氏名	シオタ　サヤカ **塩田　さやか**
学位の種類	博士（工学）
学位記番号	博第840号
学位授与の日付	平成24年3月23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 MODELS BASED ON MULTIPLE MODEL STRUCTURES FOR SPEECH RECOGNITION （複数のモデル構造を用いた統計モデルによる音声認識）
論文審査委員	主査　教授　徳田恵一 教授　北村正 教授　松尾啓志 准教授　李晃伸

論文内容の要旨

The topics of automatic speech recognition (ASR) has been active areas of research focus. Hidden Markov models (HMMs) are one of widely used statistical models for representing time series by well-defined algorithms. They have successfully been applied to acoustic modeling in speech recognition when training data can be sufficiently available. However, it is difficult to obtain a large number of clean training data (e.g., clean voice data, correct text, and correct time alignment). Recently, although I can easily obtain large and many databases from the Internet, the databases contain noises or miss transcriptions and the qualigy are low. Thus, acoustic modeling technique without considering a quality of given data is important for improving speech recognition performance. In this paper, frameworks of improving acoustic modeling were proposed for HMM-based speech recognition.

First, I propose a simultaneously optimization of model structure and model parameters. In the use of context-dependent models, decision-tree-based context clustering is applied to find an appropriate parameter tying structure. However, context clustering is usually performed on the basis of unreliable statistics of hidden Markov model (HMM) state sequences because the estimation of reliable state sequences requires an appropriate model structures, that cannot be obtained prior to context clustering. Therefore, context clustering and the estimation of state sequences essentially cannot be performed independently. To overcome this problem, I propose an optimization technique of state sequences based on an annealing process using

multiple decision trees. In this technique, a new likelihood function is defined in order to treat multiple model structures, and the deterministic annealing expectation maximization (DAEM) algorithm is used as the training algorithm. Speech recognition experiments show that the proposed method achieved a higher performance than the conventional methods.

Next, training criterion has been focused. The maximum likelihood (ML) criterion has usually been used for training statistical models for HMM-based speech recognition systems. However, since the ML criterion produces a point estimate of model parameters, the estimation accuracy may degrade when little training data is available. The Bayesian method is a statistical technique for estimating reliable predictive distributions by marginalizing model parameters, and it can accurately estimate observation distributions even if the amount of training data is small. However, the local maxima problem in the Bayesian method is more serious than in the ML-based approach, because the Bayesian method treats not only state sequences but also model parameters as latent variables. The deterministic annealing EM (DAEM) algorithm has been proposed to improve the local maxima problem in the EM algorithm, and its effectiveness has been reported in HMM-based speech recognition using ML criterion. In this paper, the DAEM algorithm is applied to Bayesian speech recognition to relax the local maxima problem.

In speech recognition based on generative models, there are many efforts to find appropriate model structures to predict observation vector sequences (e.g., multi-mixture models, clustering techniques and more complicated models). Even though a better prediction obtained by these methods leads to improve recognition performance, they still aim to find only one model structure. However, in most practical cases, it is insufficient to represent a true model distribution using only ``one" model structure, because a family of such models usually does not include a true distribution. Therefore, it is necessary to increase model complexity efficiently without inaccurate estimation cased by the over-fitting problem. Thus, I focuses on model structure integration based on the Bayesian framework. In the previous work, I proposed the marginalization of model parameters based on the Bayesian framework. Next, the model structures should be marginalized. Therefore, I proposed a new likelihood function for using multiple model structures. Since the proposed framework is regard model structures as a latent variables, the local maxima problem is caused. The basic idea of the Bayesian approach is to treat all parameters as random variables. Therefore, I proposed a novel framework of using multiple model structures based on the Bayesian framework. The conventional VB method sometimes suffers from the local maxima problem, because the conventional VB method treats not only state sequences but also model parameters as latent variables, that makes the estimation problem complicated. To overcome this problem, I have proposed the training algorithm applying the deterministic annealing framework to the Bayesian speech recognition, and reported the effectiveness for the local maxima problem. Since the proposed technique also treats the multiple model structures as a latent variable, the local maxima problem is more serious than in the conventional VB method. Therefore, the DAEM algorithm is applied to the proposed technique as a training algorithm. The proposed method can consistently perform model estimation and model selection based on the VB method.

論文審査結果の要旨

近年，音声を情報伝達の手段としたシステムの需要が高まっており，音声認識・音声合成などの音声に関する研究が盛んに行われている．音声認識における代表的な枠組みとして，音響モデルに統計モデルの一種である隠れマルコフモデル(Hidden Markov Model; HMM)を用いる枠組みがある．HMMはモデルの推定に十分な学習データ量が与えられれば高い認識性能を示すことが知られている．しかし，高精度な音響モデルを学習するためには学習データとして雑音や言い間違いなどが存在しない音声データと音声データと対になる発話内容のテキストデータが必要となる．高精度な音響モデルを構築するための最適なデータを揃えることは困難であり，テキスト情報が不正確なデータから汎化性能の高い音響モデルを推定することは重要な課題であると言える．そこで，本論文はテキスト情報の精度に依存しにくい音響モデルの性能を向上する枠組みを提案する．

まず，音響モデルのモデル構造とモデルパラメータ推定の同時最適化について提案を行う．従来のHMM音声認識の分野において広く用いられているコンテキストクラスタリングでは決定木構造と呼ばれるモデル構造を構築し，各モデルに割り当てられる学習データ量を増やすことでより信頼性の高いモデルパラメータの推定を行う．しかし，高精度な決定木構造を構築するためには初期のモデルパラメータとして用いられるモデルパラメータの信頼性が高い必要がある．逆に，信頼性の高いモデルパラメータを推定するためには高精度な決定木構造が必要となる．このようにモデルパラメータの推定と決定木構造の構築には相互に強い依存関係があるため，同時に最適化されることが望ましい．しかし，決定木構造の構築とモデルパラメータの推定の間に強い依存関係を持っているために同時最適化は計算量的に困難である．そこで，提案法では，複数のモデル構造を用いてモデルパラメータの推定を行うことでモデル構造とモデルパラメータ推定の同時最適化の近似を表現する枠組みを提案し，連続音声認識実験においても複数のモデル構造を考慮することの有効性を示した．

次に，音響モデルの学習基準について考察する．HMMに基づく音響モデリングでは，尤度最大化(Maximum likelihood; ML)基準が広く用いられている．しかし，ML基準は学習データが十分に得られない場合，モデルの推定精度が低下するという問題がある．これに対し，ベイズ基準では学習データが少ない場合においても高い汎化性能が得られることが知られている．さらに，近年変分ベイズ法が提案され音声認識においてもその有効性が確認されている．しかし，ベイズ基準を用いた学習では，隠れ変数が増加することからML基準よりも初期値に依存する局所最適性問題の影響を受けると考えられる．そのため，学習アルゴリズムを改善するために確定的アニーリングEM (Deterministic Annealing EM; DAEM)アルゴリズムを学習アルゴリズムとして導入することで重要な課題である局所最適性問題に対処することが出きることを示した．

さらに，ベイズ基準による音響モデリングにおいてモデル構造に関する提案を行う．従来の生成モデルによる音声認識システムでは，適切なモデル構造を観測系列から推定するために様々な提案が行われてきた．これらの手法により，モデル構造をより複雑に表現することができるが，音声信号の真の分布を表現するための表現としては不十分であった．そこでベイズ基準において複数のモデル構造を用いることを提案する．ベイズ基準において複数のモデル構造を扱うと言うことはつまり，モデル構造に関しても周辺化を行うということを意味している．ベイズ基準の基本概念は全てのパラメータを周辺化することであるため，モデルパラメータだけでなくモデル構造についても周辺化することは順当な考え方である．連続音声認識実験において，ベイズ基準という統一的な枠組みにおいてモデルパラメータの推定を行い，複数のモデル構造を用いることの有効性を示した．

以上のように，本論文では音声認識システムの性能向上を目的とした統計モデルの高性能化が提案されており，その有効性を示した．また，本論文の内容は国内外の論文誌・国際学会にて公表されている．よって，本研究は情報工学の分野において寄与するところが多大であり，博士論文として十分価値あるものと認める．