Real-time Audio Event Detection: An Integrated Swarmlet Approach

http://terraswarm.org/

Ilge Akkaya[1], Jeffrey A. Bilmes[2], Richard Rogers[2], Edward A. Lee[1]
[1]University of California at Berkeley, [2]University of Washington

September 9th-11th, 2015

Motivation

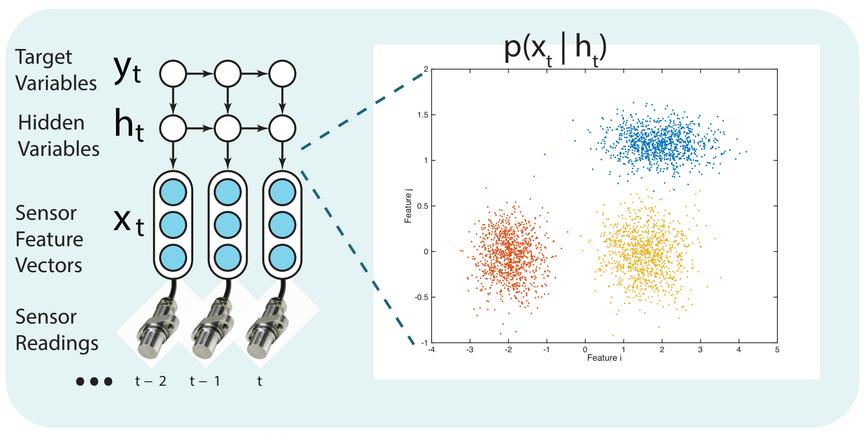
Emerging internet-of-things (IoT) frameworks integrate a wide range of heterogeneous aspects such as real-time sensing, physical dynamics, control and inference components.

In many cooperative mobile swarm applications, it is of interest to capture real-time sensor data, perform inference on streaming feature vectors and to make actuation decisions accordingly, potentially in a distributed setting. Integrating such heterogeneous components on a compositional and scalable platform to design real-time learning and inference tasks remains a challenge for the IoT.

We present an integrated Accessor framework to perform real-time audio event detection. The end goal of our demonstration is to actuate a Scarab robot in real-time, which in turn reacts to detected 'applause' events.

Demo Workflow

- We build a **Swarmlet** (i.e., a service that integrates networked sensors and devices with cloud services) using Ptolemy II, which is an actor-oriented modeling and simulation tool that enable interfacing with web services, sensors, and actuators via specialized actors called **Accessors** [1].
- Captured audio data is streamed through a set of signal processing actors to extract audio features, which are then streamed through the WebSocket protocol via the WebSocketClient.js Accessor.
- An instance of gmtkOnline receives the JSON formatted feature streams and performs online decoding



GMTK Training Model

Audio Capture

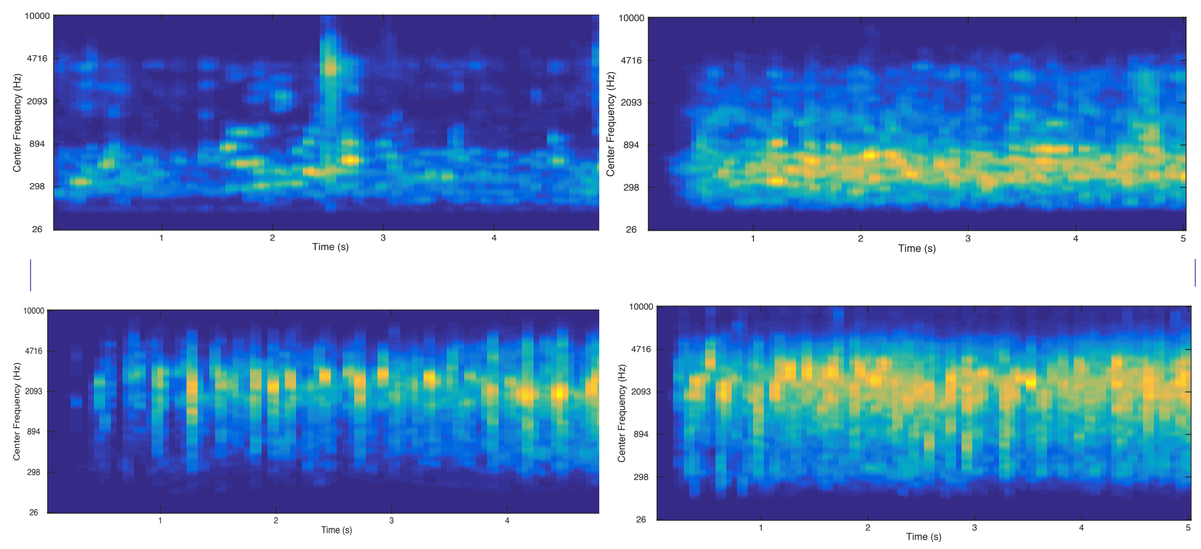
Real-time Feature Extraction and Training

Bayesian Network Models

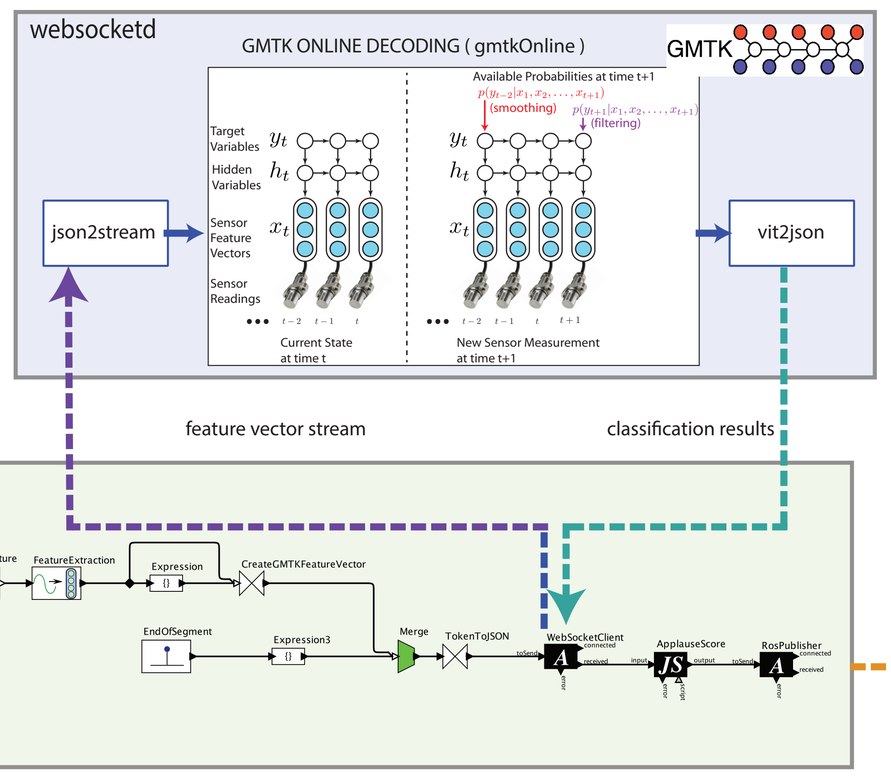
- Graphical Models Toolkit (GMTK) is an open-source framework that aims at prototyping statistical models using dynamic Bayesian Networks (DBNs).
- Audio feature training, as well as online decoding is performed using GMTK [2].
- We model the audio events using a two-state Hidden-Markov Model (HMM) with Gaussian-Mixture Model emissions. The hidden states represent applause and no applause (speech/ambient noise) events, respectively.

Audio Feature Selection

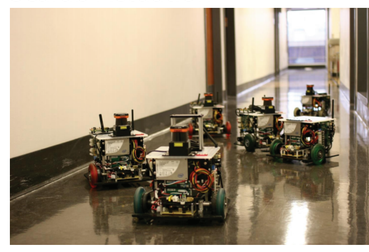
We consider several audio features for real-time applause detection via a class of discriminative audio features that are based on *auditory filterbank temporal envelopes and linear prediction filter coefficients*.



Cochleagrams depicting the log-energy of of an auditory filterbank output of two [Top] speech event samples and [Bottom] applause event samples



Scarab Actuation



Action

References:

1. Latronico, E., Lee, E., Lohstroh, M., Shaver, C., Wasicek, A., & Weber, M. (2015). A Vision of Swarmlets. Internet Computing, IEEE, 19(2), 20-28.
2. Bilmes, J., & Zweig, G. (2002, May). The graphical models toolkit: An open source software system for speech and time-series processing. In Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on (Vol. 4, pp. IV-3916). IEEE.

Sponsored by the TerraSwarm Research Center, one of six centers administered by the STARnet phase of the Focus Center Research Program (FCRP) a Semiconductor Research Corporation program sponsored by MARCO and DARPA.