| | |
|---|---|
| 氏　　名 | Xin Wang |
| 学位(専攻分野) | 博士(情報学) |
| 学 位 記 番 号 | 総研大甲第 2039 号 |
| 学位授与の日付 | 平成 30 年9月28日 |
| 学位授与の要件 | 複合科学研究科　情報学専攻<br>学位規則第6条第1項該当 |
| 学 位 論 文 題 目 | Fundamental Frequency Modeling for Neural-Network-Based Statistical Parametric Speech Synthesis |
| 論 文 審 査 委 員 | 主　査　准教授　山岸　順一<br>教授　越前　功<br>教授　宮尾　祐介<br>教授　峯松　信明<br>東京大学　工学系研究科<br>教授　徳田　恵一<br>名古屋工業大学 |

（様式３）

# 博士論文の要旨

氏　　名
Xin Wang

論文題目
Fundamental Frequency Modeling for Neural-Network-Based Statistical Parametric Speech Synthesis

The fundamental frequency (F0) of speech, which determines the perceived relative highness or lowness of the sound, plays an indispensable role in both the segmental and suprasegmental aspects of human languages. How to generate natural F0 contours from linguistic features of a text is one of the cruces in text-to-speech (TTS) systems, especially in the TTS system that has been able to synthesize speech with a natural segmental quality. A TTS system may produce unnatural speech if the generated F0 contour of that speech is incompatible with the prosodic system of the language (e.g., an incorrect F0 pattern in a tonal language), or if it is bland and monotonous (e.g., an over-smoothed F0 contour). Even though a TTS system correctly plans the prosodic structure and linguistic specification for the text to be uttered, it may generate an unnatural F0 contour when the internal F0 model fails to fulfill the plan. This problem is crucial in common TTS systems that use statistical F0 models, particularly using the so-called black-box neural networks. This thesis focuses on the neural-network-based F0 models for TTS systems, with the goal to identify potential limitations of conventional neural F0 models and propose better solutions. Specifically, this thesis treats F0 modeling as a sequential conversion problem where the input linguistic feature sequence is converted by a neural F0 model into an F0 contour frame by frame. Through interpreting and analyzing common neural F0 models or models that generate F0 with other spectral features, this thesis identifies three potential limitations: (1) whether it is appropriate to jointly model F0 and other spectral features using a normal neural network as many TTS back-ends do; (2) whether a normal neural F0 model ignores the temporal correlation of F0 contours. If yes, how a model can learn the correlation; (3) whether it is efficient for a neural F0 model to process linguistic features frame-by-frame. If not, how a more efficient and interpretable model can be designed. On issue (1), this thesis conducted experiments using highway neural networks and found that the network prioritized the high-dimensional spectral features over the F0. Analyses on the hidden features further illustrated different network behaviors in modeling the F0 and the spectral features. The results provide a rationale to separate F0 modeling from spectral feature modeling in order to improve F0 modeling performance. On issue (2), this thesis used random sampling to visualize

the limitation of the conventional neural F0 models such as the RNN. It then introduced the autoregressive (AR) dependency and defined a new model called shallow AR (SAR) model. Although the model definition is simple, this thesis gives two interpretations of the SAR, one based on the signal and filter and the other one based on the framework of normalizing flow. Interestingly, the first interpretation revealed the issue of model stability and motivated three methods to ensure the stability of the SAR. On the other hand, the second interpretation allowed us to extend the original SAR into a more general AR model using an invertible and long-term AR dependency function. This thesis further identified the limitation of the SAR and proposed a deep AR (DAR) model. This model uses the non-linear non-invertible transformation in the neural network to model the AR dependency. The basic idea is to feed back the previous F0 observation as the input to a uni-directional recurrent layer. To make the model practical, this thesis proposes quantized F0 representation, a hierarchical softmax output layer, and a data dropout strategy to train the DAR. As experiments showed, the DAR outperformed previous models and enabled random F0 contour sampling, which has never been achieved by other F0 models. On issue (3), this thesis borrowed the idea of variational auto-encoder (VAE) and decomposed a neural F0 model into an F0 contour coding part and a linguistic association part. The coding part efficiently represents the F0 contour of a linguistic unit using one codeword, and the association part directly links the F0 code space and the linguistic space for each linguistic unit. Experiments found that the VAE-based F0 model learned an interpretable F0 code space and achieved a better objective performance than the DAR even though the VAE-based model was smaller and faster. Although this thesis deals with the F0 and conducted experiments mainly on the English and Japanese data, it does not assume a specific linguistic theory about the F0. The proposed methods and models can hopefully be applied to other speech corpora and other acoustic feature sequences.

Results of the doctoral thesis screening

# 博士論文審査結果

Name in Full
氏　　名　Xin Wang

Title
論文題目　Fundamental Frequency Modeling for Neural-Network-Based Statistical Parametric Speech Synthesis

出願者は、音声の抑揚に相当する基本周波数(Fundamental frequency, F0)を、テキスト情報からニューラルネットワークで予測する時系列モデルの高精度化に関する研究を行い、その成果を博士論文としてまとめた。この基本周波数の予測は、テキストから音声波形を生成するテキスト音声合成における重要な機械学習課題の一つである。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章では基本周波数の概説および既存のモデル化手法について説明している。第３章では深層学習について説明している。第４章では、まず、音声合成で利用される音響モデルでは、基本周波数単独でモデル化を行うのが良いか、それとも他の音響特徴量と同時にモデル化するのが良いのかについて、Highway ネットワークの可視化等を通して分析を行い、基本周波数単独でモデル化を行う方が適切であることを示した。次に、第５章では、通常の RNN の出力系列は確率的に独立であるという問題に着目し、線形自己回帰モデルとリカレント型 mixture density network とを組み合わせた新たなニューラルネットワーク構造を提案し、F0 モデル化における有効性を実験から示した。また、提案モデルの信号処理的解釈についても示した。そして、第６章では上記モデルを更に、非線形自己回帰モデルへと拡張した Deep Autoregressive F0 モデルを提案し、その優位性を客観的および主観的実験から示した。また本モデルにおいてランダムサンプリングを行うことで、同じテキスト入力に対しても毎回異なる抑揚を生成可能であり、生成された抑揚はどれも自然に聞こえることも示した。第７章では、離散潜在変数を学習する VQ-VAE と言う教師なし学習手法と提案非線形自己回帰 F0 モデルとを組み合わせることで、音声のフレーム単位ではなく、音素やモーラ単位で効率よくモデルの学習および予測を実現できることを示した。また、他の非線形自己回帰モデルとの比較実験も行い、その優位性も示した。日本語女性話者の音声データベースを用いた実験では、最終的な予測モデルの相関係数は 0.91 となった。第８章では、以上の結果をまとめ、将来の課題について議論している。

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、査読付きジャーナル論文３編、国際会議論文６編（そのうち２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、本論文が学位の授与に値すると判断した。