Vol.2018-MUS-119 No.31
Vol.2018-SLP-122 No.31
2018/6/16

DRM を用いた唇動画像と音声の双方向変換

塚本伸，中鹿亘（電通大）

1. はじめに

現在，画像から音声への変換技術は，音声が欠落した映像からの発話復元や，音声・聴覚障害者の補助などに幅広く用いられており，また，一方で音声から画像への変換技術を用い，音声作品からその唇の動きを再現することで，あたかも本人が話しているかの映像を作ることができる．

これまで，画像を音声へ, もしくは音声を画像へと変換させる際に，Deep Neural Network (DNN) を用いる場合，画像か音声かのどちらか片方を入力とした DNN をそれぞれについて構築する必要があり，学習にかかるコストが余分にかかってしまう. そこで，本研究では，学習コストの削減および精度向上を目的とし，Deep Relational Model (DRM) を用いて，入力された唇動画像および音声のそれぞれの関係性を学習させることで，画像生成器と音声合成器の構築を同時に達成することができる手法を提案する．

関連研究として，混合正規分布モデル (Gaussian Mixture Model: GMM) を用いた最尤推定に基づく唇動画像からの音声生成 [3] や，唇動画像から音声への変換の逆問題についても，聴覚障害者の補助を目的として，隠れマルコフモデルを用いて音声から口唇動作の作成をする手法 [4] がある．

2. 関連手法

2.1 Deep Relational Model

画像から音声へ，もしくは音声から画像への変換を Deep Neural Network (DNN) を用いて行なった場合，DNN をそれぞれのタスクごとに一つずつ用意する必要があり，学習にかかるコストが高くなってしまう．本研究では Deep Relational Model (DRM) [1] を用いることによって上記の問題を解決する．DRM は可視層を x と y の 2 つ用意し，複数の隠れ層を可視層で挟み込んだ形のネットワークであり， x と y 間の関係性をより上位のレベルで表現するモデルである．

DRM は 1 つ目の可視変数 $\boldsymbol{x} \in \{0,1\}^X$ と 2 つ目の可視変数 $\boldsymbol{y} \in \{0,1\}^Y$, そして隠れ変数 $\boldsymbol{h}^{(l)} \in \{0,1\}^{J_l} (l = 1, \cdots, L)$ によって与えられる．ここで L は隠れ層の数である． DRM の各ユニットは隣接する層のユニットにのみ結合をもち，同じ層のユニットには結合をもたない．

3. Gaussian-Gaussian DRM

本研究では画像と音声から得られるそれぞれの特徴量は連続値からなる．したがって，提案モデルでは，[1] のように Bernoulli 分布は用いずに，Gaussian 分布のみを扱う．本研究では，[2] の GCDRM のエネルギー関数を参考に，GGDRM のエネルギー関数を次のように定める．

$$
\begin{aligned}
&E(\boldsymbol{x}, \boldsymbol{y}, \forall \boldsymbol{h}^{(l)}; \theta) \\
=&\frac{(\boldsymbol{x}-\boldsymbol{b})^T(\boldsymbol{x}-\boldsymbol{b})}{2\sigma^{(x)2}} - (\frac{\boldsymbol{x}}{\sigma^{(x)2}})^T \mathbf{W}^{(1)g} \boldsymbol{h}^{(1)} \\
&- \sum_{l=1}^{L} \boldsymbol{c}^{(l)^T} \boldsymbol{h}^{(l)} - \sum_{l=2}^{L} \boldsymbol{h}^{(l-1)T} \mathbf{W}^{(l)} \boldsymbol{h}^{(l)} \\
&+ \frac{(\boldsymbol{y}-\boldsymbol{d})^T(\boldsymbol{y}-\boldsymbol{d})}{2\sigma^{(y)2}} - \boldsymbol{h}^{(L)T} \mathbf{W}^{(L+1)} \frac{\boldsymbol{y}}{\sigma^{(y)2}} \qquad (1)
\end{aligned}
$$

ここで，$\boldsymbol{x} \in \mathbb{R}^X$ ，$\boldsymbol{y} \in \mathbb{R}^Y$ はそれぞれ多変量正規分布に従うユニットを表す．$\sigma^{(\boldsymbol{x})} \in \mathbb{R}^X, \sigma^{(\boldsymbol{y})} \in \mathbb{R}^Y$ はそれぞれ可視変数 $\boldsymbol{x}, \boldsymbol{y}$ の偏差を表し，$\boldsymbol{b} \in \mathbb{R}^X$, $\boldsymbol{c}^{(l)} \in \mathbb{R}^{J_l}$, $\boldsymbol{d} \in \mathbb{R}^Y$ はそれぞれ 1 つ目の可視層，隠れ層 l, 2 つ目の可視層のバイアスを， $\mathbf{W}^{(1)} \in \mathbb{R}^{X \times J_1}$, $\mathbf{W}^{(l)} \in \mathbb{R}^{J_{l-1} \times J_l}$, $\mathbf{W}^{(L+1)} \in \mathbb{R}^{J_L \times Y}$ はそれぞれ 1 つ目の可視層と隣接する隠れ層，隠れ層 $l-1$ と隠れ層 l, 隠れ層 L と 2 つ目の可視層間の重みを表し，いずれも推定すべきパラメータである．また，式中の除算は要素ごとの除算を表す．

可視層の条件付き確率分布は次のようになる．

$$
p(x_i = x | \boldsymbol{h}^{(1)}) = \mathcal{N}(x | b_i + \mathbf{W}^{(1)}_{i:} \boldsymbol{h}^{(1)}, \sigma_i^{(x)2}) \qquad (2)
$$

$$
p(y_k = y | \boldsymbol{h}^{(L)}) = \mathcal{N}(y | d_k + \mathbf{W}^{(L+1)T}_{:k} \boldsymbol{h}^{(L)}, \sigma_i^{(y)2}) \qquad (3)
$$

ここで，$\mathcal{N}(\cdot|\mu, \sigma^2)$ は平均 μ，分散 σ^2 の正規分布を表す．また，1 層目，L 層目の隠れ層について，条件付き確率分布はそれぞれ

$$
\begin{aligned}
&p(\boldsymbol{h}_j^{(1)} = 1 | \boldsymbol{x}, \boldsymbol{h}^{(2)}) \\
=&\sigma(\boldsymbol{c}_j^{(1)} + \mathbf{W}^{(1)T}_{:j} \frac{\boldsymbol{x}}{\boldsymbol{\sigma}^{(x)2}} + \mathbf{W}^{(2)}_{j:} \boldsymbol{h}^{(2)}) \qquad (4) \\
&p(\boldsymbol{h}_j^{(L)} = 1 | \boldsymbol{y}, \boldsymbol{h}^{(L-1)}) \\
=&\sigma(c_j^{(L)} + \mathbf{W}^{(L)T}_{:j} \boldsymbol{h}^{(L-1)} + \mathbf{W}^{(L+1)T}_{j:} \frac{\boldsymbol{y}}{\boldsymbol{\sigma}^{(y)2}}) \qquad (5)
\end{aligned}
$$

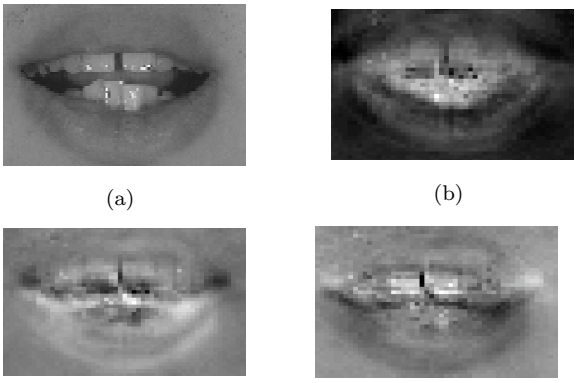
となる．$2, \ldots, L-1$ 層目の隠れ層の条件付き確率分布は

$$
\begin{aligned}
&p(h_j^{(l)} = 1 | \boldsymbol{h}^{(l-1)}, \boldsymbol{h}^{(l+1)}) \\
=&\sigma(c_j^{(l)} + \mathbf{W}^{(l)T}_{:j} \boldsymbol{h}^{(l-1)} + \mathbf{W}^{(l+1)}_{j:} \boldsymbol{h}^{(l+1)}) \qquad (6)
\end{aligned}
$$

と表される．また，GGDRM のパラメータ $\theta = \left\{ \mathbf{W}, \boldsymbol{b}^T, \boldsymbol{c}, \boldsymbol{d}^T, \boldsymbol{\sigma}^{(x)}, \boldsymbol{\sigma}^{(y)} \right\}$ は，従来の DRM と同様に

ⓒ 2018 Information Processing Society of Japan

Vol.2018-MUS-119 No.31
Vol.2018-SLP-122 No.31
2018/6/16



(a) (b) (c) (d)

図 1: (a) 元の画像 (b) 初期値ランダム DNN (c)fine-tuning 前 (d)fine-tuning 後の DRM

対数尤度 $\mathcal{L}$ が最大となるように推定される．また，分散パラメータについては，常に非負値となるように，その対数 $z_i^{(x)} = \log \sigma_i^{(x)2}, z_k^{(y)} = \log \sigma_k^{(y)2}$ を更新することで推定する．

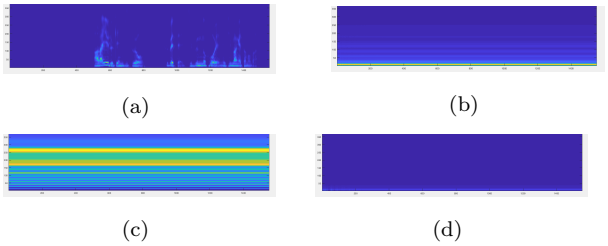
4. 実験

4.1 実験条件

実験には，男性 1 名の文章発話音声とその動画像が含まれる M2TINIT データセット [5] を用いる．元画像の全体のサイズは 720×480 ピクセルである．画像から特徴量を抽出する際には，対象領域である唇周辺の領域を 55×55 のサイズで切り出し，2 次元離散コサイン変換を行なった後に主成分分析を行うことで次元圧縮をし，画像特徴量を得る．音声からは 32 次元メルケプストラムを抽出し，それを音響特徴量として使用する．また，音声データのサンプリング周波数は 48 kHz を 12 kHz にダウンサンプリングして実験に使用した．また，隠れ層数は 2 で，隠れ素子数は 100 で実験を行なった．

4.2 実験結果

図 1 および 2 はそれぞれ，実験によって得られた画像とスペクトログラムである．図 1(b) を見ると，得られた他二枚の画像と比べて，全体的に暗く唇の輪郭が少し不鮮明になっている．また，図 2(d) は，唇の輪郭がはっきりしており，他二枚と比べて唇の画像だと判別が容易である．音声については，生成されたメルケプストラムをスペクトログラムへ復元したところ，図 2 が得られた．図 2(a) とそれ以外の三つの図を見比べると，初期値ランダム DNN および DRM により得られたスペクトログラムは全フレームにわたって平均化されており，フレームごとにあまり違いが生じていないことがわかる．



(a) (b) (c) (d)

図 2: (a) 元のスペクトログラム (b) 初期値ランダム DNN (c)fine-tuning 前 (d)fine-tuning 後の DRM

5. おわりに

本稿では， DRM を用いて DNN 音声合成器および画像生成器を同時に学習する手法について提案した．

画像については，fine-tuning により得られた結果は，fine-tuning をせずに得られた画像と比べると，唇の画像だと判断しやすいものであった．また，従来手法である DNN と比べると，元の画像に近いものが生成されることがわかった．音声については，元のスペクトログラムと得られたスペクトログラムを比較したところ，得られたものはフレーム全体で出力が平均化されていた．

今後の展望として，DRM を用いた実験において，生成される画像と音声の特徴量を目標となる画像と音声の特徴量に近づくように繰り返し実験を行う．

参考文献

[1] T. Nakashika *et al.*: “Modeling bidirectional relationships for image classification and generation”, ICASSP, pp.13271331 (2016)

[2] K. Sone *et al.*: “Pre-training Method for DNN-based Speech Recognition and Synthesis Based on Bidirectional Conversion between Text and Speech”, IPSJ SIG Technical Report (2017)

[3] R. Ra *et al.*:“Visual-to-speech conversion based on maximum likelihood estimation”, 2017 Fifteenth IAPR International Confetence on Machine Vision Applications (MVA), pp.518-521 (2017)

[4] E. Yamamoto *et al.*:“Lip movement synthesis from speech based on hidden Markov models”, Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 154-159 (1998)

[5] S. Sako *et al.*: “HMM-based text-to-audio-visual speech synthesis image-based approach,” in Proc. ICSLP, Vol. 3, pp.2528 (2000)

ⓒ 2018 Information Processing Society of Japan