学位報告４

報告番号	※甲　　　第　　　号

主論文の要旨

論文題目　Incorporating Prior Knowledge on Speech Production Mechanism into Neural Speech Waveform Generation
（深層音声波形生成における音声生成過程に関する事前知識の導入）

氏　　名　　呉　宜樵　（WU, YI-CHIAO）

論文内容の要旨

Speech generation techniques including text-to-speech (TTS), speech enhancement, and voice conversion ones are widely applied to current daily applications such as a personal mobile assistant and car navigation. The naturalness of synthesized speech and the flexibility of acoustic controllability are the main challenges of speech generation. That is, high-fidelity synthesized speech sounding like natural speech and speech components being flexibly manipulated are important to a speech generation system. A speaker voice conversion (VC) task is adopted in this thesis as a paradigm of speech generation systems, and the VC task involves converting the speaker identity of input speech to a specific target speaker while keeping the same speech content. A general VC system is composed of analysis, manipulation, and synthesis modules, and the thesis focuses on improving the synthesis module using prior knowledge of speech.

A baseline VC system for the non-parallel VC (SPOKE) task of voice conversion challenge 2018 (VCC2018) has been established in this study. The analysis module of the VC system parameterizes speech into spectral and prosodic features using the WORLD vocoder, which is a conventional source-filter-based vocoder. Since the training corpus is non-parallel, the speech contents of the source and target utterances of the SPOKE task are different. A two-stage spectral conversion model with TTS-generated reference speech has been adopted to map the non-parallel source and target utterances. In contrast to conventional VC systems, the baseline system replaces the synthesizer of the conventional vocoder with a neural-based speech generation model, WaveNet (WN). The WN as a vocoder directly transfers the converted acoustic features to speech waveforms without many ad˜hoc designs of speech production imposed on the conventional vocoder. Both the internal and external evaluation results in this thesis show the better performance of the WN vocoder than the conventional vocoder.

However, because of the data-driven nature, generic network architecture, and lack of speech-related prior knowledge, the WN vocoder sometimes generates unexpected outputs such as non-speechlike noise while the input acoustic features are unseen or distorted. To avoid the collapsed speech problem of the WN vocoder, a collapsed speech detection and suppression approach has been studied in this thesis. The method is based on the prior knowledge of speech continuity and the stability of conventional vocoders. Specifically, although the naturalness of the WORLD-generated speech is worse, the speech is more stable than the WN-generated speech. The proposed detection method segmentally compares the waveform envelope difference between the WORLD- and WN-generated utterances to detect the collapsed speech segments. The WN vocoder regenerates the detected segments with a waveform-based constraint derived from the continuity extracted from the WORLD-generated speech.

On the other hand, because of the implicit pitch modeling of the WN vocoder, the lack of pitch controllability is a problem. That is, regardless of whether the input fundamental frequency feature F_0 is scaled or not in the F_0 range of the training data, the WN vocoder usually cannot generate speech with accurate pitches. To improve the pitch controllability of the WN vocoder, which is an essential vocoder feature, a pitch-dependent dilated convolutional neural network (PDCNN) and a quasi-periodic (QP) structure have been studied in this thesis. The PDCNN introduces the prior periodicity knowledge of speech to the WN vocoder for dynamically adapting the network architecture according to the input F_0. The QP structure based on the prior knowledge of speech production applies a source-filter-like structure to the WN vocoder for modeling pitch- and spectral-related components. With the PDCNN and QP structure, QPNet has been proposed, which markedly improves the pitch controllability and speech modeling efficiency of the WN vocoder.

Furthermore, to achieve real-time generation, a non-autoregressive neural-based speech generation model, parallel WaveGAN (PWG), with a compact network has also been studied in this thesis. Since the training process and network designs of the PWG are similar to those of the WN, the PWG also suffers from the difficulty of pitch controllability. With the proposed PDCNN and QP structure, the proposed QPPWG also markedly improves the pitch controllability and speech modeling efficiency of the PWG. Because of the direct waveform output, the internal generative mechanisms are easily revealed by the intermediate outputs of the PWG and QPPWG. The visualized results confirm the effectiveness of the QP structure to make the QPPWG like a source-filter model with a unified neural network. That is, the QPPWG simultaneously attains high-fidelity speech generation as the PWG with better tractability and interpretability.

To summarize, the studies show that applying the speech-related prior knowledge to the neural-based speech generation models significantly improves the robustness against the distorted and unseen acoustic features, pitch controllability, and speech modeling efficiency of these speech generation models.