氏　　名	Hieu-Thi Luong
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 2193 号
学位授与の日付	2020 年 9 月 28日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Deep learning based voice cloning framework for a unified system of text-to-speech and voice conversion
論 文 審 査 委 員	主　査　教授　山岸　順一 教授　杉本　晃宏 准教授　稲邑　哲也 教授　戸田　智基 名古屋大学　大学院情報学研究科 教授　篠田　浩一 東京工業大学　情報理工学院

Summary of Doctoral Thesis

Name in full Hieu-Thi Luong

Title Deep learning based voice cloning framework for a unified system of text-to-speech and voice conversion

Speech synthesis is the technology of generating speech from an input. While the term is commonly used to refer to text-to-speech (TTS), there are many types of speech synthesis systems which handle different input interfaces such as voice conversion (VC), which converts speech of a source speaker to the voice of a target, or video-to-speech, which generates speech from an image sequence (video) of facial movements.

This thesis focuses on voice cloning task which is the developing of a speech synthesis system with an emphasis on speaker identity and data efficiency. A voice cloning system is expected to handle less than ideal data circumstance of a particular target speaker. More specifically, we do not have control over the target speaker, recording environment, or the quality and quantity of speech data. Such system will be useful for many practical applications involve generating speech with desired voices. However, it is also vulnerable to misuse which can cause significant damages to the society by people with malicious intentions. By first breaking down the structures of conventional TTS and VC systems into common functional modules, we propose a versatile deep learning based voice cloning framework which can be used to create a unified speech generation system of TTS and VC with a target voice. Given such unified system, which is expected to have consistent performance between its TTS and VC modes, we can use it to handle many application scenarios that are difficult to tackle by just one or the other, as TTS and VC have their own strengths and weaknesses.

As this thesis is dealing with two major research subject, which are TTS and VC, to provide a comprehensive narrative, its content can be considered as comprising of two segments which tackle two different issues: (1) developing a versatile speaker adaptation method for neural TTS systems. Unlike VC which existing voice cloning methods are capable of producing high-quality generated speech, existing TTS adaptation methods are lacking behind in performance and scalability. The proposed method is expected to be capable of cloning voices using either transcribed or untranscribed speech with varying amount of adaptation data while producing generated speech with high quality and speaker similarity; (2) establishing a unified speech generation system of TTS and VC with highly consistent performance between two. To achieve this consistence, is is desirable to reduce the differences between the methodology and use a same framework for both systems. Beside for convenient purposes, such system also has the ability to solve many unique speech generation tasks, as TTS and VC are operated under different application scenarios and complement each other.

On the first issue, by investigating the mechanism of a multi-speaker neural acoustic model, we proposed a novel multimodal neural TTS system with the ability to perform crossmodal adaptation. This ability is the fundamental for cloning voices with untranscribed speech on the basis of backpropagation algorithm. Comparing with existing unsupervised speaker adaptation methods which only involve a forward pass, a backpropagation-based unsupervised adaptation method has significant implication on performance as it allows us to expand the speaker component to other part of the neural networks beside the speaker bias. This hypothesis is tested by using speaker scaling together with speaker bias, or the entire module as adaptable components. The proposed system unites the procedure of supervised and unsupervised speaker adaptation.

On the second issue, we test the feasibility of using the multimodal neural TTS system proposed previously to bootstrap a VC system for a particular target speaker. More specifically, the

proposed VC system is tested on standard intra-language scenarios and cross-lingual scenarios with the experiment evaluations show promising performance in both. Finally given the proof-of-concept provided by earlier experiments, the proposed methodology is incorporated with relevant techniques and components of modern neural speech generation systems to push performance of the unified TTS/VC system further. The experiments suggest that the proposed unified system has comparable performance with existing state-of-the-art TTS and VC systems, at the time this thesis was written, but higher speaker similarity and better data efficiency.

At the end of this thesis, we have successfully created a versatile voice cloning system which can be used for many interesting speech generation scenarios. Moreover, the proposed multimodal system can be extended to other speech generation interfaces or enhanced to provide controls over para-linguistic features (e.g., emotions). These are all interesting directions for future works.

Form8・Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　Hieu-Thi Luong

Title
論文題目　　Deep learning based voice cloning framework for a unified system of text-to-speech and voice conversion

本学位論文は、少量の音声で目標話者の声を模倣する複数の機械学習タスクを融合し、どのタスクにも適応可能な汎用的深層学習モデルを構築する研究であり、全９章から構成され、英語で記述されている。

第 1 章では、本論文で扱う問題の重要性、位置付けおよび貢献について説明している。第２章ではテキスト音声合成、声質変換、これらの関係性と違いについて説明している。テキスト音声合成は、テキスト入力から所望の話者を合成する技術であり、声質変換は他人の入力音声から、所望の話者の音声を合成する技術である。前者は音声対話システムの要素技術であり、後者はエンターテイメント分野において期待されている技術である。これらを統合し、品質や話者の類似性を損なうことなく融合することができれば、音声生成技術の応用分野がさらに拡大する。

第３章では、テキスト音声合成において、少量の音声で目標話者の声を模倣する話者適応タスクについて記述している。これは本研究の開始点であるが、音声の書き起こしラベルを利用しており、他のタスクとの融合を妨げる。そこで、第４章では、書き起こしラベルを利用せずに話者適応を行う Multimodal ニューラルネットワークという新たな構造およびその適応方法を提案している。

第５章では、次に、話者適応の関数をどう定義し、転移学習を行うのが最も効率的か調査している。事前学習済みの Multimodal ニューラルネットワークの重みを単純なファインチューニングで更新するだけでなく、ニューラルネットワークの隠れ層自身を話者埋め込みベクトル等により因子化させることで、たった数文章という少量のデータで目標話者に類似した音響特徴量を出力可能なモデルを構築する方法を報告している。

第６章では第４章の内容を発展させ、Multimodal ニューラルネットワークを音声入力と文字入力から共通の確率的潜在変数を計算させる変文型エンコーダ・デコーダ型ニューラルネットワークと再定義し、これにより、さらに頑健にモデル学習が可能になる事を示した。

第７章では、第６章で提案したニューラルネットワーク構造を利用し、テキスト音声合成を事前学習タスクとし利用し、声質変換タスクへ転移学習する方法を示した。これにより、テキスト音声合成タスク用データベースが、声質変換タスクにおいても利用可能になり、異言語間における声質変換などにおいて有効活用できる事を示した。

第８章では、第６章で提案したニューラルネットワーク構造に、Wavenet、非線形自己回帰等の技術と組み合わせることで、テキスト音声合成タスクと声質変換タスクのどちらにおいても高品質音声を一貫した話者性で合成可能である事を示した。第９章では、以上

の結果をまとめ、将来課題について議論している。

公開発表会では博士論文の章立てに従って発表が行われ、その後に行われた論文審査会及び口述試験では、審査員からの質疑に対して適切に回答がなされた。

博士論文審査の結果、出願者は情報学分野の十分な知識と研究能力を持つと認められ、また研究内容は学位論文として十分なレベルの新規性や有効性があると認められた。また、本論文の内容に関し、国際会議論文５編（そのうち３編はトップ会議）を出版済みである。以上の理由により、審査委員会は、本学位論文が学位の授与に値すると判断した。