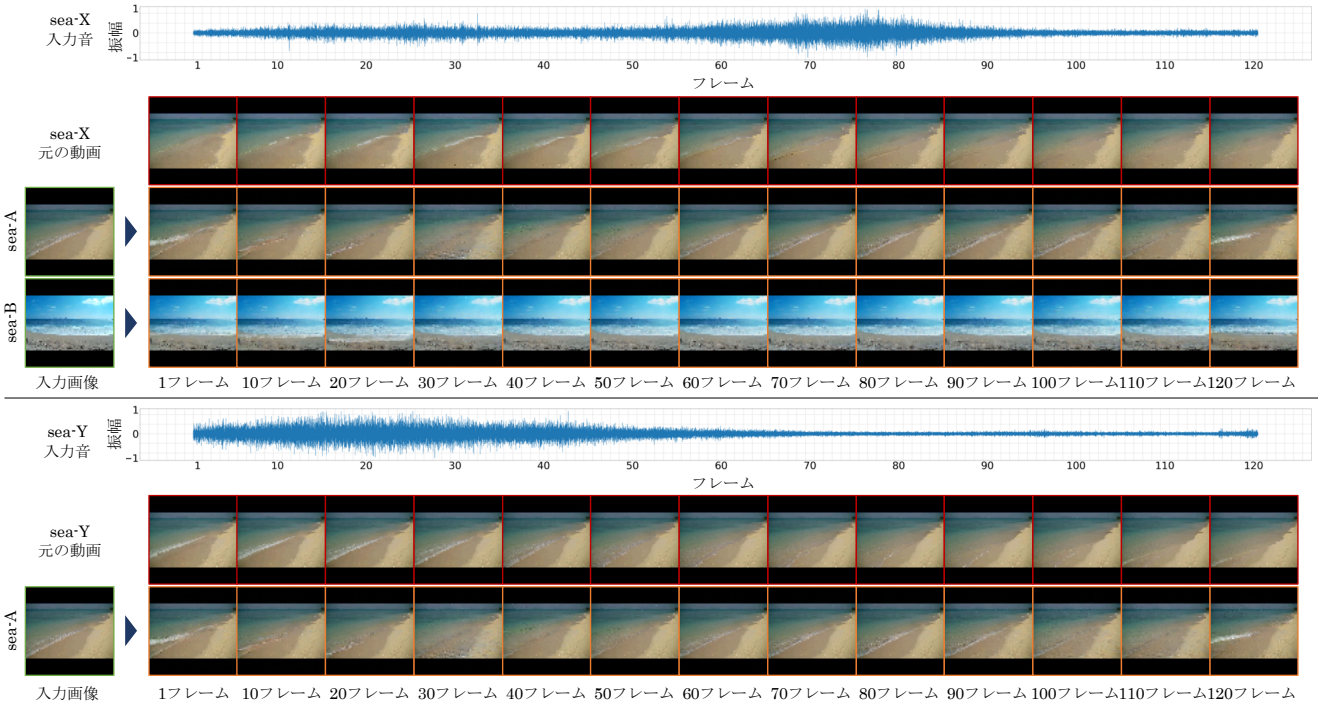

図 5 入力画像と入力音を変化させた時の元の動画と生成動画の比較（海）

に，元の動画の姿勢と同じような姿勢を生成できていることから，見た目を反映した動画生成ができていると言える．

次に人の口の結果について図 3 より，入力音の振幅が大きくなるフレームで口を開けている画像が生成できている．生成動画と元の動画を比較すると，口を開けるタイミングだけでなく，どの発音がされているかも表現できていることが分かる．人の音声は，手の打撃音とは異なり，”あ”，”い”，”う”，”え”，”お”の 5 種類の音があるため，BLSTM によりタイミングだけでなく，音の種類に合わせてどの口形を生成するかも学習されたと考えらえる．

人が手を叩くという単純な動きでは生成画像は自然であったが，入力画像からの変化が大きい口の動画では，口の部分の生成画像の自然さが保たれていないものがあった．これの一つの要因として考えられるのは，用意した

データセットのデータ数が十分でないことである．手を叩く動画は一人の人物が 3 種類の服装をして撮影した 30 本 (944 秒) の動画，人の口は一人の人物が発音の順番を変えた 6 本 (408 秒) の動画を用意した．手を叩く動画では，見た目のバリエーションが 3 種類あるのに対して，人の口は 1 種類しかないため，見た目の特徴を抽出する学習が十分ではなかったと考えられる．

6.2 花火と海

次に YouTube に公開されている動画を用いて学習を行った花火，海について考察する．まずは，花火の結果について図 4 より，異なる入力画像に対して，同じ花火の音を入力とした時に，入力画像の花火の場所を保持した動画を生成できていることが分かる．また，音に対応している花火の部分は光るが，それ以外の部分は暗いままであることから，画像のどの部分が音に対応する対象であるかを学習できていると言える．入力音 fireworks-X の場合，花火の光が 40 フレームまでは明るく，その後暗くなるという光り方が共通していることから，BLSTM によって入力音の時間的な変化の特徴量を抽出できたと考えられる．

次に海の結果について図 5 より，同じ入力音 sea-X に対して異なる海の画像を入力とすると，それぞれ海の部分に白波が生成できていることから，画像のどの部分が音に対応している対象であるかを認識できていると言える．それに対して，入力画像 sea-A に対して異なる音を入力とした時の結果を比較すると，1 から 20 フレーム，60 から 90 フレーム，120 フレームで同じ場所に類似した白波が生成されていることから，入力音に対して物理的に自然な動きを生成できていないことが分かる．これは，画面外の波の音が収録されていることが原因であると考えられる．音が変化しているにも関わらず，画像が変化していない場合があることで，音に対応した動きを生成できなかったと考えられる．この結果から，人の手や口，花火などの画面内で音と動きが対応している対象は学習できるのに対して，海の波のように動画外で起きていることを学習できないことの確認ができた．

7. まとめと今後の課題

本稿では，音と動きが連動している対象に対して，一枚画像と数秒の入力音からそれらに合う動画を生成する手法を提案した．本手法の有効性を検証するために，4 種類の対象に対して実験を行った．本稿で提案したネットワークでは，用意したデータセットによって，人の手や口や花火のように学習ができる対象と，海のように学習ができない対象があった．今後の課題としては，人の手や口では動作や人数を増やすことで，音の時間的な変化だけでなく，画像の見た目を学習できるデータセットを構築することが挙げられる．

謝辞 本研究の一部は，JST ACCEL(JPMJAC1602) の支援を受けた．

参考文献

[1] Aifanti, N., Papachristou, C. and Delopoulos, A.: The MUG facial expression database, *International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)* (2010).
[2] Aytar, Y., Vondrick, C. and Torralba, A.: SoundNet: Learning Sound Representations from Unlabeled Video, *Neural Information Processing Systems (NIPS)* (2016).
[3] Chen, L., Li, Z., Maddox, R. K., Duan, Z. and Xu, C.: Lip Movements Generation at a Glance, *European Conference on Computer Vision (ECCV)* (2018).
[4] Chen, L., Srivastava, S., Duan, Z. and Xu, C.: Deep Cross-Modal Audio-Visual Generation (2017).
[5] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C. and Bengio, Y.: Generative Adversarial Nets, *Neural Information Processing Systems (NIPS)* (2014).
[6] Graves, A., rahman Mohamed, A. and Hinton, G. E.: Speech Recognition with Deep Recurrent Neural Networks, *Computing Research Repository* (2013).
[7] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, A., Suleyman, M. and Zisserman, A.: The Kinetics Human Action Video Dataset, *Computing Research Repository* (2017).
[8] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization., *Computing Research Repository* (2014).
[9] Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H. and Freeman, W. T.: Visually Indicated Sounds, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
[10] Ronneberger, O., Fischer, P. and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2015).
[11] Shlizerman, E., Dery, L. M., Schoen, H. and Kemelmacher-Shlizerman, I.: Audio to Body Dynamics, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
[12] Soomro, K., Zamir, A. R. and Shah, M.: UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, *Computing Research Repository* (2012).
[13] Suwajanakorn, S., Seitz, S. M. and Kemelmacher-Shlizerman, I.: Synthesizing Obama: learning lip sync from audio, *ACM Transactions on Graphics (TOG)* (2017).
[14] Tulyakov, S., Liu, M.-Y., Yang, X. and Kautz, J.: Mocogan: Decomposing motion and content for video generation, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
[15] Zhou, Y., Wang, Z., Fang, C., Bui, T. and Berg, T. L.: Visual to Sound: Generating Natural Sound for Videos in the Wild, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).