論文 / 著書情報
Article / Book Information

題目(和文)	音声と映像の関連性を活用した現実的な音声駆動型話者顔合成とその応用
Title(English)	Realistic Speech-Driven Talking Face Synthesis via Audio-Visual Association Exploitation and its Applications
著者(和文)	SunYasheng
Author(English)	Yasheng Sun
出典(和文)	学位:博士(学術), 学位授与機関:東京科学大学, 報告番号:甲第25号, 授与年月日:2024年12月31日, 学位の種別:課程博士, 審査員:小池 英樹,篠田 浩一,岡崎 直観,齋藤 豪,井上 中順
Citation(English)	Degree:Doctor (Academic), Conferring organization: Institute of Science Tokyo, Report number:甲第25号, Conferred date:2024/12/31, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名		Yasheng SUN	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	小池英樹	教授	審査員	井上中順	准教授
	審査員	篠田浩一	教授			
		岡崎直観	教授			
		齋藤豪	准教授			

論文審査の要旨（2000字程度）

本論文では「Realistic Speech-Driven Talking Face Synthesis via Audio-Visual Association Exploitation and its Applications」と題し、発話音声から顔アニメーション、Talking Faceを生成する手法について述べている。本論文は英文6章から成る。

第1章「Introduction」では、Talking Face研究の背景を紹介し、現実感のあるTalking Face生成に向けた重要な研究方向性について述べている。先行研究のほとんどが、話者の感情やアイデンティティといった情報を付加するために参照画像やラベルといった外部の参照手がかりを用いており、発話音声だけを用いた現実感のある顔アニメーションの生成は十分に探求されていないことを指摘している。これに対し、発話音声は話者の感情やアイデンティティという暗黙の情報と深く関係しており、発話音声からこれらの情報を抽出できる可能性があることが述べられている。本研究は、暗黙情報を含む視覚的手がかり（顔画像）を用いて学習された特徴空間内の暗黙情報に対応する箇所を、発話音声から得られる暗黙情報に置き換えることで、発話情報だけを用いて現実感のあるTalking Faceを生成する枠組みについて述べている。そして、この枠組みを用いることで2次元および3次元Talking Faceへの応用が可能であることが述べられている。

第2章「Related Work and Preliminary」では、2次元Talking Face合成および音声駆動型3次元Talking Head合成に関する先行研究が、データセットと手法の二つの側面から述べられている。特にデータセット部分では、本研究で使用した4つのデータセットの詳細が説明されている。手法部分では、現実感のあるTalking Faceを対象とした研究や、感情認識や音-画像を用いた学習に関連する研究が述べられている。さらに、3次元パラメトリック顔モデルの定義、StyleGANアーキテクチャ、拡散モデルの基本理論、事前学習モデルなど、本研究に関連する基礎知識が紹介されている。

第3章「Research Proposal」では、発話音声と生成された顔画像の一貫性を考慮した現実感のあるTalking Faceシステムが提案されている。本手法は、(1)対照学習による特徴強化と(2)脱もつれ(disentangled)空間における特徴統合に分けられる。前者では、より感情やアイデンティティに関する強い情報を持つ顔画像を用いた学習を行う。後者では、生成された特徴空間における暗黙情報の位置を特定し、これを発話音声から得られる特徴情報で置き換える。結果として発話音声だけで

現実感のある Talking Face を生成することができることが述べられている。

第 4 章では、第 1 の応用例として Speech2TalkingFace が紹介されている。本手法は、発話音声のみから Talking Face を合成することを目的とし、音声と唇の動きが同期し、かつ話者アイデンティティと矛盾のない動画を生成する。実験の結果、話者アイデンティティ生成において、音声から生成された顔画像の定量的評価において、Similarity(cosine:0.397, L1:60.84), Retrieval(R@1:10.65), Quality(VFS:39.00+-2.90)とすべてにおいて既存手法(Speech2Face, Voice2Face)より優れた値を示した。また音声-唇の同期性能については既存研究(ATVG, Wav2Lip)と同等かそれ以上の性能を示した。以上、本手法は、話者アイデンティティ、唇の同期、頭部姿勢の制御を統合した唯一のアプローチとして優れていることが示されている。

第 5 章では、発話音声と合成された顔画像における感情の一貫性をより改善したシステムとして AVI-Talking が紹介されている。AVI-Talking は感情特徴を強化するために大規模言語モデル (LLM) を活用した。発話音声はまず LLM で処理され、顔の表情に関する詳細な表現が出力される。次にこの表情に関する表現と発話音声が Talking Face システムに与えられ顔の３次元アニメーションが生成される。既存研究(MeshTalk, EmoTalk, CodeTalker 等)との定量的および定性的比較実験により本手法の有効性が示された。

第 6 章「Conclusion」では、この提案手法の主要なアプローチを要約すると同時に、計算の複雑さや古くなった深層学習モジュールといった本研究の限界について述べている。そして、今後の方向性として、より高度な技術で現在の枠組みのモジュールを更新することが強調され、また、発話音声から全身アニメーションを生成するなど、広範な応用の可能性についても議論されている。

以上、本論文では、人間の発話音声に含まれる暗黙の情報を抽出・活用することで、参照画像やラベルといった外部手がかりを使用することなく、発話音声だけでより現実感のある顔アニメーションを生成する新しい枠組みが提案された。開発された 2 つの応用例は、先行研究に対する優位性を示しており、実験によって検証されている。以上から、本論文は高い学術的貢献が認められ、博士（学術）の学位として十分価値があると認められる。