様式：課程博士用

2015 年　1 月　16 日

電子・情報工学専攻	学籍番号	033711
申請者 氏名	木村優志	

指導 教員	桂田　浩一

論　文　要　旨　(博士)

論文題目	マルチモーダルタスク推定のための調音運動 HMM 音声認識合成方式

(要旨　1,200 字程度)

音声認識合成技術は近年の技術発展に伴い様々なシーンで利用されるようになってきた．しかし，知的エージェントなどの実現には，周囲環境から対話相手の状態や取り組んでいるタスク，更には行動を推定する技術が必要になると考えられる．本研究ではこうした応用が現実となる状況に備えるべく，特に音声合成，およびタスク推定技術に焦点を当て，それらの性能向上に取り組む．

人間の音声生成と音声知覚が 1-system か 2-systems かは，長年論争され未だ決着がついていないが，近年の脳研究は 1-system 説を支持する結果を示しつつある．著者はこの見解を基に，音声認識と音声合成の双方に共通の音響モデルを使用する「ワンモデル音声認識合成方式」を開発した．本論文では，音声認識のための調音運動モデルを HMM (Hidden Markov Model) で実現し，同じモデルを使用して音声を合成する方式を提案する．これまでに提案された標準的な HMM 音声合成はスペクトル由来の特徴を使用するため特定話者の多量の音声を必要としていた．提案方式は，話者共通の調音運動を HMM で表現すると同時に，HMM から得られる調音特徴系列を，多層ニューラルネットを用いて声道音響パラメータである線スペクトル対 (LSP: Line Spectral Pairs) に変換した後，LSP 合成フィルタにより音声を合成する．したがって少量の音声のみで音声を合成できる．提案方式は主観・客観両評価で従来手法を上回る結果を示した．

また，本論文では知的エージェントが人間と共生するための様々な課題の中で人間との共同作業課題（タスク）を実行するために，耳と眼でこれを判別する機能に焦点を絞って検討した．家庭内でロボットを利用するシーンでは，従事すべき多種多様なタスクについて，その都度指令を与えるアプローチは現実的でない．人間が従事するタスクを音声対話や映像を通してロボットに判断させる研究は，これまであまり実施されてこなかった．そこで本論文では，ロボットやエージェントに人間が従事するタスクを認識させることのできる，より一般性のある方法を提案する．提案方法は最初に，タスク従事中に現れる画像オブジェクトと発話単語という二つのメディアの出現頻度をベクトル空間上に表現し，潜在意味解析 (LSA: Latent Semantic Analysis) を適用してベクトルを圧縮した特徴空間上に表現する．未知タスクの映像と発話からなる入力ベクトルは特徴空間に写像され，予め学習した参照ベクトルとの類似度を計算することでタスクが推定される．本手法は，人間の動作や意図に関する推定を行う必要がないため，高速にタスクを推定することが可能である．本手法では，タスク開始から 40 秒で 90%のタスク正解率を示した．

year month day
2015 1 16

Department	Electronic and Information Engineering	ID	033711
Name	Masashi Kimura		

Supervisor	Kouichi Katsurada

Abstract

Title	Articlatory Movement HMM Speech Recognition and Synthesis System for Multi-Modal Task Estimation

(800 words)

In recent years, automatic speech recognition and synthesis systems are used in various situations. However, it is essential to develop task estimation facilities that enable intelligent agents to interact with human-beings in the real world. This thesis focuses on two important functions of speech synthesis and task estimation to extend the application range of intelligent agents.

Whether human speech production and perception run on a one-system or on two-systems is an unsolved problem yet, however, recent brain studies stand for one-system. Based on the idea of one-model system, the author has developed a "one-model speech recognition and synthesis system" that uses a common acoustic model of articulatory movement HMM (hidden Markov model) in both speech recognition and synthesis. Conventional standard HMM-based speech synthesis systems require a large amount of speech data collected from each speaker. On the other hand, the proposed speech synthesizer uses speaker-unspecific, or independent, articulatory movement HMMs that generate an articulatory-feature (AF) sequence. The AF sequence is converted into a line spectrum pair (LSP) sequence that represents vocal tract acoustic parameters by using a multilayer neural network (MLN), then, the LSP sequence is fed into a digital filter, or LSP synthesizer. The proposed method can reduce the amount of speech data when designing the HMM synthesizer because speaker-independent HMM phone models and a speaker- dependent vocal tract model are separated. The proposed method outperforms a conventional method both in subjective and objective evaluations.

The author also focused on the design of "task estimation" facilities using the information obtained from image data and spoken words. It is an essential function for an intelligent agent to work in various situations through the collaboration with humans. In the real world applications including home robot, we couldn't always instruct, or program, robots, however, a few studies have conducted previously. In this thesis, the author proposes general approach that can estimate human's tasks. The proposed method represents a task vector using the frequencies of image objects and spoken words in utterances appeared in the task. Then, the task vector is converted into feature space of singular vectors by using LSA (Latent Semantic Analysis). When recognizing an unknown task that contains image objects and spoken words, an input task vector is mapped onto the same feature space and then, the similarity between the input singular vector and the reference singular vectors of tasks. The proposed method has the merit that can detect a task in a short time without estimating human behavior and intentions. The experimental results showed the accuracy rate of 90 % within 40-seconds after starting the task.