【主論文の要旨・記入例】

学位報告 4

報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　High-Quality and Flexible Voice Conversion Techniques based on Statistical Spectral and Waveform Modeling
（統計的スペクトル変換・波形生成モデルに基づく高品質かつ柔軟な音声変換技術）

氏　　名　Patrick Lumbantobing

論 文 内 容 の 要 旨

This thesis presents an in-depth study for high-quality and flexible voice conversion (VC) using statistical spectral and waveform modeling techniques. Using VC, voice characteristics of a source speaker can be transformed into that of a target speaker while preserving the linguistic contents. To develop a VC, representations of vocal-tract resonance characteristics of the source speaker, such as vocal-tract spectrum parameterizations, have to be converted into that of the target speaker. Along with the source-excitation (pitch) characteristics, the converted speech waveform is generated using the transformed speech parameterizations. A high-quality and flexible VC can be beneficial for many speech applications in research and in daily life. In this work, the high-quality aspect is examined through the use of data-driven statistical waveform modeling, whereas the flexible aspect is examined through the use of data-driven statistical spectral modeling.

The statistical spectral modeling is developed to model the mapping function of the vocal-tract spectrum parameterizations, e.g., spectral envelope parameters, between the source and the target speakers. Another possible representation of resonance characteristics modeling is by the means of physical approach, such as through the use of articulatory (speech organs) configurations. The latter has the advantage of being more flexible in terms of direct control. Whereas, the former has the advantage on being more flexible in terms of system development, owing to the more straightforward methods to obtain spectral envelope parameters. On the other hand, the statistical waveform modeling is developed to model the speech waveform signal generation from speech parameterizations, such as spectral envelope and excitation parameters. Another way to generate speech waveform is to employ conventional rule-based approach for the generat

ion procedure based on the source-filter theory of speech production (vocoder). The latter has the advantage of not requiring any model training, though the quality is limited. On the other hand, the former has the advantage of being able to produce natural sounding synthetic speech, thanks to the use of data-driven approach, such as with neural network model, i.e., neural vocoder. In this thesis, the high-quality aspect is emphasized by the usage of neural vocoder in VC, while the flexibility aspect is emphasized on the development of statistical model for spectral mapping.

To achieve the goal of this thesis, four main frameworks are elaborated, which correspond to the contribution for high-quality and flexible VC. First is the development of voice modification with articulatory manipulation, which enables flexible control of speech sounds by means of the intuitive representations of articulatory information. Second is the development of VC with neural-network (NN)-based spectral and waveform modeling, which uses spectral envelope parameters as vocal-tract representations. Third is the improvement of NN-based VC framework to achieve high-quality converted speech by performing fine-tuning of neural vocoder. Finally, in the fourth system, the flexibility in the development of VC is achieved by means of nonparallel spectral mapping model framework, which does not require parallel (paired) data between source and target speakers.

In order to develop a voice modification with articulatory manipulation system, Gaussian mixture model (GMM)-based statistical modeling is employed to perform both of the acoustic-to-articulatory (inversion) mapping and the articulatory-to-acoustic (production) mapping. A sequential mapping procedure between inversion and production mappings is developed to enable the manipulation of intermediate articulatory representations for performing speech modification. To yield higher quality of modified speech, the vocoder-based excitation generation is avoided through direct filtering of input speech waveform with the use of differential spectrum that is calculated between the input and the modified spectral parameters. The experimental results demonstrate that the system is capable of producing modified vowel sounds by manipulation of tongue positions, and the use of direct waveform modification yields significant quality improvements for varying modification of articulation efforts, i.e., hypo- and hyper-articulations.

In the second system, a VC framework is developed by means of NN-based modeling for spectral mapping. Further, an NN-based modeling is also used to directly model the speech waveform signal by conditioning on spectral and excitation parameters, such as a WaveNet vocoder. However, owing to the use of converted spectral parameters instead of natural parameters in the synthesis time, there exist mismatches between the spectral and the waveform modeling. To reduce these mismatches, a postprocessing method based on the direct waveform modification is used to obtain refined converted spectral envelope parameters. The experimental results demonstrate that the NN-based VC is capable of a

chieving higher quality and speaker similarity in cross-gender conversion compa red to using conventional vocoder-based excitation generation and achieving hig her speaker similarity in same-gender conversion compared to using direct wav eform modification method.

In the third system, the NN-based VC is improved through the use of recu rrent neural network (RNN)-based architecture for spectral mapping modeling a nd the finetuning of the WaveNet vocoder. This is because the use of postproc essing method does not directly address the mismatches within the WaveNet v ocoder. However, it is not straightforward to achieve fine-tuning of WaveNet vo coder in VC due to the difference of temporal structure between source and ta rget speakers. In other words, the converted spectral parameters from the sour ce speaker cannot be used to fine-tune a WaveNet vocoder using the target sp eech waveform. To perform WaveNet finetuning in VC, a cyclic RNN structure i s introduced, which can produce both of the converted source spectral paramet ers and the appraisal of estimated (oversmoothed) target spectral parameters. T he oversmoothed target spectra obtained from the cyclic flow is used for Wave Net fine-tuning. The experimental results demonstrate that the CycleRNN-based spectral mapping model makes it possible to perform proper WaveNet fine-tuni ng in VC by significantly improving the quality and the speaker-similarity of th e converted speech compared to the previous VC framework.

Finally, in the fourth system, the flexibility for developing voice conversion is achieved by means of nonparallel spectral modeling using variational autoen coder (VAE) framework. In the previous two systems, the statistical spectral mo dels are developed with a parallel dataset between source and target speakers where they utter a same set of sentences. However, nonparallel speech datasets of source and target are more practical to be obtained, especially for speakers with different language. To achieve that, a VAE-based VC is employed, where t he shared characteristics between speakers, such as phonetics, are to be captur ed within a latent space, and the speaker-dependent characteristics are to be d etermined by time-invariant speaker-coding features. However, owing to the ina bility to explicitly optimize converted spectral features (only reconstruction is c onsidered), the performance of VAE-based VC is limited. To improve the VAE-b ased VC, a cyclic flow is introduced to recycle the converted spectra back into the system to obtain cyclic reconstructed spectra that can be directly optimized. The experimental results demonstrate that the CycleVAE-based VC gives signifi cant improvements in quality and speaker-similarity of converted speech, especi ally for cross-gender conversions, as well as improvements in the disentanglem ent of speaker-independent and speaker-dependent traits from the latent space.

In summary, in this thesis, an investigation for achieving high-quality and f lexible voice conversion system is conducted through the aforementioned four f rameworks. The possibility of having flexible control is investigated through the development of articulatory controllable speech modification system, where spe ech sounds can be intuitively modified by manipulating the intermediate articul

atory representations. To achieve VC in this way, in the future, it is necessary to model the physical vocal-tract shape with articulatory representations. Next, flexibility in the development of VC system is utilized through the use of NN-based spectral mapping model with parallel (paired) data, where spectral envelope parameters are more straightforward to be obtained compared to articulatory data. Then, the use of neural vocoder in VC is thoroughly investigated to achieve high-quality output, where the mismatches of spectral parameters are dealt with a spectral postprocessing method or with neural vocoder fine-tuning. Finally, to achieve flexible VC system for practical real-world applications, a nonparallel spectral modeling, which utilizes latent space to capture shared traits between different speakers, is presented. The last approach can be easily extended with the neural vocoder fine-tuning approach to ultimately achieve high-quality and flexible VC.