	ナカムラ　カズヒロ	
氏　　名	**中村　和寛**	
学位の種類	博士（工学）	
学位記番号	博第966号	
学位授与の日付	平成26年12月17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	STATISTICAL APPROACH TO SPEECH AND SINGING VOICE SYNTHESIS （音声合成・歌声合成のための統計的アプローチ）	
論文審査委員	主　査	准教授　李　　晃伸
		教　授　北村　正
		准教授　竹内　一郎
		准教授　南角　吉彦

論文内容の要旨

Speech is the most important way for human communication and is expected to be used as a new human-machine communication interface with the emergence of miniature and speedy computers. Speech synthesis is one of the core technologies of the speech communication interface, and it is used for text-to-speech (TTS), singing voice synthesis, speech translation, speech dialogue, etc. State-of-the-art research on speech synthesis is based on hidden Markov models (HMMs). HMMs are statistical models that are widely used for speech recognition by well-defined algorithms to estimate model parameters. It is well known that they have high recognition performance if they are trained with enough data. HMM-based speech synthesis has also grown in popularity over the last several years. This framework makes it possible to model different voice characteristics, speaking styles, or emotions without recording a large speech database. Although the quality of the synthesized voices is high enough in some cases, the core technology needs to be improved to synthesize high quality voices about the same as human voices. Additionally, the applications of speech synthesis should be used to make our life more convenient. In particular, some important functions, such as

multilingualization and emotional synthesis are required by many applications. In this paper, I improve speech synthesis technology from both sides, i.e. the core technology and the applications.

First, to improve the core technology, I propose a novel approach to integrate spectral feature extraction and acoustic modeling for HMM-based speech synthesis. The statistical modeling process of speech waveforms is typically divided into two component modules: the frame-by-frame feature extraction module and the acoustic modeling module. In the feature extraction module, the statistical mel-cepstral analysis technique has been used, and the objective function is the likelihood of mel-cepstral coefficients for given speech waveforms. In the acoustic modeling module, the objective function is the likelihood of model parameters for given mel-cepstral coefficients. It is important to improve the performance of each component module in order to achieve higher quality synthesized speech. However, the final objective of speech synthesis systems is to generate natural speech waveforms from given text, and improving each component module does not always lead to an improvement in the quality of synthesized speech. Therefore, ideally, all objective functions should be optimized on the basis of an integrated criterion that well represents the subjective speech quality of human perception. In this paper, I propose an approach to model speech waveforms directly and optimize the final objective function. Experimental results show that the proposed method outperformed the conventional methods in objective and subjective measures.

Next, I propose a mel-cepstral analysis technique that restores missing high frequency components from low-sampling-rate speech. In HMM-based speech synthesis, the samplingrate of the synthesized speech depends on that of training speech data. Low-sampling-rate training speech data degrades the quality of the synthesized speech. Recently, speech databases have come to be recorded at a high sampling rate, e.g., 48 kHz. The sampling rates of many speech databases recorded in the past are low. With the popularization of speech synthesis techniques, the demand for using databases recorded in the past is growing bigger. Additionally, in some cases, such as speaker adaptive training (SAT), which trains a model with speech data uttered by different speakers, the amount of the training data can be increased significantly by using speech databases recorded at different sampling rates. Therefore, I train a model of speech waveforms from a high sampling-rate speech database in advance and use it for analyzing mel-cepstral coefficients whose high frequency components are restored from

low-sampling-rate speech databases. Experimental results show that the proposed method restored high frequency components and improved the quality of the synthesized speech.

Finally, as an important function for the applications, the multilingualization for HMMbased singing voice synthesis is attempted. An English singing voice synthesis system is proposed and compared with the Japanese one. In this approach, the spectrum, excitation, and vibrato of singing voices are simultaneously modeled by using context-dependent HMMs, and waveforms are generated from HMMs themselves. Japanese singing voice synthesis systems have already been developed and used to create variable musical content. To expand this system to English, contexts that can be used in Japanese and English singing voice synthesis systems are designed. Furthermore, methods for matching musical notes and the pronunciation of English lyrics are proposed and evaluated in subjective experiments. Then, Japanese and English singing voice synthesis systems are compared.

As described above, in this paper, I propose a core technology and an application for HMM-based speech synthesis, and they are evaluated in objective and subjective experiments.

論文審査結果の要旨

統計モデルに基づく音声合成の普及を目指し，コア技術の更なる改善とアプリケーションとしての機能の充実という両方の観点から検討が行われている．音声合成とは機械に喋らせたり歌わせたりする技術の総称であり，近年の音声合成研究分野では，隠れマルコフモデル（HMM）に基づく手法が主流となりつつある．本論文では，HMM音声合成のコア技術，そして応用技術に関して研究が行われている．

まず，コア技術の更なる改善として，合成音声の音質改善のためにメルケプストラム分析と音響モデリングの統合手法が提案されている．近年では様々な分野で，複数の統計モデルを組み合わせることで複雑なシステムが構成されている．HMM音声合成においても，音声から特徴量を抽出するメルケプストラム分析と，特徴量を統計モデル化する音響モデリングは，それぞれが統計モデルで表現され，別々の目的関数の最適化問題として定義されていた．本論文で提案されている統合モデリングの枠組みでは，それらのモデルを統合して音声を1つの統計モデルとして表現することで本来の目的関数を最適化しており，音声の自然性に対する主観評価実験により，著しい音質改善を達成している．

次に、統合モデリングの応用として，低周波数標本化音声データの高帯域成分の復元を考慮したメルケプストラム分析手法が提案されている．統計的音声合成においては，高音質な大量の学習データを用意することで高音質な音声を合成することが可能である．しかし，高音質な音声を収録し，音声合成のためのラベルを付与する作業は多大なコストを伴うため，既存のデータを有効活用する枠組みが求められている．本論文では，合成音声の音質を左右する要因として標本化周波数に着目し，低周波数標本化音声の失われた高帯域成分を統計モデルにより復元してHMMの学習に用いることで，高周波数標本化音声を合成する手法が提案されている．合成音声の自然性に対する主観評価実験では，提案法は目標とする高周波数標本化音声と同等の非常に高い性能を示している．

次に，アプリケーションとしての機能の充実という観点から，HMMに基づく英語歌声合成が提案されている．HMM歌声合成は，楽譜と歌声の関係をHMMでモデル化し，任意の楽譜が与えられたときにHMMから歌声を生成するシステムであり，これまでに日本語の楽譜に対応したシステムが提案されていた．本論文では，歌声合成技術の海外展開の第一歩として，英語の歌声合成システムが提案さていている．日本語楽譜では歌詞が「かな」で記述されるのに対し，英語楽譜では歌詞が単語で記述されており，複数の音符に跨って記述されることも多いため，音符に対する発音の割り当て方が問題となるが，提案手法は最大で92%の割り当て誤り削減率を達成している．更に，今後の多言語化を見越したコンテキスト設計が提案されている．

以上のように，本論文では音声合成のコア技術の更なる改善として統合モデリング手法が，アプリケーションのための機能の充実として英語歌声合成が提案されており，実験により有効性が示されている．論文の内容は国内外の論文誌・国際学会にて公表されおり，また，アプリケーションとしても一般に公開されていることから，音声研究分野並びに社会への貢献度は高く，本研究は博士論文として十分な価値を持つものと認める．