氏　　名　加藤 集平

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2243 号

学位授与の日付　2021年3月 24日

学位授与の要件　複合科学研究科 情報学
学位規則第6条第1項該当

学 位 論 文 題 目　Rakugo Speech Synthesis: Toward Speech Synthesis That Entertains Audiences

論 文 審 査 委 員　主　査　山岸 順一
情報学専攻 教授
越前 功
情報学専攻 教授
佐藤 いまり
情報学専攻 教授
小林 隆夫
東京工業大学 名誉教授
森 大毅
宇都宮大学 大学院工学研究科 准教授

（様式３）

博士論文の要旨

氏　　名　加藤　集平

論文題目　Rakugo Speech Synthesis: Toward Speech Synthesis That Entertains Audiences

Conventional speech synthesis research has focused on transferring information which the speech should have, such as content and speakers' emotions, personality, intention, accurately to listeners. Setting this purpose is reasonable considering that speech is a kind of media. Today, some speech synthesis systems can successfully produce speech as natural as human speech, albeit in the case of using well-articulated read speech.

However, the role of speech is not just information transfer. For example, verbal entertainment, including rakugo, on which we focus in this thesis, entertains audiences through the medium of speech. In other words, speech has a role of stirring listeners' emotion. This role has not been focused on enough in speech synthesis research, but we believe it is a good time to attempt to realize speech synthesis that entertains audiences because some modern speech synthesis systems have an ability to produce speech as natural as human speech, as mentioned above, albeit in the case of read-aloud speech.

In this thesis, we attempt to build rakugo speech synthesis as a challenging example of speech synthesis that entertains audiences. Rakugo is a traditional Japanese form of verbal entertainment similar to a combination of one-person stand-up comedy and comic storytelling. Although rakugo has a more than 300-year history, it is popular even today in Japan. In rakugo, a performer plays multiple characters, and conversations or dialogues between the characters make the story progress.

First, we built a large rakugo speech database for our study because there was no rakugo speech databases usable to train speech synthesis models. Most commercial rakugo recordings, thousands of which we can easily access, are live recordings that include noise and reverberation, whereas even modern speech synthesis cannot yet properly model such noisy and reverberant speech; therefore, we needed to build a rakugo speech database. We recorded performances by a shin-uchi (first-rank professional) performer to train speech synthesis models, and performances by professional performers at various levels including the shin-uchi performer to evaluate synthesized speech. We not only transcribed the pronunciation of the recorded speech but also appended context labels to each sentence for better modeling of the speech.

Using the database, we modeled rakugo speech using segment-to-segment neural transduction (SSNT) based speech synthesis. The SSNT-based model has no soft

attention network. An attention network maps the encoder and decoder time steps in a sequence-to-sequence speech synthesis model. Sequence-to-sequence models greatly improve the quality of speech synthesis, but attention networks occasionally cause unacceptable errors during synthesis. Since rakugo speech is far more diverse and casually-pronounced than speech ordinarily used for building speech synthesis, an attention network may cause errors more frequently; therefore using SSNT-based speech model, which has no attention networks, will be reasonable for modeling rakugo speech. We also used global style tokens (GSTs), which is a style transfer mechanism for sequence-to-sequence models, or manually labeled context features to enrich speaking styles of synthesized rakugo speech. Although the combination of the SSNT-based model and GSTs produced somewhat natural, character-distinguishable, and content-understandable speech, the mean opinion scores for this speech were just around 3 through a listening test.

For further improvement, we attempted Tacotron 2, a state-of-the-art speech synthesis model, and an enhanced version of it with self-attention to better consider long-term dependency. We also used GSTs, manually labeled context features, or the combination of them. Through a listening test, we found that state-of-the-art TTS models could not yet reach the professional level, and there were statistically significant differences in terms of naturalness, distinguishability of characters, understandability of the content, and even the degree of entertainment; nevertheless, the results of the listening test provided some interesting insights: 1) we should not focus only on naturalness of synthesized speech but also the distinguishability of characters and the understandability of the content to further entertain listeners; 2) the fo expressivity of synthesized speech is poorer than that of human speech, and more entertaining speech should have richer fo expression.

Lastly, we proposed a novel methodology for evaluating rakugo speech and conducted a listening test to investigate how the level of rakugo speech synthesis compares to professional rakugo performers at various levels. Through a listening test, we found that the level of speech synthesis did not reach that of human professionals. On the other hand, the results suggested that we also should at least improve the fo expression of speech synthesis to catch up with human professionals.

Although there is room for improvement, we believe this thesis is an important stepping stone toward achieving entertaining speech synthesis at the professional level.

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　加藤 集平

Title
論文題目　Rakugo Speech Synthesis: Toward Speech Synthesis That Entertains Audiences

本学位論文は、日本の伝統話芸である落語音声を機械学習の対象として取り上げ、落語の実演データから深層学習モデルを学習、あたかもプロの噺家の様に、噺を読み上げる落語音声合成システムを構築することで、機械が人を楽しませる事が可能になるのか？と言う科学的な問いを調査することを目標としている。

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。入力文章から音声を生成する音声合成技術は、近年深層学習の発展により大きく進展し、(1)新聞記事の読み上げ等ごく限られた条件下ではあるが、人間に非常に近い自然な音声を生成可能となっていること、(2)音声にはニュアンスや意図なども含まれているが、情報伝達や質問回答を主たる目的としてきた従来の音声合成研究においては、この様な観点はこれまで全く重要視されてこなかったこと、また評価方法も確立されていなかったこと、(3) 話の内容がある程度決まっている古典落語が本問いへの調査に適していることの説明がなされている。２章では落語の歴史、噺の構造、噺家の格付けなど、落語に関する一般的な概説が記載されている。

第３章では、江戸落語の真打（最高位の格付け）の柳家三三師匠の協力のもと、古典落語の実演を無響室で収録し構築した新たな２種類のデータベースの詳細について記述している。１つ目のデータベースが、ニューラルネットワークを最適化する際に利用する音声データベースであり、様々なアノテーションを行なった江戸落語 25 演目が含まれる。２つ目のデータベースが、異なる格付けの噺家による実演を収録した音声データベースである。これを評価データとして用いる事で、人間の実演と機械学習による生成結果とを比較し、噺家としての格付けを行う事が可能になる。また簡易な音響分析結果も示している。

第４章では、従来のパイプライン型テキスト音声合成と end-to-end 音声合成の違い、および Sequence-to-sequence 型ニューラルネットワークについて概説している。その後、古典落語には古い言葉遣いが利用され、かつ口語である事から、現代語を想定したテキスト解析ツールや辞書を利用することは難しく、パイプライン型テキスト音声合成の適用が難しいこと、それゆえ、データベースに付与されたアノテーション以外の外部知識を使わない end-to-end 学習が適している事に言及している。

第５章では、収集した落語実演データのモデリング手法の初期検討として、segment-to-segment neural transduction という hard attention 型 end-to-end 時系列ニューラルネットワークに、発話様式を参照音声からエンコードする global style token ネットワークを組み合わせた構造を提案し、これにより課題はあるものの、落語音声をある程度モデル化でき、音声も合成可能であることを示した。音声の品質、役の区別、内容理解に関する

主観評価実験結果も示した。

第６章ではモデリング手法をさらに改良し、encoder-decoder ネットワーク、非線形自己回帰ネットワーク、soft-attention ネットワーク、self-attention ネットワーク、および global style token ネットワーク、およびアノテーション情報を組み合わせ利用するモデリング結果も示し、落語の小噺単位で真打の実演と比較する評価、および、問題点の分析を行なった。

第７章では、異なる格付けの噺家（真打・２つ目・前座）による実演を収録した２つ目の音声データベースに、第６章で提案したシステムで共通の噺を合成し加えた上で、音声の自然性、役の区別、内容理解に加え、噺家として技量、および、聞き手が感じた面白さについても比較をする主観評価実験も行なった。その結果、自然な落語音声を合成可能であるものの、プロの噺家とは差があり、その差を埋めるためには、抑揚の多様さや役の区別の仕方をさらに改善する必要がある事がわかった。また真打と２つ目・前座では役の区別の仕方が音響的に違う事も実験から判明した。第８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。

この様に本論文は、音声合成技術を情報伝達以外の用途に適用する事を目標に、落語音声合成という新たなタスクを提案し、データベースの構築、モデリング、評価方法の提案まで行い、その可能性を先駆的に示したものである。音声合成の学術分野を発展させる内容である事はもちろん、音声情報処理を利用したサービス等にも直結する内容であり、その貢献は大きいと言える。

博士論文審査の結果、申請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、ジャーナル論文１編、国際会議論文２編を出版（もしくは採択）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。