Human perception and listening by machines

Cleo Pike (Chair)
School of Psychology and Neuroscience, University of St Andrews, UK

Amy Beeston (Panelist)
Department of Computer Science, University of Sheffield, UK
Research background

Cleo Pike

Academy of Contemporary Music Production
University of Surrey MSc Psychology
University of Surrey PhD Psychoacoustics
University of St Andrews Research: Multi-Sensory Perception
Research background

Amy Beeston

University of Edinburgh
BMus Music Technology

Koncon, The Hague
MMus Sonology

University of Sheffield
PhD Computer Science
Section 1: What is machine listening?

Section 2: What are the processes involved in machine listening and the problems encountered?

Section 3: How do humans do it better?
Introduction

What is a machine listener?

What is a machine?

A machine receives input commands and follows rules in software to perform an action.

What is listening?

Registering audio input (Hearing) + an effort to interpret (recognize/attend to) input.

Machine listeners:

<table>
<thead>
<tr>
<th>Hear</th>
<th>Listen</th>
<th>Act</th>
<th>Learn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanical transduction</td>
<td>Categorisation</td>
<td>Follow rules/norms for acting</td>
<td></td>
</tr>
<tr>
<td>Division by frequency</td>
<td>Recognition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transduction to neural firing</td>
<td>Streaming</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature selection</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Applications of Machine listening

ASR

Automatic Speech Recognition:

Siri (Apple)  

Cortana (Windows)
Applications of Machine listening

ASR

Siri

Siri

Siri

Cortana

Karen Jacobson (Aus)  Jon Briggs (UK)  Karen Jacobson (USA)  Jen Taylor (USA)

Applications of Machine listening

ASR

Speech recognition:

Dictation systems
Translation systems

English speech → English word? --- French word? → French speech

Speaker recognition:

Verification (check it’s you)
identification (work out who you are compared to N other people)
Applications of machine listening
Beyond ASR

- para-linguistics: emotion recognition, conversation analysis
- event detection: sonic interaction, alarm notification
- engagement: sort and search, information retrieval

www.thatwhitepaperguy.com/images/using-voice-recognition.png
www.maximumpc.com/files/dh6627/shout.jpg

Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016
Applications of machine listening

Example 1. para-linguistics

Example 1. para-linguistics

https://www.newscientist.com/article/mg22229683-800-speech-analyser-monitors-emoHon-for-call-centres/

– helps customer-service reps build better rapport with customers

http://vms.mit.edu/cogito
Applications of machine listening

Example 2. event detection

Applications of machine listening

Example 3. engagement

Shazam – digital fingerprint

Helps people recognize and engage with the world around them
http://www.shazam.com/company

Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016
Conceptual explanation of ML

Processes behind machine listening

For any ML we need:
Hardware and software

**Machines:**
Engineered peripherals (microphones)
Engineered software - algorithms
Engineered peripherals (loudspeakers)

**Humans:**
Evolved peripherals (ears, nerves)
Evolved software - algorithms
Evolved peripherals (nerves, mouth)

Materialist vs Dualist Philosophy
Statistical ASR

Conceptual explanation of ML

Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016

Conceptual explanation of ML

In general: task and context dependencies

1-D time-domain audible signal

n-D feature vectors control parameters

combining features (multimodal)

classification (supervised)

clustering (unsupervised)
Temporal feature (1D)

Intensity tracking (Praat)

1. Track the intensity envelope
   – Open file, show intensity
   – Extract visible intensity Contour

2. Segment signal (voice activity detection)
   – Praat Objects window, Intensity > To TextGrid (silences)

3. Sound and TextGrid > View & Edit

Praat
– doing phonetics by computer
– www.praat.org/

Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016
**Temporal feature (1D)**

### Intensity tracking fail

<table>
<thead>
<tr>
<th>Time (mm:ss)</th>
<th>Amplitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:00</td>
<td></td>
</tr>
<tr>
<td>00:30</td>
<td></td>
</tr>
<tr>
<td>01:00</td>
<td></td>
</tr>
<tr>
<td>01:30</td>
<td></td>
</tr>
<tr>
<td>02:00</td>
<td></td>
</tr>
</tbody>
</table>

### Prime (bottom-up) grouping
- simultaneous (vertical) – common on/offset, harmonicity
- sequential (horizontal) – continuity, proximity

### Schema-driven (top-down)
- prior knowledge, semantics, pragmatics

---


Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016
Spectral features (2D)

Pitch estimation (Sonic visualiser)

1. Reveal the harmonic structure
   - Open file
   - Pane > Add spectrogram
2. Estimate fundamental frequency in harmonic regions
   - Transform > Aubio Pitch Detector

Sonic Visualiser
   - for viewing and analysing the contents of music audio files
   - http://www.sonicvisualiser.org
Aubio Pitch Detector
   - http://www.vamp-plugins.org/

Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016
Spectral features (2D)

Pitch estimation ‘fail’

- Owen Green’s Now for some music (2007)

- Implementation
  - First pitch tracker adaptively divides input sound into two classes (pitch or noise)
  - Amount of disagreement between first and second pitch tracker controls signal processing resulting in more/less perceived roughness/disruption of input

Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016

Multi-dimensional features (nD)

Timbral description (Max)

- ... perceived dissimilarity despite same loudness, pitch and duration
- Brightness => spectral centroid
- Noisiness  => spectral flatness

Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016
Spectro-temporal features (nD)

Timbral description ‘fail’

- humans adapt to the room (and fast!!)
- our machine listeners typically don’t

---


---

Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016
Problems with ML applications

Machine listening is flawed

- ML breaks with common environmental problems (noise, channel coloration, reverb)

- We are only using one source of information to classify sounds. Real systems can use multiple sources

- Most cues have problems with reverberation/background noise and coloration

- Human listeners have a means of overcoming these problems and machine listening can incorporate this

- Humans take the context into account
Overview

• Reverberation degrades speech intelligibility
  – acoustic content differs with distance
  – but phonetic content persists

• We compensate for reverberation
  – monaural/binaural

• Compensation is reliant on contextual sound
  – what factors promote/inhibit compensation?

• Can machine listeners use equivalent cues?
  – same mistakes as humans?

Compensation for reverberation

Late reverberation

- Late reverb => noise-like effects
  - Increases noise floor
  - Reduces dynamic range of temporal envelope
- Stop consonants => very sensitive to reverb
  - Identification depends on rapid amplitude modulation, e.g. [t] dip
  - Peaks prolonged, dips filled

Compensation for reverberation

Watkins — Next you’ll get \{sir, stir\} to click on

Human audition

Peripheral and central processing

- Continual recalibration: feedback (centrally and to the periphery)
  - low-level/stimulus driven \textit{and} high-level/attentional effects

\textbf{Guinan (2011). Auditory and vestibular efferents}

Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016
Efferent-inspired auditory model

- Efferent processing => reduce response to energy in reverberant tail

Compensation for reverberation

Findings

• Monaural replication and extension of Watkins’ work
  – real speech, multiple talkers, incl. s+{t, k, p}+vowel

• Human compensation
  – is apparent for {p, t, k} when high freqs are present
  – is abolished with time-reverse reverberation
  – uses intrinsic info when extrinsic context is ambiguous
  – is rapid (c. 500 ms)

• Compensation model
  – does not require phonetic processing
  – uses efferent processing to help recover [t] dip
  – best version derives info from reverberant tails

Beeston and Brown (2014). 7th Forum Acusticum, Krakow, Poland.
Aim

Do humans compensate for spectral distortion (colouration) caused by environment?

What are the perceptual mechanism involved?

Can we apply any to benefit ML?
Compensation for spectral distortion

Rationale

Spectrum - key to recognition
e.g. /e/ or /a/

Environment - rooms, loudspeakers, microphone

Spectral distortion/colouration - /e/ physically becomes /a/

Compensation - we still hear the intended /e/ vowel

Fig. 5.15. Canadian Vowels [free]
Redrawn by UKT from Russell: http://www.umanitoba.ca/faculties/arts/linguistics/rusell/l38/sec4/formants.htm
Compensation for spectral distortion

Experiment 1

http://www.esm.rochester.edu/concerts/halls/hatch/

Cleo Pike


Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016
Results

Condition 1

| 1 | 2 | 3 | 2 | 3 | 1 |

| 1 | 2 | 3 | 2 | 3 | 1 |

Compensation for spectral distortion

Explanation of results

A memory effect? (Olive et al. 1995)
Compensation for spectral distortion

Experiment 2

Is time between listening a cause of compensation?

Condition 1

\[ 1 \quad 2 \quad 3 \quad 2 \quad 3 \quad 1 \]

→ Time

Condition 3

\[ 1 \quad 2 \quad 3 \quad 1 \]

→ Time
Compensation for spectral distortion

Results

Condition 1

Condition 3

Results

CondiHon*Room F=187.22 p<.001
Effect size = 29 points, a “Large” effect

Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016
A memory effect? (Olive et al 1995)

Memory loss should cause noise in ratings not contraction
Compensation for spectral distortion

Explanation

Aim:
Find mechanisms to explain compensation due to time gap

The ‘auditory enhancement’ effect
Aim:
Find mechanisms to explain compensation due to time gap

The ‘auditory enhancement’ effect

![Diagram showing two spectra over time](image-url)
Compensation for spectral distortion

Explanation

Aim:
Find mechanisms to explain compensation due to time gap

The ‘auditory enhancement’ effect

Cleo Pike and Amy Beeston · AES140 · W17 · Human perception and listening by machines · Paris · 7 June 2016
Overall findings

**Enhancement** enhances spectral change in running speech or music

This raises spectral change in speech/music above a ‘colouration floor.’

Additionally, the colouration floor can be removed with a similar but longer time course process...

**The spectral compensation effect**
Compensation for spectral distortion

Application to ML

Is this process implemented in machines?

Machine listeners do remove colouration:
‘Speaker vocal tract compensation’
Can also be used to remove colouration by any channel

Vocal Tract Length Normalisation

Cepstral mean subtraction

Are these even needed for colouration by reverb?
Dereverberation processes should also remove colouration

Co-articulation effects could also be compensated for with an ‘enhancement’ process
Computational Auditory Scene Analysis

**Cocktail party problem** we receive mix of sound
How do we pick out any one?

**ASA** – principles in human listening to segregate auditory streams

**Knowledge based (Schema) grouping** – Prior experience, top down

**Primitive grouping** - low level, bottom up
Examples:
  - Common onset
  - Common AM, FM
  - Harmonicity
CASA mimics the auditory system from the beginning:

- Just two microphones
- Gammatone filterbank
- Cochleogram
- Segmentation into TF units
- Segregation vertically and horizontally - based on ASA principles
- Binary/soft mask applied to isolate target from noise
The ideal binary mask is a 0 or 1 separation of noise from the background.

Similar to occlusion in vision:

However, attention is not absolute so further psychoacoustic principles could be added.....

“Humans can hold the noise streams in mind but this is not often implemented in Machines.” (Wang 2005)
Other important human perceptual mechanisms that have been implemented in machine listening:

**Categorical perception**

The listener hears distinct categories of musical or speech sounds rather than a continuum

CP can remove variation caused by distortion

Clustering algorithms mimic this

**Neural networks** - inspired by human perception
Thank you

Please ask us questions:

Amy
a.beeston@sheffield.ac.uk

Cleo
cdp5@st-andrews.ac.uk


