知能メディア講座　音声メディア分野

教授　奥乃博　助教　糸山克寿　特定助教　西出俊

志望区分:知－7

概要

本研究室では，音声だけでなく，楽音や環境音などのさまざま音，さらにそれらの混合音に含まれる情報の知覚・理解を通じた音環境理解の研究を，知能情報学の立場から行っている．研究のキーワードは,**「音を聞き分ける」，「N次創作を可能にする音楽情報処理」**である．とくに，実環境でのロボット聴覚や音楽情報処理を実現するためには，事前知識の制約を減らすノンパラメトリックベイズ推定によるマルチチャネル音響信号処理，楽器演奏音分析合成，実時間組込みシステム，同時発話認識，擬音語認識の課題に取り組み，音楽共演ロボットや，人とロボットとのインタラクション手法などに展開をしている．

研究テーマ

1. **音環境理解とロボット聴覚**:
   ロボットを実環境に配備し，人との共生を進めるには，必要な事前知識が極力少ないロボット聴覚機能が不可欠である．これまでに同時発話が『聞き分けられる』ロボット聴覚ソフトウエアHARKの高機能化のために，ノンパラメトリックベイズ手法による音源定位・音源分離・音源追跡，擬音語認識，さらには，屋内外でのSLAM (Simultaneous Localization and Mapping) などに取り組んでいる．また, HARKを活用したヒューマンロボットインタラクションの設計も重要な課題である．
2. **CGM (Consumer Generated Media) のための音楽情報処理・音楽共演ロボット**:
   膨大なディジタル音楽を自由に加工し，N次創作を促進し，音楽の新しい楽しみ方を支援するために，多重奏音楽音響信号から最小限の事前知識だけで楽器音を分離し，残響抑制を行う機能をノンパラメトリックベイズ手法により取り組んでいる．また，ハードウエアに依存しない電子楽器テルミン演奏ロボットの開発を行い，人のリードによる音楽共演ロボットの開発にも取り組んでいる．
3. **ノンパラメトリックベイズ法による統計的音響信号処理・認識**:
   これまでに，音源定位，動的変化する音源の追跡，マイク数よりも音源数の方が多い劣決定条件音源分離，音源数推定と音源分離の同時処理，複数の方言発話を許容する方言音声認識，楽器音モデルの揺らぎを許容する楽器音分離，人の演奏揺らぎを許容する楽譜追跡などに取り組んでいる．
4. **ロボットの感情認識・生成による人とのマルチモーダルインタラクション**:
   ロボットの感情表現では，感情認識と感情生成に共通するマルチモーダルなモデルが不可欠である．認識と生成が同じモデルで可能となることによって，ロボットの感情表現が無矛盾になると期待される．これまでに，DESIRE (Description of Emotion by Speed, Intensity, Regularity, and Extent) モデルを提案し，音声，手振り，身振りといったマルチモーダルな感情の統一的な認識生成に取り組んでいる．

分野基礎問題出題範囲

上記のような研究を行うに際しては，人工知能，コミュニケーションモデル，パターン認識と機械学習,音響・ディジタル信号処理，統計的信号処理，聴覚心理学等(すべてを要求しない) に関する基礎的な素養とともに，人間が音を知覚し，理解する過程に対する深い洞察と旺盛な好奇心が望まれる．

具体的には以下の書籍から出題する．

1. 日本音響学会編『音のなんでも小事典』(講談社ブルーバックス),
2. 長尾他著『文字と音の情報処理』(岩波講座マルチメディア情報学第4巻) の音に関する章.
3. 【参考】Al Bregman: “*Auditory Scene Analysis*” (1991, MIT Press) の第1章.

問合せ先
京都大学総合研究7号館408号室　奥乃　博(tel: 075-753-5376)
電子メイル：okuno@i.kyoto-u.ac.jp
研究室ホームページ：http://winnie.kuis.kyoto-u.ac.jp/

Sound Information Processing Group, Intelligence Media Division

Professor: Hiroshi G. OKUNO; Assistant Professors: Katsutoshi ITOYAMA and Shun NISHIDE

Application Code: IST-7

Description

The laboratory, aka Okuno Laboratory, takes a broader view of sound information processing including speech, music, environmental sounds, and their mixtures, investigating from the perspective of intelligence informatics. The key words in our research are *sound source separation* and *music information processing for N-order value-added creation or consumer generated media (CGM)*. Particular emphasis is placed on applications for system built into robots. The laboratory develops technology that enables real-time processing by hierarchically integrating multiple sensor inputs real-world environments.

Research Topics

1. **Computational auditory scene analysis and robot audition**:
   Since we hear mixed sounds in a daily life, sound source localization, i.e., where sounds come from, sound source separation, i.e., what sounds are included, and recognition of each separated sound are mandatory for sound information processing in a real-world environment. This research area is called *computational auditory scene analysis, CASA*. We are developing a robot audition open-sourced software called HARK. Some key capabilities include CASA functions based on a non-parametric Bayesian approach, self-generated sound cancellation, and audio-visual integration for auditory scene analysis. For human-robot interactions, we research simultaneous speech recognition for Shotoku-Taishi robots and automatic onomatopoeia recognition for environmental sound manipulation.
2. **Music information processing for CGM and music co-player robots**:
   A new style of music appreciation is that people enjoy music by separating and remixing some parts of existing music performance. For this *active* music appreciation and CGM, we adopt *analysis-and-synthesis* of musical instrument sounds. Musical instrument sounds separation with echo cancellation is being developed by non-parametric Bayesian methods. We also develop co-player musical robots that sing and play the Theremin, an electronic musical instrument. For recognizing partner's musical behaviors in an ensemble performance, we study real-time beat-tracking, score following, and human gesture recognition based on non-parametric Bayesian approach.
3. **Non-Parametric Bayesian Methods for Statistical Singal Processing and Recognition**:
   Since real-world application of CASA requires concurrent solution of several processing, non-parametric Bayesian methods are exploited to solve sound source localization and separation, speech recognition for multiple dialects, estimation of the number of sound sources and separation of them, and score following.
4. **Multi-modal interaction with robots based on emotion recognition and synthesis model** :
   Emotion model should be bidirectional and multi-modal, because the bidirectional emotion model will enable the system/robot to recognize human's emotional expression and synthesize robot's emotion between different modalities. We have developed DESIRE (Description of Emotion by Speed, Intensity, Regularity and Extent) and comfirmed transmodal emotion between voiced utterance, gesture and behavior.

Scope of Area-specific Basic Questions (Master's Program)

This research requires a basic understanding of at least some of the following: artificial intelligence, communcation modeling, pattern recognition and machine learning, acoustic digital siganal processing, statistical signal processing, psychoacoustics, psychophysics, etc. Applicants are also reqeusted to have keen observation skills and avid curiosity regarding human sensation and understanding of sound. More specifically, questions will come from the following two books.

1. The Acoustic Society of Japan ed., *A Small Dictionary of Sounds* (Kodansha Blue Backs Series).
2. NAGAO, M. et al., *Information Processing of Texts and Sounds* (Iwanami Lecture Series on Multimedia Informatics, Vol.4) (chapters regarding sound).
3. [Reference] Al Bregman: *Auditory Scene Analysis* (1991, MIT Press) Chapter 1.

Contact

OKUNO, Hiroshi G., Room No.408, Research Building No.7, Kyoto University (Tel: 075-753-5376)
E-mail: okuno@i.kyoto-u.ac.jp
Laboratory website: http://winnie.kuis.kyoto-u.ac.jp/