氏　　名　安田 裕介

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2241 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　Lexical pitch accent and duration modeling for neural end-to-end text-to-speech synthesis

論 文 審 査 委 員　主　査　山岸 順一
情報学専攻 教授
相澤 彰子
情報学専攻 教授
高須 淳宏
情報学専攻 教授
徳田 恵一
名古屋工業大学 情報工学専攻 教授
篠崎 隆宏
東京工業大学 工学院 准教授

（様式３）

博士論文の要旨

氏　　名　安田　裕介

論文題目 Lexical pitch accent and duration modeling for neural end-to-end text-to-speech synthesis

Text-to-speech synthesis (TTS) is a task to transform texts into speech. End-to-end TTS is one of the TTS frameworks, and its unique approach is characterized by using only a single model to generate speech directly from texts. This feature of end-to-end TTS makes contrast to conventional TTS framework called pipeline which consists of multiple models specialized to one function to realize TTS. End-to-end TTS reduces the burden of constructing TTS model without relying on complex labeling, and is therefore promising to expand its applicability to many languages. On the other hand, the end-to-end approach has limitation to learn every language or speech features implicitly.

This thesis focus on two lexical features of speech to improve end-to-end TTS: pitch accents and phoneme duration. These two features have been treated by dedicated models in traditional TTS frameworks. In contrast, existing end-to-end TTS methods do not model these features explicitly and try to capture these features implicitly without relying on any mechanism specialized for them.

In preliminary experiments, we show that some typical end-to-end TTS systems tend to produce flat pitch when they are applied to English, which is intonation languages in terms of prosodic description. In addition, end-to-end TTS systems fail to produce correct pitch accents without any supervision for pitch accents when they are applied to Japanese, which is pitch accent language in terms of prosodic description. We also show that soft-attention, which is used for alignments in most end-to-end TTS methods, is hard to avoid alignment errors which sometimes result in unacceptable quality of synthetic speech. These experimental results motivated us to focus on pitch accent and alignment modeling in end-to-end TTS.

This thesis proposes methods to incorporate a dedicated module to end-to-end TTS to model these two features by using latent variable. For pitch accents, we introduce pitch accent modeling to end-to-end TTS. We represent pitch accent latent variable in the form of abstract tone contour patterns, which makes the latent space discrete. To learn and model pitch accents, tone recognizer and tone predictor are introduced, and all components are optimized jointly. Our method enables to predict pitch accents without relying on labels during test phase. To enforce latent variable to represent pitch accent information, we use supervision with tone contour labels during training. In experiments, our method can generate moderately correct pitch accents without relying on tone contour labels during test phase.

For alignment modeling, we develop end-to-end TTS method using discrete latent alignments. As a first step, we handle alignments in acoustic frame level. We design monotonic alignment by using transition variable to avoid fatal alignment errors. The transition variable has two values, stay or proceed, which makes alignment structure monotonic. The whole TTS model can be optimized by maximizing marginal likelihood by marginalizing with respect to all possible alignments. Our experiments show that this method successfully eliminates fatal alignment errors, but it still gives other kinds of alignment errors such as overestimation of phoneme duration.

We advance the end-to-end TTS method incorporating alignment modeling by constructing discrete latent alignments based on phoneme duration. Phoneme duration ensures monotonic alignment structure, and it enables to model alignments more efficiently compared with alignments represented in acoustic frame level because phoneme duration sequences are much shorter than acoustic feature sequences. We represent phoneme duration as discrete symbols in the number of acoustic frames. Linguistic feature inputs can be aligned to acoustic feature outputs by upsampling them based on phoneme duration. A duration recognizer, duration predictor, and forced aligner are introduced to model phoneme duration, and all components are optimized jointly. In experiments, our TTS method using phoneme duration modeling shows higher naturalness of synthetic speech compared to existing TTS method using duration predictor.

We use an identical framework to handle pitch accents and phoneme duration. Both pitch accents and phoneme duration are lexical feature depending on linguistic units, and can be represented in discrete symbols. We thus use conditional VQ-VAE (vector quantized variational autoencoder) to model the two features as latent variable. We show both pitch accents and phoneme duration can be handled with the conditional VQ-VAE. We construct the pitch accent modeling by defining approximate posterior with tone contour pattern labels, using tone recognizer as encoder, and utilizing tone predictor as prior for VQ-VAE. We construct phoneme duration modeling by defining approximate posterior with forced aligner, using duration recognizer as encoder, and utilizing duration predictor as prior for VQ-VAE. There components have been used for conventional TTS methods consisting of multiple models dedicated to a specific functionality of TTS. Our method based on the conditional VQ-VAE enables for end-to-end TTS to incorporate the conventional modules designed for a specific feature of speech. This approach connects end-to-end TTS to conventional TTS methods to compensate its disadvantages by using established ways of the conventional TTS while keeping its advantage.

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　　安田 裕介

Title
論文題目　Lexical pitch accent and duration modeling for neural end-to-end text-to-speech synthesis

本学位論文は、文字系列から音声を生成する音声合成において広く利用されている Sequence-to-sequence 型ニューラルネットワークの二つの重要な問題に取り組んでいる。その一つの課題が、日本語や中国語と言ったピッチアクセント型言語にいかに適用するかという課題であり、二つ目の課題が、Sequence-to-sequence 型ニューラルネットワークに利用されるソフトアテンションネットワークのエラーをいかに原理的に改良するという課題である。

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章では従来のテキスト音声合成と end-to-end 音声合成の違い、および其々の代表的手法について概説している。その後、end-to-end 音声合成において通常利用される Sequence-to-sequence 型ニューラルネットワークの説明、および、その要の技術であるソフトアテンションネットワークについて概説している。

第３章では、Tacotron と呼ばれる RNN 型の Sequence-to-sequence ニューラルネットワークを様々な条件で学習し、十分なパラメータ数がある際は文字や音素系列から適切な発音を学習できるものの、音声の抑揚を表す基本周波数のモデリングには問題がある事、および、入力文字系列と出力音声系列の時間対応関係を表したソフトアテンションネットワークが致命的なエラーを起こし、このエラーを完全に無くす事が困難である事をそれぞれ実験から示した。１つ目の問題はピッチアクセントを言語的特徴として利用する日本語や中国語ではクリティカルな課題である。

そこで、第４章と５章では、ピッチアクセント情報を追加入力情報として与えることが可能な end-to-end 音声合成方式の提案、および、ピッチアクセント情報を潜在変数として扱うモデル化方式の提案を行い、その有効性を示した。具体的には、日本語のピッチアクセント情報(アクセント型)を追加入力情報として外部から与える事が可能な RNN 型の Sequence-to-sequence ニューラルネットワーク構造を提案し、適切なパラメータ数でモデル学習を行う事により、ピッチアクセント情報以外にも様々なアノテーションを行なった従来のパイプライン型のテキスト音声合成と同等もしくはそれ以上の品質の合成音声を実現できる事を示した。更に、離散型変分自己符号化ネットワークによりピッチアクセントを潜在変数化することで、ピッチアクセント情報を外部入力情報として与えなくても、言語的に正しいアクセント変化を伴った音声を生成する事が可能な新たな枠組みも提案した。

また、第６章と７章では、ソフトアテンションネットワークが致命的なエラーを起こしこれを完全に防ぐことが容易ではない事に着目し、ソフトアテンションを利用しない新しい sequence-to-sequence モデルを２種類提案している。第６章では、入力と出力系列の関

係を表したトレリス上の状態遷移を潜在変数として用いる sequence-to-sequence モデルを提案し、第７章では、各入力文字記号もしくは音素記号の継続長時間を離散潜在変数として学習する sequence-to-sequence モデルを提案し、それぞれの有効性および計算効率上の利点を示した。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、音声情報処理、特に音声合成の分野において、４つの手法の提案を行い、その有効性を示したものである。本論文の貢献は、機械学習を利用した音声合成分野を学術的に発展させる内容であると同時に、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、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、ジャーナル論文１編、国際会議論文３編を出版済みである。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。