上智大学情報理工学科・日本音声学会 共催講演会

Title: What does a machine learn that learns to speak?

Speaker: Dr. Reiner Wilhelms-Tricarico (Haskins Laboratories)

Date: July 21, 2014 (Mon.)

Time: 15 : 30 - 16 : 45

Place: Sophia University, Yotsuya Campus, Central Library, Room L-821

Language: English

Abstract: Machine learning and pattern discovery methods, often subsumed as "deep learning methods", has resulted in many quite remarkable improvements in speech recognition and recently also in speech synthesis. This happened in part to the detriment of classical approaches to modeling, both in linguistics and articulatory phonology. Building articulatory speech synthesizers was once the objective for many research projects. It gave many insights into human speech production and motor control, but had very little influence on the development of commercially viable speech synthesizers. For speech recognition, the expected gains of using information from articulatory phonology were mostly superceeded by machine learning methods, whose main strength is not a better insight in speech production but simply an efficient process that makes use of massive data.

The question is if it is still possible to build better models that take actual knowledge about speech production into account while fully exploiting the advantages of deep learning.

I am going to describe several components from machine learning that can be used and combined with the classic way of modeling: A voice source that makes use of a generalized recurrent dynamic model of the voice source, a model that attempts to represent articulatory dynamics, by estimating parameters of recurrent systems which describe the short term dynamics of spectral parameters (or articulation itself if data are available), a model for intonation contour and phrasing that is based on several time scales, namely phonological, syllable/word and phrase based. Some of the components for the analysis and for obtaining model parameters are implemented by means of a generalized Kalman filter, the cubature Kalman filter. Finally I propose a system that tries to extract prosodic parameters and markers by analyzing text that is aligned to the acoustic signal, and makes use of deep learning methods.

※本講演会は日本学術振興会の科学研究費補助金 (#263004) の支援を受けています.

問い合わせ先：上智大学 理工学部 情報理工学科　荒井隆行 (arai@sophia.ac.jp)