Audio recognition, context-awareness, and its applications

Yoonchang Han
Co-founder & CEO, Cochlear.ai
26 March, 2018
Rule-based methods ➔ Deep learning
See

Computer vision

Understand language

Natural language processing

Listen

Speech recognition

(Source: Softbank Pepper)
Taking an umbrella

Closing the window
Foot step sound

High heels
Easy for Humans

Hard for Machines
Evolution of data processing technique

Early days

- Feature engineering

Traditional ML

- Feature engineering
- ML Classifier

Deep learning

- Deep learning

More human effort

More automatic

Better performance

Data

Prediction
To tackle each topic
(make some “rules”)

To simulate how
human understand the sound
(and prepare data)
Required domain knowledge

- Signal Processing
- Cognitive Sciences
- Music
- Psychoacoustics
- Acoustics
- Machine Learning
“Modern” audio identification pipeline

Audio → Time-frequency representation → Neural Network → Output

objects in an image ≈ instruments in a spectrogram

flower → butterfly

voice → violin → piano
“Machine listening” is the use of signal processing and machine learning for making sense of natural / everyday sounds, and recorded music.

- Machine listening lab, Queen Mary, Univ. of London
### Acoustic scenes
- bus
- library
- train
- home
- cafe
- park
- city centre
- driving
- market

### Acoustic events
- glass break
- car horn
- footstep
- gun shot
- bird chirping
- sneeze
- knock
- dog bark
- water boil
- snoring
- crying

“Any” sound we hear everyday
Computer vision

- Optical Character Recognition (OCR)
- Facial recognition
- Object detection

Machine listening

- Voice recognition
- Music search
- Speaker identification
- Acoustic scene/event detection

(Sources: Tensorflow, Facebook, Microsoft, Apple, Shazam)
Scene classification accuracy (IEEE DCASE)

2013: 76%
2017: 92%

Perceive
Think
Act
Five, Zero

Cat
Know what it is (with input restriction)

Know what it is

Know what/where it is

Know what/where it is + why
**cochlear.ai Sense** (closed alpha release in April)

- **Activity detection**
- **Music, Speech, Others**
- **Music analysis**
  - Genre / Mood / Key / Tempo
- **Speech analysis**
  - Age / Gender / Emotion
- **Scene classification**
  - Indoor / Outdoor / Vehicles
- **Acoustic event**
  - Dog bark / Baby cry
  - Car horn / Snoring ...
Why do we need...

Activity detection

Unified model
It is really challenging because...

Recording environment

Recording device

Noises

Local characteristics

Overlapped / Polyphonic
Probability or Saliency?
Example: AI speakers

Simple voice control

“Alexa, turn on the light”
“Alexa, play dance music”
“Alexa, turn on TV”

IoT control-tower with context-awareness

(footstep sound, door slam, cough, Someone got back home, got a bad cold)

turn on light / TV
play suitable music
adjust room temperature warmer
(not just a pattern, there is a “reason”)
ask to take cold medicine before sleep
Example: Humanoid robots

See things
Understand speech
+
Listen things other than voice
Know who they talk to

(Source: Atlas, Boston Dynamics)
Example: Autonomous car

Outside - Car horn (normal, air horn), Siren (fire truck, police, ambulance)

Inside - Music mood, snoring, baby, anomaly detection (malfunction warning)
ATMO: Generative music for spatial atmo-sphere
Generative Music with contextual information
Generative Music with contextual information
Analysis Result: Typing in a rainy day...

Contextual Information:
Typing...
Reading a book...
Raining outside...
contact@cochlear.ai