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)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	Computer Science 系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 (Philosophy) Doctor of
学生氏名： Student's Name	Yasheng Sun	審査員主査： Chief Examiner	Hideki Koike

要旨（英文 300 語程度）
Thesis Summary (approx.300 English Words)

This paper, titled "Realistic Speech-Driven Talking Face Synthesis via Audio-Visual Association Exploitation and its Applications," presents a method for exploiting the underlying human speech features to facilitate speech-consistent talking face generation using explicit visual cues, as well as its applications in realistic talking face synthesis. The paper consists of 10 chapters in English.

Chapter 1, "Introduction" introduces the background of talking face studies and one important research direction is realistic talking face generation targeting lifelike synthesis. It points out that most prior talking face synthesis approaches rely on external reference cues to animate a talking face, while direct exploitation of underlying human speech information for synthesis is under explored. But the emotional information and speaker identity features are highly correlated with the human voice according to the principle of acoustic pronunciation. This study devises a framework to exploit this implicit information within human speech via explicit visual cues. This framework facilitates two novel applications and showcases its effectiveness in both 2D and 3D talking face settings.

Chapter 2, "Related Work" includes the prior works on 2D talking face synthesis and speech-driven 3D talking head synthesis. They are described in two veins, datasets, and approaches. The dataset part, in particular, provides the details of the four datasets used in this study. The approach part describes the various series of studies targeting realistic talking faces and some research closely relevant with ours such as emotional recognition and audio-visual association learning. In addition, it introduces the preliminary knowledge relevant to this study, including the definition of 3D parametric facial models, the StyleGAN architecture, the basic theory of diffusion models, and the pretrained models, which serve as the feature extraction models in this study.

Chapter 3 covers the "Research Proposal", which reviews the key aspects in realistic talking face synthesis. A realistic talking face system is proposed to take into account consistency between human speech and visual generation. A high-level overview of the philosophy behind this system is provided, including research motivation, design choice of our framework and in-depth analysis of common points behind our promoted applications. To fully leverage the underlying information present in the human voice and incorporate it into the generative process, a framework is carefully devised to firstly identify a disentangled space via explicit visual cues and then replace the visual cues with audio features. Two case studies involving speech information from the speaker identity and talking status are discussed.

Chapter 4 introduces the first application, Speech2TalkingFace, to improve the speaker identity consistency between visual synthesis and human speech input. This approach is focused on the task setting of synthesizing a talking face solely from a clip of audio, and aims to generate both lip-synchronized and speaker identity consistent videos. Experiments demonstrate that the proposed framework not only generates speaker identity consistent with human speech input but also achieves comparable lip synchronization performance. In contrast to most prior studies, the Speech2Talking-Face work stands out

as the only approach that integrates speech identity-based generation, lip movements, and pose control within a single unified framework.

Chapter 5 introduces the second proposed application, AVI-Talking, to improve the emotional consistency between visual synthesis and human speech input. This approach introduces the texts to reinforce the feature extraction of speech emotion and leverages Large Language Models (LLMs) to instruct the 3D talking face synthesis process. Instead of directly predicting the talking face, this paper decomposes this task into two components, with language instruction as an interface. This strategy not only allows users to specify their desired facial motions but also enhances the extraction of talking status through cross-modal learning. Moreover, the intermediate output enables easy identification of whether any issues stem from the audio-visual instruction module or the talking face instruction module, thereby increasing the model's explainability.

Chapter 6 concludes this thesis by summarizing the major components of the proposed approaches. Meanwhile, it states the limitations of this work like computation complexity and outdated backbone. In future direction, it highlights updating the modules of the current framework with more advanced techniques. Furthermore, other follow-up research directions that have potential for broader applications are also discussed, such as driving full-body movements instead of face regions with human speech.

In summary, this paper targets to exploit implicit information within human voice. A novel framework is proposed to leverage explicit visual cues to facilitate audio-visual association learning. Two kinds of implicit vocal information, speaker identity and talking status, are explored in this work. They are enhanced by visual images and language descriptions, respectively. Incorporating this implicit information from human voice, the talking face system is able to achieve realistic talking face synthesis with consistent audio-visual demonstration. The developed applications demonstrate its unique superiority over previous approaches, which are validated by extensive experiments.