Acoustic Scene Classification

Marc-Christoph Gerasch

Outline

- Acoustic Scene Classification - definition
- History and state of the art
- Two approaches
  - Statistic
  - Human
- Conclusion
- Further research
- Questions and Answers

Acoustic Scene Classification

- Computational Auditory Scene Analysis (CASA)
- Classifying the environment of an audio record
- Acoustic event classification

- Cherry (1953): ‚Cocktail party problem.‘
- Human vs. machine
- Application:
  - Hearing aids
  - Speech recognition
  - Context aware computing applications

Shutterstock.com

History and state of the art

- 1932 • Speech recognition at Bell labs
- 1953 • Cherry: ‚Cocktail party problem.‘
- 1982 • David Marr - information processing of the brain from a computational view
- 1990 • Bregman – ‚Auditory Scene Analysis‘
- • Development of digital hearing aids pushed CASA
- 1997 • Sawhney and Maes – first exclusive CASA method
- 1998 • Hidden Markov Models
- 2003 • TrecVid started
- • Mel Frequency Cepstral Coefficients
- 2013 • IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)
- 2015 • IEEE WASPAA (forthcoming)

Two approaches

Statistic	Human
• pure physical information	• Brainwork
• Low-level grouping	• Low-level grouping
• Monaural	• High-level grouping
• Brute force all data analysed	• Binaural
	• Attention
	• Filters (Band-pass,…)

Two approaches
-similarities-

- Preparation of the audio stream
  - E.g. windowing,…
- Physical features of the audio stream are extracted
  - E.g. MFCC, F_0 ,..
- Events are hints to the scene
- Training and classification phase

Technical methods

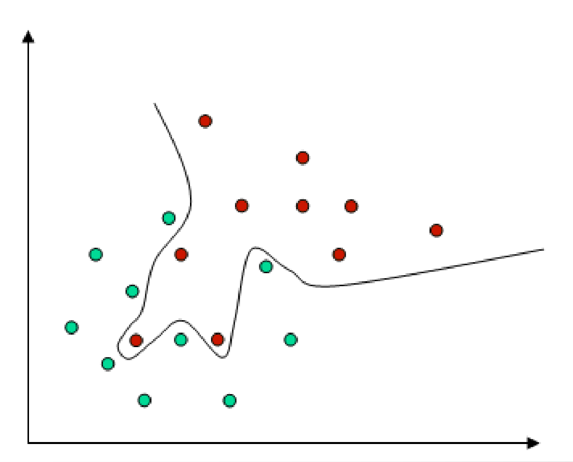
- F_0 (fundamental frequency)
  - Detection and summation of harmonics for finding f_0
  - Speech recognition
  - Multi speaker problem
- MFCC (Mel Feature Cepstral Coefficients)
  - Transformation of audio invented for speech recognition
  - Mel: perceptual scale of pitches
  - Cepstrum: Inverse Fourier transform (IFT) of the logarithm of the estimated spectrum of a signal

  → possibility to divide vocal excitation (pitch) and vocal tract (formants)

Technical methods

- LPI (Latent Perceptual Indexing)
  - Similar to latent semantic indexing for text analysis
  - Points out the super ordinated attributes/key attributes
  - For huge amounts of data
  - Needs lot of training

- SVM (Support Vector Machine)
  - Representation of acoustic events as vectors
  - Certain vectors (support vectors) construct a hyper plane dividing scene classes



Ennepetaler86 from www.wikipedia.org

Statistic approach
Geiger et. al.

- Audio preparation
  - Monaural
  - Windows (overlapping)
- openSMILE:) feature extractor
  - MFCC (Mel Feature Cepstral Coefficients)
  - F_0 (sub harmonic summation and probability of voicing)
  - …
- Classification
  - SVM (Support Vector Machines)
  - LPI (Latent Perceptual Indexing)

Statistic approach
-results-

- More training data needed for LPI
- SVM obtained best results
- Window size matters
- MFCC does the main part (68% combined with SVM)

- 71% on training data
- 69% on evaluation data

Human based approach
Kalinli et. al.

- How does the ear perceive sounds?
- What is happening in the brain while listening?
- What influence has experience?
- How does attention work?

→ LISA (Latent Indexing using SAliency)

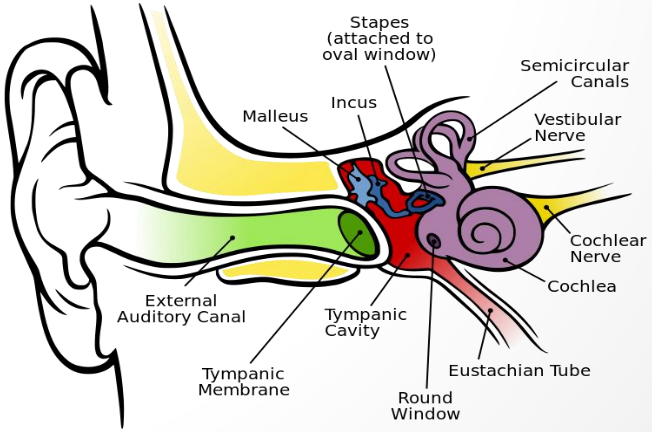
Human based approach
-sound perception-

Human

- Usually two ears (binaural hearing)
- Sounds have spectral harmonics
- Frequency dependent perception of the cochlea
- Constant noises are partially supressed

Implementation

- Two microphones
- F_0
- Band-Pass filter
- Noise reduction



"Anatomy of the Human Ear", A. Brockmann

Human based approach
-brainwork-

Human

- Auditory cortex (feature extraction)
- Comparison and grouping of cues
- Experience
- Information storage

Implementation

- MFCC, F_0
- High-level cue grouping
- Context awareness
- Neural network

Human based approach
-attention-

Human

- Like a spotlight
- Suppression of noise without attention (binaural)
- Microphone → just cacophony
- Direction and movement detection (binaural)

Implementation

- Salient event detector
- Saliency feature filter
    - Intensity
    - Frequency contrast
    - Temporal contrast
    - Orientations/latency

Human based approach
-results-

- Goal was not to reach best results
- Comparison LISA vs. Baseline (40%)
- 74% reduced data for better results (50% using top 35 salient events)
- Up to 98% reduced data for baseline results (40% using top 10 salient events)

Conclusion

- Basic methods are similar (MFCC, LPI,…)
- Different audio databases (no direct comparison)
- Statistical methods seem to be more accurate
- Human mimicking methods vastly reduce data and computing effort
- Both approaches do not hit the mean human accuracy (71%)

Further research

- Algorithms for devices with limited computational power
- Independent systems for unlabelled scenes
- Including external information e.g. Geo location

References

- Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell and Mark D. Plumbley, Senior Member, IEEE. School of Electronic Engineering and Computer Science, Acoustic Scene Classification, November 17, 2014.
- Ozlem Kalinli, Shiva Sundaram, Shrikanth Narayanan. Saliency-Driven Unstructured Acoustic Scene Classification Using Latent Perceptual Indexing. MMSP'09, October 5-7, 2009.
- Jürgen T. Geiger, Björn Schuller, Gerhard Rigoll. Large-Scale Audio Feature Extraction And Svm For Acoustic Scene Classification. 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 20-23, 2013, New Paltz, NY.
- Malcolm Slaney. The History and Future of CASA. In Perspectives on Speech Separation, Editor: P. Divenyi, Kluwer, 2006.
- Deliang Wang and Guy J. Brown. Fundamentals of Computational Auditory Scene Analysis. 2006.
- Ben Milner and Dan Smith. Acoustic Environment Classification. ACM Transactions on Speech and Language Processing, Vol. 3, No. 2, July 2006, Pages 1–22.