7W-01

画像・音声データを用いた対話エージェントのキャラクター設計支援*

栗山 秀平†,　伊藤 克亘‡,

1　まえがき

我々は会話をする際に, 相手の声や話し方の細かな差から相手の特徴を無意識のうちに区別する. そこで本研究では, キャラクターの音声や画像からキャラクターの外見や話し方を推定することで, 対話システム開発時のキャラクター決定補助に役立てる. 本研究では従来研究を参考に, キャラクターの動画から得られる音声と画像から, キャラクターの外見画像を出力として学習させる. また, キャラクターの会話から得られたテキストから抽出した文章特徴を利用し, VQA のような手法を用いてキャラクターの話し方を推定する.

2　研究概要

音声データと画像の関連性の学習を行うために, 本研究では, Speech2Face[1] の手法を参考にする.

まず, AVSpeech[2] や VoxCeleb[3] で集めた動画から得られた話者の画像データに VGGFace[4] を用いて顔データを 4096 次元のベクトルに変換する.

また, 音声データからは音声波形を抽出し, 得られた音声波形のスペクトログラムをボイスエンコーダを用いて同様にベクトル型のデータに変換する.

このボイスエンコーダは畳み込みニューラルネットワーク (CNN) である.

得られたデータをフェイスデコーダ [7] にかけることで, 音声特徴に応じた顔画像の出力を行う.

本研究ではこのモデルを構築したあと, 映像データから用意した三種類の学習データを用いて学習を行う.

学習した結果に関しては, 比較評価実験を行うために, 出力結果をキャラクターの画像に統一するための処理を行う. 人間の音声から人間の学習データを用いた結果の出力画像を, GAN による既存の手法を参考にキャラクター画像に変換する.

話し方から人物の話し方を推測するために, VQA[9] のような方法を用いる. キャラクターのセリフのテキスト情報を印象語 [5][6] や役割語 [8] を参考に分類しそれを画像の特徴量と関連付けて学習を行う. 表 1 では従来研究において分類された役割の例であり, キャラクターの話す言葉はこれらのような分類方法で区分することができる. 入力された人物画像に対して, どのようなセリフの特徴が現れるかを結果として確率的に表すことで, キャラクター設定における文章特徴の決定の補助を促す.

表 1. キャラクタ印象語の例

キャラクタ印象後	
アイドル的な	中性的な
色気のある	厨二病の
オタクな	ツッコミ役の
厳しい	冷たい
華奢な	天才的な
食いしん坊な	毒舌な
下品な	ドジな
元気な	生意気な
サバサバした	ナルシストの
しっかりした	馬鹿な
嫉妬深い	派手な
神経質な	恥ずかしがり屋の
素直な	腹黒い
短気な	病弱な
知的な	飄々とした
不愛想な	プライドの高い
不器用な	不良の
平凡な	真面目な
マゾヒスティックな	わがままな
優しい	弱気な
冷静な	ミステリアスな
暗い	威圧的な
意地悪な	穏やかな
運動神経の良い	感情的な
泣き虫な	狂気的な
胡散臭い	好奇心旺盛な
攻撃的な	純粋な
上品な	包容力のある
高飛車な	大人っぽい
さわやかな	若い
クールな	

3　予備実験

得られた学習データのうち, VoxCeleb の動画 1513 個に対して Speech2Face のモデルを 40 エポックで学習させた. テストデータとして, 学習用データではない動画や簡単な文章を話す合成音声のデータを 10 個ほど用意し, それらの音声からどのような画像が生成されるかを調べた.

4　結果

テスト用動画の音声から推測された人物, 合成音声から推測された人物どちらも, 似た音声を持つ人物を選べている傾向にはあったが, 女性の音声で男性を推測してしまうことがどちらのテストデータでも生じた.

---

*: Character design support for dialogue agents using image and voice data *Shuhei Kuriyama (Hosei Univ.) et al.*

†法政大学大学院 情報科学研究科
‡法政大学 情報科学部

図 1 では入力音声の本来の人物画像と推測された上位 5 人の人物画像を並べている．上二段がテスト用の女性の動画で下二段が音声合成で作成した女性の音声である．また, 複数の推測結果に同じ人物が現れることがあった.

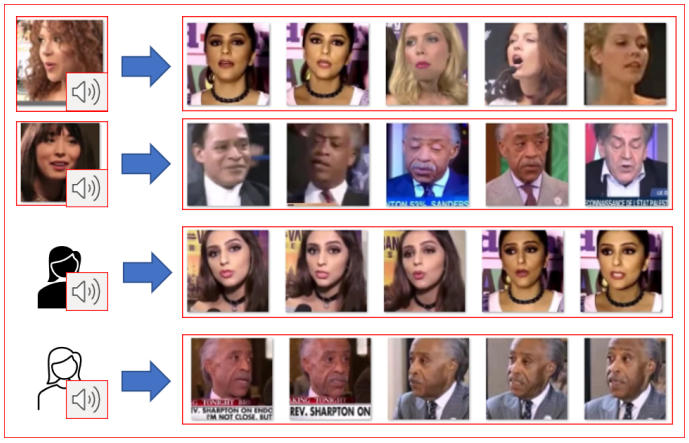

図 1. 人物推測結果

5 考察

今回の結果では, 推測結果にばらつきはあったものの, 一部のテストデータからの推測では, 似た音声の人物を推測することができていた．推測結果の偏りや精度の悪さは, 用いた学習データの動画の量の偏りや, 学習に用いたデータ量の不足による分類不足が原因だと考える.

6 あとがき

今後の予定としては残りの実験を行いつつ, Speech2Face のモデルの学習を行う際のデータ数を増やし, 結果の精度にどのような変化が起こるかを調べていく．学習データの一人あたりの動画の量が一定では無かったことが実験結果に大きく影響していると考えたため, そのような点も注意して実験を重ねることで, 予備実験で得られた知見を活かしたい．また, 推測した人物の顔画像を Face++[10] などで評価し，従来研究と比べてどのくらいの精度が出せるのか実験していく．セリフの学習についても実験を重ね，どのような手法がセリフの推測に適しているのか調べていく.

参考文献

[1] Tae-Hyun Oh and Tali Dekel and Changil Kim and Inbar Mosseri and William T. Freeman and Michael Rubinstein and Wojciech Matusik. Speech2Face: Learning the Face Behind a Voice, 2019

[2] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Has-sidim, W. T. Freeman, and M. Rubinstein. Looking to listen atthe cocktail party: A speaker-independent audio-visual modelfor speech separation.ACM Transactions on Graphics (SIG-GRAPH), 37(4):112:1-112:11, 2018.

[3] Nagrani, Arsha and Chung, Joon Son and Zisserman, Andrew. VoxCeleb: a large-scale speaker identification dataset. ISCA, Interspeech Aug.2017.

[4] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep facerecognition. InBritish Machine Vision Conference (BMVC),2015.

[5] 真木恵, 菊池英明. 声優演技音声の声質に基づいたキャラクタ印象認知モデルの構築. 日本音響学会春季研究発表会講演論文集, 3-P-56, pp.563-566, Sep.2013.

[6] 真木恵, 菊池英明, 声優演技音声の声質のステレオタイプ -キャラクタ印象評 価尺度の構築と声質の印象表か実験を通して. 日本漫画学会第 13 回大会, July.2013.

[7] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, andW. T. Freeman. Synthesizing normalized faces from facialidentity features. InIEEE Conference on Computer Visionand Pattern Recognition (CVPR), 2017.

[8] 金水敏（2003）『ヴァーチャル日本語役割語の謎』岩波書店

[9] Aishwarya Agrawal and Jiasen Lu and Stanislaw Antol and Margaret Mitchell and C. Lawrence Zitnick and Dhruv Batra and Devi Parikh. VQA: Visual Question Answering. 2016.

[10] L. face-based identity verification service. Face++.https://www.faceplusplus.com/attributes/.