	ヨシムラ　タケノリ		
氏　名	吉村　建慶		
学位の種類	博士（工学）		
学位記番号	博第1139号		
学位授与の日付	平成30年9月5日		
学位授与の条件	学位規則第４条第１項該当　課程博士		
学位論文題目	ACOUSTIC AND WAVEFORM MODELING FOR STATISTICAL SPEECH SYNTHESIS （統計的音声合成のための音響・波形モデリング）		
論文審査委員	主査	教授	徳田　恵一
		教授	竹内　一郎
		教授	本谷　秀堅
		教授	李　晃伸
		准教授	南角　吉彦

論文内容の要旨

Text-to-speech (TTS) is the artificial generation of human speech from texts. Since speech is one of the most important ways of human communication, TTS systems have been widely used in many applications, e.g., bus announcements, smartphone applications, smart speakers, and speech-to-speech translation systems. With the spread of services using TTS systems, they are expected to generate synthetic speech with not only high naturalness but also various speaker characteristics, emotions, and speaking styles.

To achieve that goal, many researchers have been tackled the issues of TTS over the past decades. One of the most successful approaches is corpus-based statistical parametric speech synthesis (SPSS). In this approach, the relationship between the linguistic features extracted from texts and the acoustic features extracted from the corresponding speech waveforms is modeled by a statistical model. The major advantage is that the characteristics of synthetic speech can be easily controlled by manipulating the parameters of the statistical model. As the statistical model, the hidden Markov model (HMM) has been widely used thanks to the well-defined algorithms and its flexibility for modeling sequential data. One of the major problems of the HMM is that it cannot fully exploit large collections of heterogeneous speech data. In order to solve the problem, the factor analyzed HMM (FAHMM), which is a probabilistic version of eigenvoice, has been proposed.

In the framework of the FAHMM, the speech characteristics can be altered by changing a low-dimensional variable, which is called factor, rather than very high-dimensional model parameters. It should be noted that the factor is automatically extracted through model training of the FAHMM. While the effectiveness of the SPSS methods has been shown in various experiments, novel deep-neural-network (DNN)-based TTS approaches have been recently proposed. They attempt to directly model the relationship between the linguistic features and the raw audio speech waveforms using a specially-designed neural network architecture. One of the most successful approaches is the WaveNet generative model. By using a stack of convolutional neural networks (CNNs), the WaveNet can capture long-term temporal dependencies in speech signals. The WaveNet model outperformed the state-of-the-art TTS systems in subjective evaluation tests.

In the paper, techniques for improving the naturalness of synthetic speech and controlling the speech characteristics are proposed. For controlling the speech characteristics, the model structure to appropriately model heterogeneous speech data is studied using the FAHMM. Although the model structure of the FAHMM has no constraints under its framework, a single binary decision tree is typically used so that a simple algorithm for building a model structure can be used. However, it would prevent to capture the complex speaker/emotion/speaking-style-dependent relationship between the linguistic and acoustic features. To flexibly model the relationship, a more complex multiple-tree structure is proposed. The multiple trees are grown simultaneously rather than sequentially because building trees one by one makes it difficult to find the optimal model structure. However, since the possible combinations of tree structures exponentially increase according to growing trees, simultaneously optimizing the model structures is computationally infeasible. In order to avoid the computational explosion of the optimization, two computational complexity reduction algorithms inspired by techniques used in HMM-based speech synthesis are introduced.

For improving the naturalness of synthetic speech, the quantization used in neural-network-based speech waveform synthesis is investigated. One of the key techniques of neural-network-based speech waveform synthesis is modeling speech signals composed of a set of discrete values instead of continuous ones. In other words, speech waveform modeling is formulated as a classification problem rather than as a regression one. This enables more flexible waveform modeling because a categorical distribution has no assumptions about the shape. Simple linear quantization or nonlinear quantization with $¥mu$-law companding is typically used to obtain the discrete-valued speech signals. However, the quantization scheme introduces white noise into the original signals. Since the white noise is uniformly distributed over the entire frequency range, the quantization noise is easily perceived by human listeners. This paper presents a quantization noise shaping method in which a time-variant mask derived from mel-cepstrum is applied to the white quantization noise. Since mel-cepstrum can be based on the human auditory system, some of the quantization noise should be difficult for a human listener to perceive.

論文審査結果の要旨

論文の専攻内審査会を5月21日、公聴会を8月21日に行った。

論文のテーマは、与えられた任意のテキストに対応する音声を機械によって生成する音声合成に関するものである。音声合成においては、人間らしい自然な音声を合成できることと、個人性や感情・発話スタイルを柔軟に表現できることが特に求められている。申請者は、音声合成におけるこれらの大きな２つの課題について取り組み、その成果を論文にまとめ発表を行った。

音声合成の自然性向上に対しては、音声波形を直接モデル化する際に利用される量子化法に着目して、メルケプストラムに基づくノイズシェーピング量子化法を提案した。また、柔軟性向上に対しては、因子分析に基づく隠れマルコフモデルに着目して、基底クラスタリングという技術を提案した。いずれも研究内容の新規性があり、十分な成果があることが示された。

発表資料は要点を得た簡潔な作りとなっており、分かりやすい発表が行われた。審査員から様々な質疑があったが、いずれも的確に応答しており、優れたプレゼンテーション能力をもつことが示された。

以上のことから、論文審査の結果としては合とし、学位授与の可否としては可とすることが適当である。