I-002

クロスモーダルな埋め込み空間の学習による音声を入力とした表情画像生成
Generating facial expression images from speech input by learning cross-modal embedding spaces

小関 虎太朗† 清 雄一† 田原 康之† 大須賀 昭彦†
Kotaro KOSEKI Yuichi SEI Yasuyuki TAHARA Akihiko OHSUGA

1 はじめに

マルチモーダル情報処理とは文字や音，画像などの異なる性質を持つ複数の情報を統合的に処理することによって，情報間の関連性を明らかにすることである．近年，人間の声と顔のマルチモーダル学習についての研究は，オレオレ詐欺対策，電話での会話体験の向上や，生体認証などのセキュリティの観点から重要性が注目され広く研究が行われている [1]．また，人間の認知機能において声による認知と顔による認知とのマルチモーダルな関連性は生物学と神経心理学の観点からもその存在が実験的に確認されており，人間は声と顔を関連付けて記憶することで対象人物の認識精度を向上させるということが示されている [2].

これらの前提を踏まえて，本研究では人間の音声と顔に注目し音，画像間でのデータ同士の関連性をマルチモーダル学習により獲得することによって音声情報を入力とし，入力音声に対して尤もらしい顔画像を生成するモデルを作成することを目指す.

2 提案手法

2.1 提案手法の全体像

まず，VAE と音声エンコーダを用いて入力音声から顔画像を生成するモデルの概略図を図 1 に示す．本研究では，はじめに顔画像を入力した際に入力と同一の顔画像を再構成することのできる VAE モデルを作成する．次に作成した VAE モデルのエンコーダの出力する顔特徴に対して，顔画像に対応する音声を入力したときに同一の特徴を抽出することができるように音声エンコーダを学習させる．これにより，ある話者の発話音声を音声エンコーダに入力した場合出力される音声特徴は，話者の顔画像を VAE モデルのエンコーダに入力した際に得られる顔特徴と類似したものとなるため，音声エンコーダの出力から潜在変数 z を計算し VAE モデルのデコーダを用いてデコードすることによって入力音声に対応する顔画像を生成することができる.

2.2 VAE

潜在変数に確率変数を用いた AutoEncoder で，潜在変数 z 及び損失関数は AutoEncoder におけるエンコーダ部分の出力より得られる平均 μ 分散 σ によって以下の式で定義される.

$$z \sim \mathcal{N}(\mu, \sigma) \tag{1}$$

$$\mathcal{L}_{\mathrm{VAE}}[\mathrm{q}_\varphi(\mathrm{z}|\mathrm{x})] = \mathrm{E}_{\mathrm{q}\varphi}[\log \mathrm{p}_\theta(\mathrm{x}|\mathrm{z})] - \mathrm{KL}[\mathrm{q}_\varphi(\mathrm{z}|\mathrm{x})\|\mathrm{p}(\mathrm{z})] \tag{2}$$

式 (2) において，右辺第 1 項は潜在変数 z を入力画像 x に戻した際の対数尤度を最大化させるための項であり，第 2 項は事前分布から事後分布への KL ダイバージェンス距離を表す.

VAE は他の生成モデルである GAN などと比較して平均的な画像を生成することに優位性があり，音声と顔の対応関係は一位に定まるものではない．そのため，本研究においては VAE を用いることが適当であると判断した.

2.3 音声エンコーダ

DeepTalk[3] モデルは話者認識のタスクに関して優れた精度を示している．話者認識タスクでは各話者の音声の発話内容よりもピッチやアクセントのつけ方など声調の特徴をうまく学習させることが重要とされている．本研究においても抽出する音声特徴は発話内容よりも声調特徴の方が重要であるため，DeepTalk モデルのエンコーダ部分を音声エンコーダとして使用する.

3 実験

3.1 データセット

学習データには多様性を持った人間の顔画像とそれに対応する発話音声がペアになったデータセットが必要である．そこで，データセットとして 6000 以上の話者とそれに伴う 100 万以上の発話動画を伴う VoxCeleb2[5] を用いた．VoxCeleb2 は動画のデータセットであるため mp4 データとなっているが，必要なデータは動画中の音声データと顔を捉えたフレーム画像であるため前処理が必要となる．そこで，動画から正面を向いている瞬間の画像とそれに対応する動画の音声データをそれぞれ保存しデータセットとした.

3.2 VAE の学習

VAE の学習は以下の設定で行った.

Epoch ：2000
Learning Rate ：0.001
Input Size ：150*150 pixel
Optimizer ：Adam
Batch Size ：1000

学習した VAE モデルに対して顔画像を入力し，顔画像生成をした際の結果を図 2 に示す.

3.3 音声エンコーダの学習

音声エンコーダの学習は以下の条件で行った.

Epoch ：17000
Learning Rate ：0.0001
Loss Function ：MSE
Optimizer ：Adam
Batch Size ：32

学習した音声エンコーダを用いて図 1 に基づいて顔画像を再構成した結果を入力音声に対応する本来の顔画像と共に図 3 に示す.

4 考察

† 電気通信大学大学院 情報理工学研究科 情報学専攻
Graduate School of Informatics and Engineering, The University of Electro-Communications

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

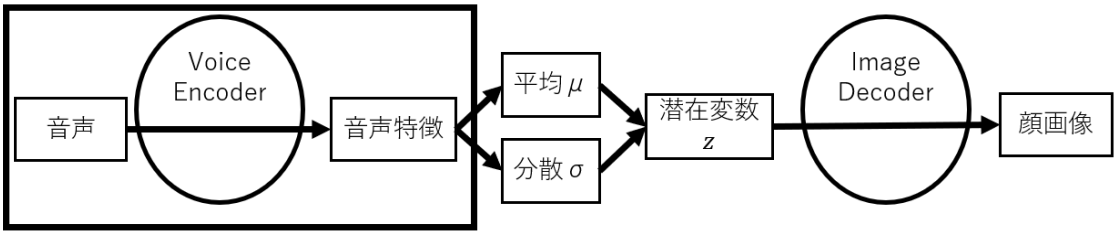

図 1　提案手法の概略図

図 2　VAE による顔生成

図 3　提案手法による顔画像生成

4.1　VAE の学習・生成結果について

図 2 の結果は顔の向きや性別などの特徴はやや捉えられているが表情に関していえば Input ではしかめっ面であるのに対し Output は柔和な表情であったりと，あまり上手く捉えることが出来ていない印象を受ける．このような結果となった原因として予想されることとして，学習データセットの顔の向きに関してのばらつきが挙げられる．顔の向きが変わると全ての顔のパーツの距離の比率が大きく変わってしまう．そのため，VAE の潜在空間に特徴がマッピングされ似た特徴を持つクラスターが形成されていく際に，輪郭や顔パーツの大きさ，形などの人間の属性をとらえるために必要な情報よりも大きな単位で顔の向きについてクラスターができてしまう．具体的には「右向き顔クラスターのタイプ A タイプ B，左向き顔クラスターのタイプ A タイプ B」の様に，本来同一のクラスターに属するはずのタイプ A 同士，タイプ B 同士の顔が別のクラスターにマッピングされてしまったことが考えられる．

4.2　音声エンコーダの学習・顔画像生成について

音声エンコーダの学習については，図 3 がどの入力に対してもほぼ同一の出力をしていることから上手くできなかったことが分かる．この原因として考えられるのは音声エンコーダモデルの出力データの形の変形方法である．音声エンコーダの出力は，学習の際に VAE の中間表現である特徴量と出力の形をそろえるために多層の畳み込みの後に最終層として全結合層によって出力次元を大きく変更している．これによって，獲得した特徴を失ってしまっていると考えられる．

5　おわりに

本研究では，話者識別タスクにおいて高い精度を持つ DeepTalk の音声エンコーダモデルと VAE を用いて音声入力から顔画像を生成する手法を提案した．結果として入力音声に対応する顔画像に似た画像を生成することはできなかったが，音声情報から人間らしい画像の生成をすることには成功している．また，VAE の品質と音声特徴と顔画像特徴との対応関係についての学習方法を突き詰めることで課題を達成することができるという提案手法の妥当性を確認することができた．現状での実験結果から，精度向上のための施策案として出力の形を合わせるために音声エンコーダの出力を変形するのではなく，VAE モデルの畳み込み処理や全結合層を増やすことで既存研究として特徴抽出の性能が担保されている音声エンコーダに手を加えずに学習を進められるようにするという工夫や，音声特徴と顔画像特徴の対応関係学習の際の損失関数を試行錯誤することが考えられる．また，VAE の損失関数に関しても β-VAE[6] や Vector Quantised-VAE2 (VQ-VAE2)[7] のような様々な VAE モデルの実装を参考にして改良していくことも考えられる．今後は，これらの手法を実装・実験することで生成精度の向上を目指す．

謝辞

本研究は JSPS 科研費 JP21H03496, JP22K12157 の助成を受けたものです．

参考文献

[1] A. K. Jain, A. Ross, and S. Prabhakar, “An introduction to biometric recognition,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 1, pp. 4–20, Jan 2004.

[2] K. von Kriegstein and A.-L. Giraud,“Implicit multisensory associations influence voice recognition,” PLoS Biology, vol. 4, pp. 1709 – 1714, 2006.

[3] Chowdhury, Anurag, Arun Ross, and Prabu David. ”DEEPTALK: Vocal Style Encoding for Speaker Recognition and Speech Synthesis.” ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

[4] Goodfellow, Ian, et al. ”Generative adversarial nets.” Advances in neural information processing systems 27 (2014).

[5] J. S. Chung*, A. Nagrani*, A. Zisserman VoxCeleb2: Deep Speaker Recognition INTERSPEECH, 2018.

[6] Higgins, Irina, et al. ”beta-vae: Learning basic visual concepts with a constrained variational framework.” (2016).

[7] Razavi, Ali, Aaron Van den Oord, and Oriol Vinyals. ”Generating diverse high-fidelity images with vq-vae-2.” Advances in neural information processing systems 32 (2019).

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.