A STATISTICAL APPROACH TO SPEECH SYNTHESIS AND IMAGE RECOGNITION BASED ON HIDDEN MARKOV MODELS

氏名	サワダ　ケイ **沢田　慶**
学位の種類	博士（工学）
学位記番号	博第1129号
学位授与の日付	平成30年3月26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	A STATISTICAL APPROACH TO SPEECH SYNTHESIS AND IMAGE RECOGNITION BASED ON HIDDEN MARKOV MODELS （隠れマルコフモデルに基づく音声合成と画像認識のための統計的アプローチ）
論文審査委員	主査　准教授　南角　吉彦 教授　徳田　恵一 教授　李　晃伸 教授　本谷　秀堅

論文内容の要旨

We human beings communicate with others by transmitting auditory and visual information. Recently, since information technology has steadily improved, not only communication between humans, but also communication between humans and computers has become feasible. In consequence, new services such as spoken dialogue system and biometrics authentication system have been developed. In order to realize communication between humans and computers, the computer need to interpret auditory and visual information. Speech recognition and speech synthesis are used as techniques for processing auditory information, and image recognition and image synthesis are used as techniques for processing visual information. Improvements in these techniques are necessary for smooth communication between humans and computers.

Hidden Markov models (HMM)-based speech recognition and speech synthesis have been proposed as a standard framework. HMMs are one of widely used statistical models for representing time series by well-defined algorithms. Additionally, two-dimensional data such as pixel values of image can be modeled by extending the HMM to two dimensions. In this paper, I propose speech synthesis and image recognition based on HMMs to realize smooth communication between humans and computers. Especially, for widening the communication, I investigate highly versatile construction methods from low-resource data.

For speech synthesis, I propose a method for constructing text-to-speech (TTS) systems for languages with unknown pronunciations. There are thousands of active written languages in the world. However, conventional methods of constructing corpus-based TTS systems for a new language not only require preparation of training corpus but also require language-specific knowledge. Especially, to marshal language-specific knowledge about pronunciation for each new language requires high cost. Therefore, a goal of the speech synthesis research is to establish a language-independent framework that can be used to construct TTS systems for any written language. To address this problem, I investigate a framework for automatically constructing a TTS system from a target language database consisting of only speech data and corresponding Unicode texts. In the proposed method, pseudo phonetic information of the target language with unknown pronunciation is obtained by a speech recognizer of a rich-resource proxy language. Then, a grapheme-to-phoneme converter and a statistical parametric speech synthesizer are constructed based on the obtained pseudo phonetic information. With these processes, it becomes possible to construct a TTS system automatically without specific knowledge on the target language.

For image recognition, I propose an image recognition method based on hidden Markov eigen-image models (HMEMs) using a Bayesian framework. The geometric variations of the object to be recognized, e.g., size, location, and rotation, are an essential problem in image recognition. Separable lattice hidden Markov models (SL-HMMs), which have been proposed to reduce the effect of geometric variations, can perform elastic matching both horizontally and vertically. However, SL-HMMs still have a limitation in that the images are assumed to be generated independently from corresponding HMM states. It is insufficient to represent variations in images, e.g., lighting conditions and object deformation. To overcome this problem, HMEMs have been proposed in which the structure of factor analysis (FA) or probabilistic principal component analysis (PPCA) is integrated into SL-HMMs. HMEMs have good properties of both SL-HMMs and FA/PPCA: invariances to the size and location of objects to be recognized and a linear feature extraction. In some image recognition tasks, it is difficult to acquire sufficient training data. Additionally, models with a complex structure such as HMEMs suffer from the over-fitting problem, especially when there is insufficient training data. This study aims to accurately estimate HMEMs using the variational Bayesian (VB) method. The VB method can utilize prior distributions representing useful prior information and is expected to have a high generalization ability due to the marginalization of model parameters. Furthermore, to relax the local maximum problem in the VB method, the deterministic annealing expectation maximization algorithm is applied to train HMEMs. Experiments on face recognition indicated that the proposed method offers a significantly improved image recognition performance.

As described above, in this paper, I propose a statistical approach to speech synthesis and image recognition based on HMMs, and they are evaluated in experiments.

論文審査結果の要旨

人とコンピュータとのコミュニケーションを実現するためには，コンピュータが聴覚・視覚情報等を解釈する必要がある．そこで，本研究では，人とコンピュータとの円滑なコミュニケーションを目指した，統計的アプローチに基づく音声合成と画像認識の検討がされている．

近年では，大量の学習データを用いた統計的アプローチは様々な分野で高い性能が報告されている．しかし，統計モデル学習に用いるための大量の音声や画像のデータを収集することが難しい場合は多く存在する．このような条件においても，高性能な音声合成・画像認識システムの構築を目指した，隠れマルコフモデル（hidden Markov model; HMM）に基づく手法を提案している．

まず，音声合成においては，発音情報が未知の言語におけるテキスト音声合成システムの構築法が提案されている．世界には数千におよぶ書記言語が存在すると考えられており，あらゆる書記言語のテキスト音声合成（text-to-speech; TTS）システムを構築することは，音声合成研究の1つのゴールである．しかし，一般的なTTSシステム構築法は，目的とする言語に関する専門的な知識を用いた人手による作業を必要とし，言語ごとに高い構築コストがかかる．そこで，提案法では発音情報が未知である言語の音声データとUnicodeテキストのみから構成されるデータベースから，言語に関する専門的な知識を利用せずにTTSシステムを自動構築する手法について検討している．提案法では，発音情報が未知であるターゲット言語の発音情報を代理言語の音声認識により獲得している．そして，疑似発音情報に基づき書記素音素変換器と統計的音声合成器を構築している．これにより，ターゲット言語固有の知識を利用することなくTTSシステムを構築することを可能とした．

次に，画像認識においては，ベイズ基準に基づく可変固有画像モデル（hidden Markov eigen-image models; HMEM）を提案している．画像認識において，認識対象の位置や大きさなどの幾何学的変動に対応可能な分離型格子HMMに固有画像のような主成分分析の構造を組み込んだHMEMが提案されている．従来，HMEMの学習には尤度最大化基準が用いられてきた．しかし，画像認識では十分な量の学習データを用いることが困難である場合も多く，このような場合に，尤度最大化基準によりHMEMのような複雑なモデル構造を学習すると過学習を起こす恐れがある．そこで，提案法では，ベイズ基準に基づく高精度なHMEMの学習を用いている．ベイズ基準は，事前情報を事前分布として用いて事後分布を推定することで過学習を緩和している．さらに，確定的アニーリング期待値最大化アルゴリズムを導入することで，初期値に依存した局所最適解の問題を克服している．

以上のように，本論文では，人とコンピュータのコミュニケーションの実現のために，HMMに基づく高性能な音声合成システムと画像認識システムを構築するための統計的アプローチを提案し，その有効性を示した．また，本論文の内容は論文誌や国際学会にて公表されている．よって，本研究は該当分野において寄与するところが多大にあり，博士論文として充分価値のあるものと認める．