# DCASE 2016: Detection & Classification of Audio Scenes and Events

Introduction and Philosophy

Mark Plumbley
Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK

# DCASE 2016: Why?

- Huge potential for automatic recognition of real-world sounds
- Up to now: relatively little research activity, compared to e.g. image, speech, or even music
- Barrier? -> Shortage of good open datasets for research
  - Data is expensive/time-consuming to collect and label
  - Commercial data may be restricted, hard to compare
- Public evaluation data challenges:
  (1) Provide open data that researchers can use
  (2) Encourage reproducible research
  (3) Attract new researchers into the field
  (4) Create reference points for performance comparisons

# Previous data challenges

- Some earlier evaluation challenges, e.g.:
  - MIREX: Music Information Retrieval (since 2005/6)
  - PASCAL CHiME: Speech Separation (since 2006)
  - CHIL CLEAR: AV from meetings (2007-8)
  - SiSEC: Source Separation (since 2008)
  - TRECVID Multimodal Event Detection (since 2010/11)
- IEEE Audio & Acoustics Sig Proc (AASP) TC support, e.g.:
  - CHiME 2, REVERB, ACE, … and DCASE 2013
- DCASE 2013: Audio Scenes and Events
  - 3 Tasks: Acoustic Scenes; Office Live; Office Synthetic
  - 18 participating teams, presented at WASPAA 2013

# DCASE 2016: Overview

- Build on and extend success of DCASE 2013
- More data, more complex, closer to real applications

Four Tasks:

- Task 1: Acoustic scene classification
  - Audio environment, e.g. "park", "street", "office"
- Task 2: Sound event detection in synthetic audio
  - Office sound events, e.g. “coughing”, “door slam”
- Task 3: Sound event detection in real life audio
  - Events in Home (indoor) and Residential area (outdoor)
- Task 4: Domestic audio tagging
  - Activity in the home, e.g. “child speech”, “TV/Video”

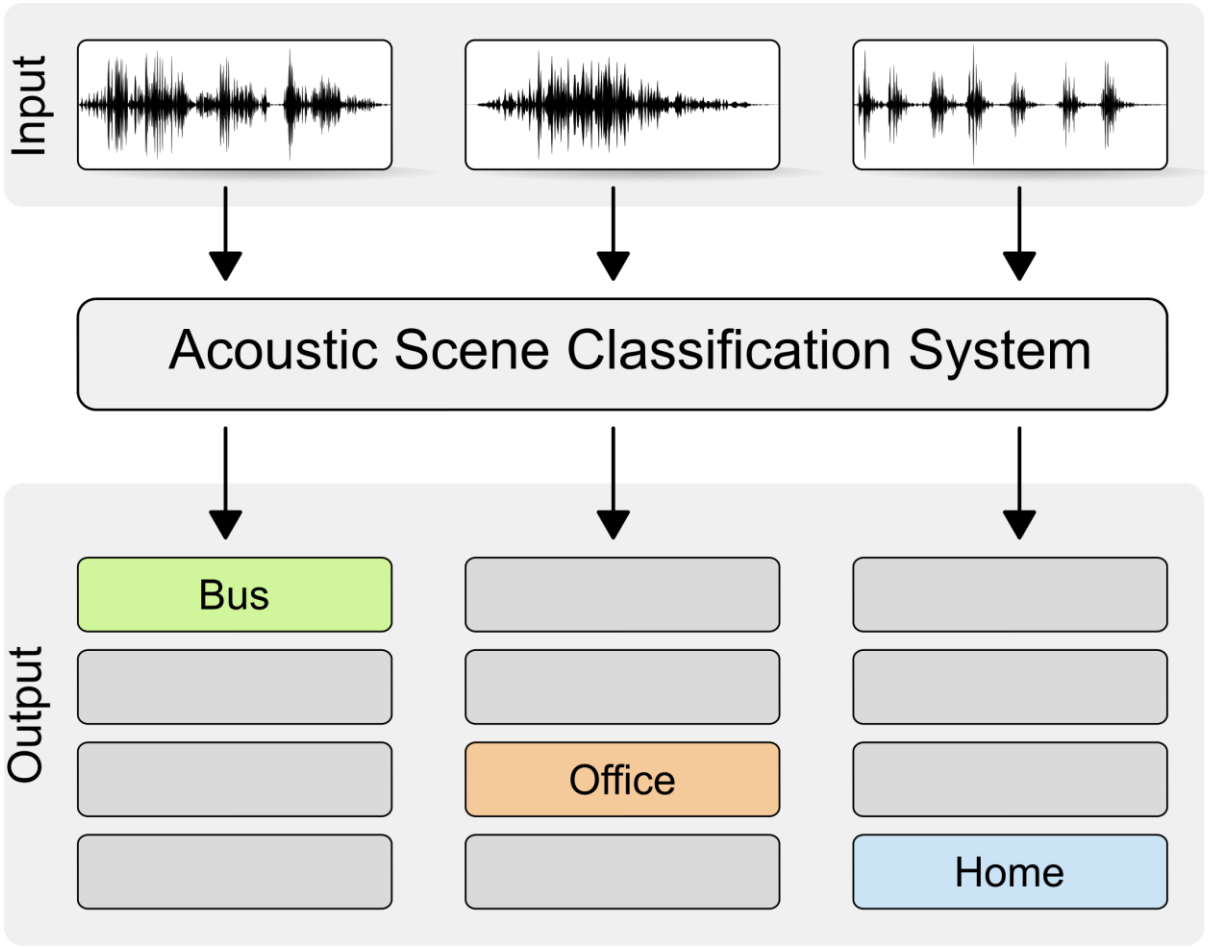
# DCASE 2016: How?

- International organizing team:
  - Tampere University of Technology (FI)
  - Queen Mary University of London (UK)
  - IRCCYN (FR)
  - University of Surrey (UK)
- Submissions
  - 82 submissions to the challenges
  - 23 Papers submitted to the workshop
- DCASE 2016 Workshop (Today)

# DCASE 2016
# Tasks and Results

Tuomas Virtanen
Tampere University of Technology
Finland

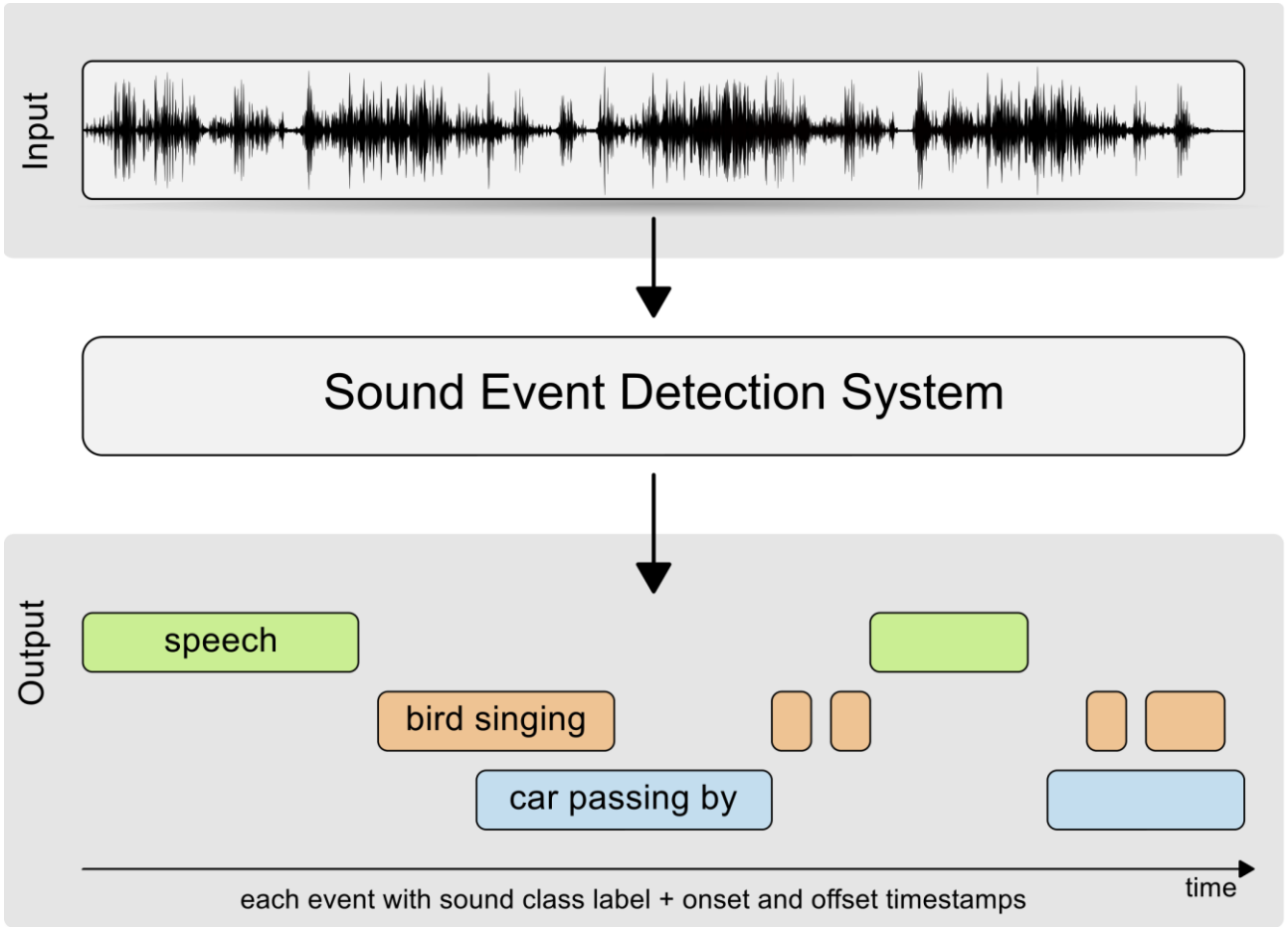
# Task 1: Scene Classification



# Task 1: Scene Classification

- 15 classes (bus / café / car / city center / forest path…)
- Binaural audio, 44.1 kHz, 24 bits
- Recorded in different locations in Finland
- Development set (9 h 45 min)
  - From each scene class: 78 segments, 30 seconds each
  - 4-fold cross-validation setup
- Evaluation set (3 h 15 min)
  - 26 segments per scene class
  - Evaluated using classification accuracy

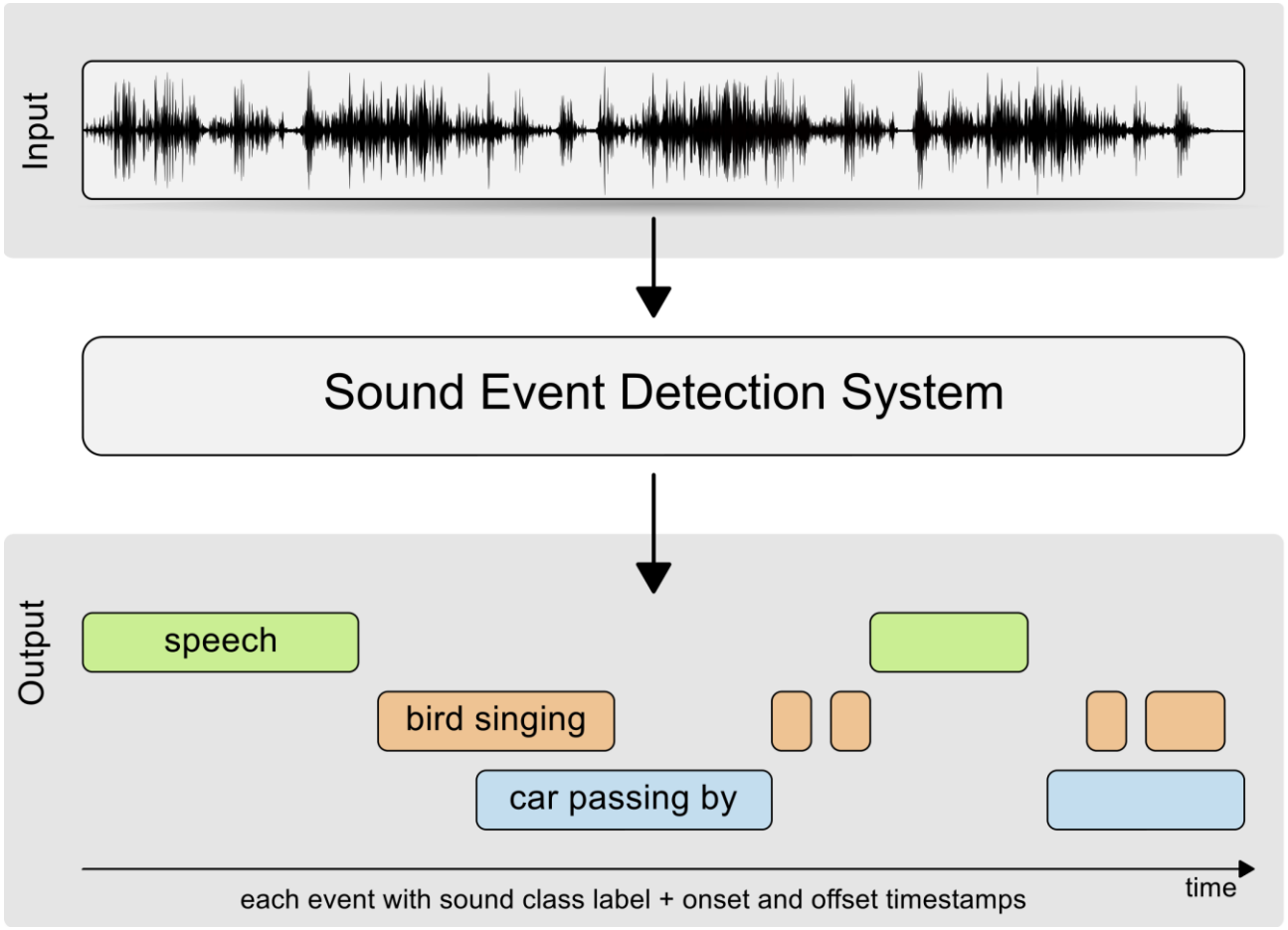
# Task 2: Event Detection, Synthetic Audio



# Task 2: Event Detection, Synthetic Audio

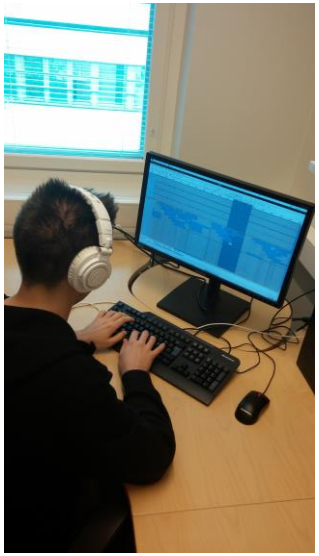
- 11 sound event classes (clearing throat, coughing, door knock, door slam, drawer, human laughter, keyboard, keys, page turning, phone ringing, speech)
- Development set:
  - 20 isolated samples per class
  - 18 minutes of generated mixtures
- Evaluation set:
  - 54 audio files of 2 min duration each
  - Multiple SNR and event density conditions

# Task 3: Event Detection, Real Life Audio

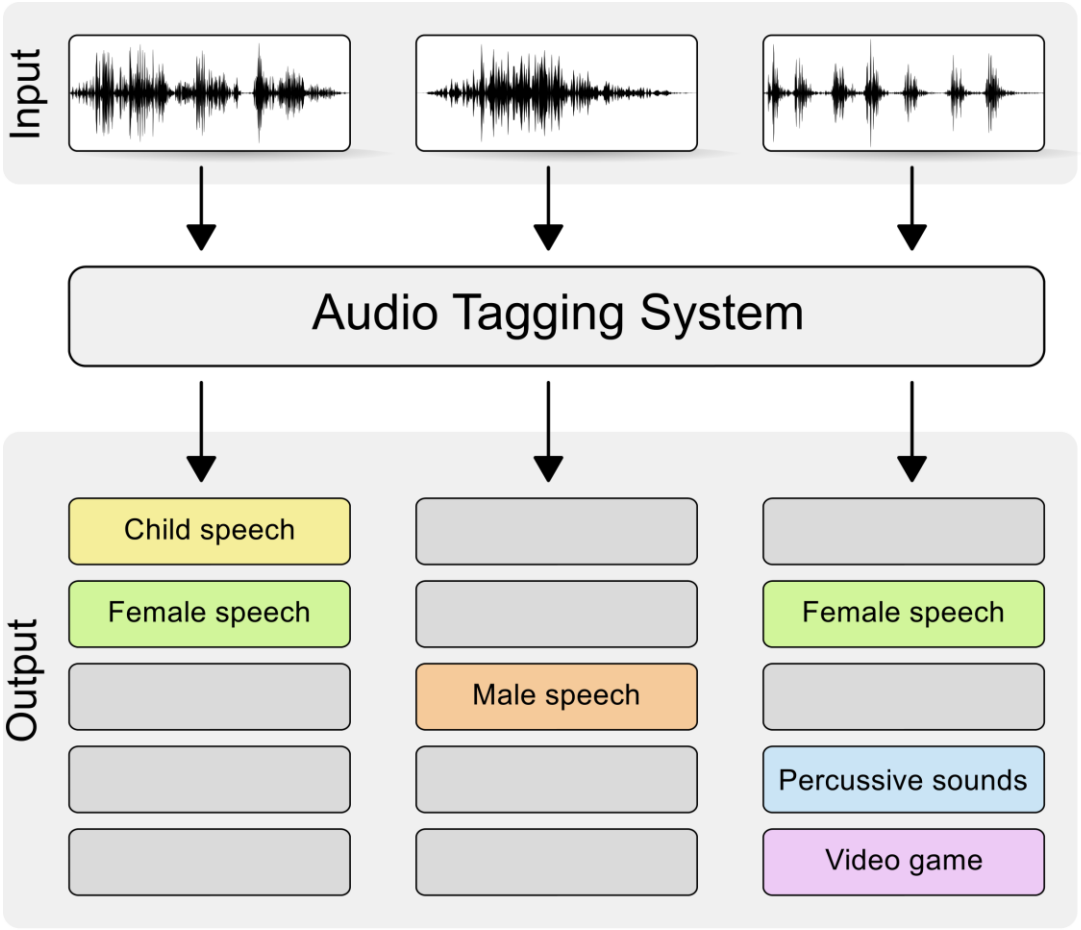


# Task 3: Event Detection, Real Life Audio

- 11 (home context) and 7 (residential area) classes (cutlery, drawer, walking / bird singing, car passing by, children shouting…)
- Manually produced annotations of real audio
- Development set
  - Home (indoor), 10 recordings, totaling 36 min
  - Residential area (outdoor), 12 recordings, totaling 42 min
  - In total 954 annotated events
- Evaluation set
  - 18 minutes of audio per context



# Task 4: Domestic Audio Tagging



# Task 4: Domestic Audio Tagging

- 7 label classes: child speech, adult male speech, adult female speech, video game/TV, percussive sounds, broadband noise, other identifiable sounds
- Annotations sourced using 3 human annotators, we indicate which 4-second audio chunks have strong annotator agreement
- To simulate commodity hardware, use 16 kHz monophonic audio
- Development set (4.9h): 4378 chunks, incl. 1946 strong agreement chunks
- Evaluation set (54min): 816 strong agreement chunks

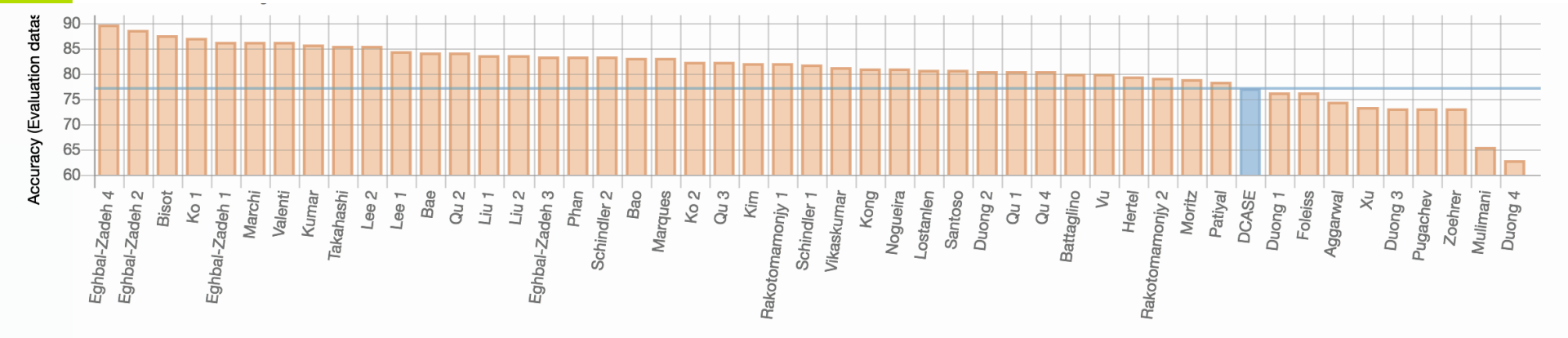
# Number of submissions

| Task | Submissions |
|---|---|
| 1 | 48 |
| 2 | 10 |
| 3 | 16 |
| 4 | 8 |
| **total** | **82** |

- Increased number of participants:
  - DCASE 2013: 24 submissions
  - DCASE 2016: 82 submissions

# Task 1 results

- 48 submissions / 34 teams / 113 authors



# Task 1 analysis of results

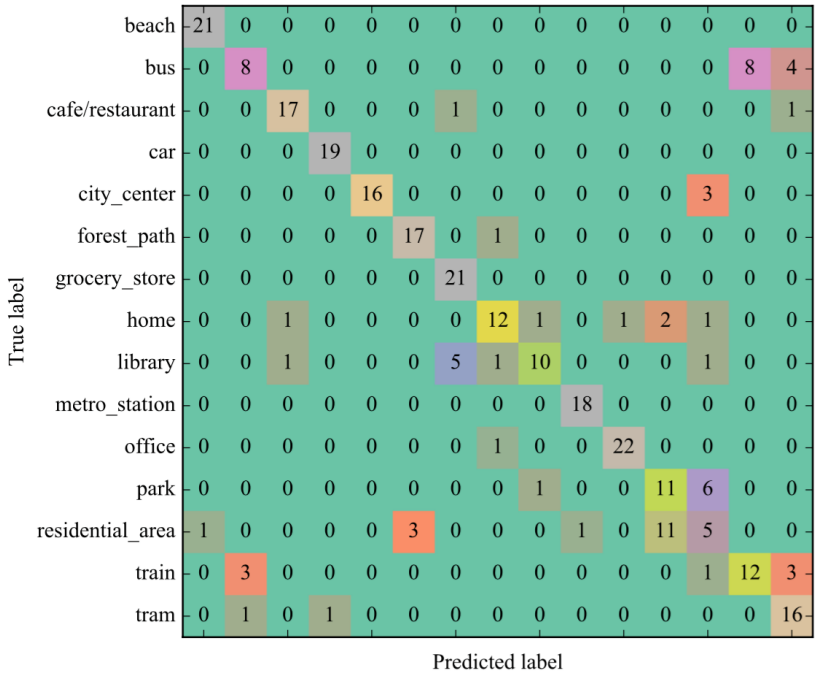
- Features: MFCCs or log-mel energies used in most systems
  - provide a reasonably good representation
- Also other features used in some systems, leading to improved results

# Task 1 analysis of results

- Most common classifiers:
  - 22 DNN based (enough data to learn deep models)
  - 10 SVM based
  - 10 ensemble classifiers
- Factor analysis methods (i-vectors, NMF) perform well
  - Each scene composed of multiple sources
- Fusion of classifiers leads to good results
- One-versus-all classifier for each class works well
- CNNs outperform MLPs or GMMs (SVMs also good, no direct comparison)

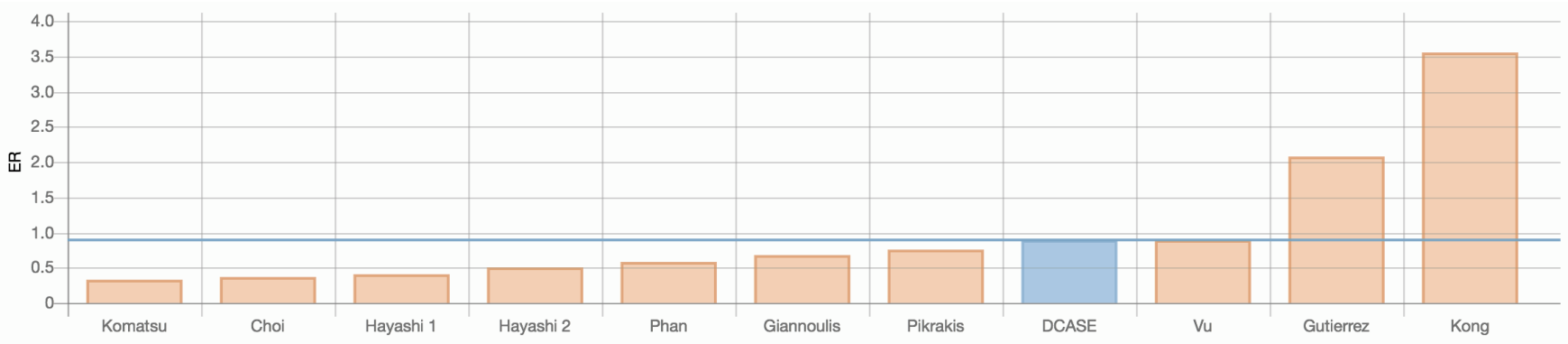
# Task 1 analysis of results

- Generalization properties
  - Most systems have comparable or better performance for evaluation compared to development dataset
  - Utilization of all development data improves results
  - The cross-validation setup needs to be carefully designed to avoid problems
- Some classes similar to each other and more difficult to recognize:
  - Bus / train / tram
  - Residential area / park



# Task 2 results

- 10 submissions / 9 teams / 37 authors



# Task 2 analysis of results

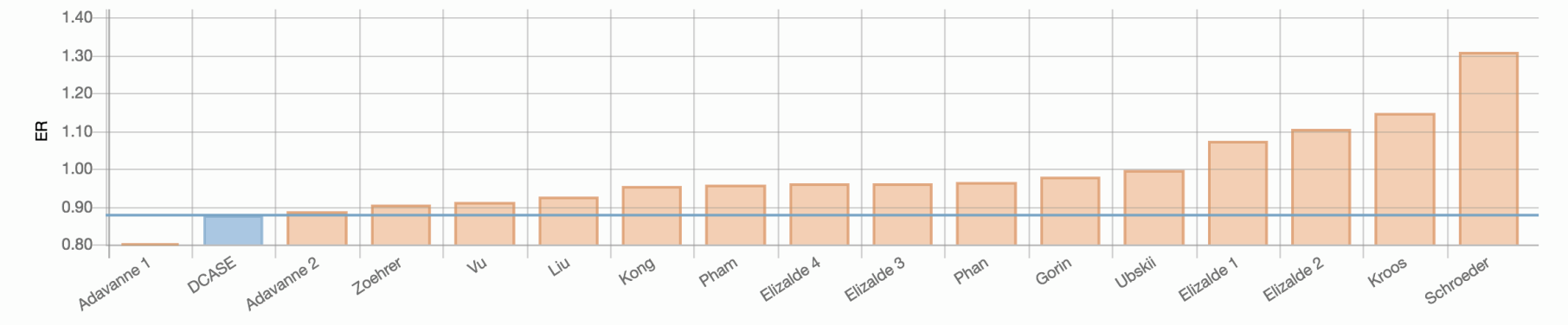
- Features: most methods use log-scale time-frequency representations (mel spectrograms, CQT, VQT)
- Classifiers
  - 5 DNN-based methods
  - 2 NMF-based methods
  - random forests, kNN, template matching
- Best results by NMF with Mixture of Local Dictionaries (Komatsu et al), followed by DNN (Choi et al) and BLSTM-based (Hayashi et al) methods
- Most systems report a drop in event-based metrics (which imply temporal tracking)

# Task 2 analysis of results

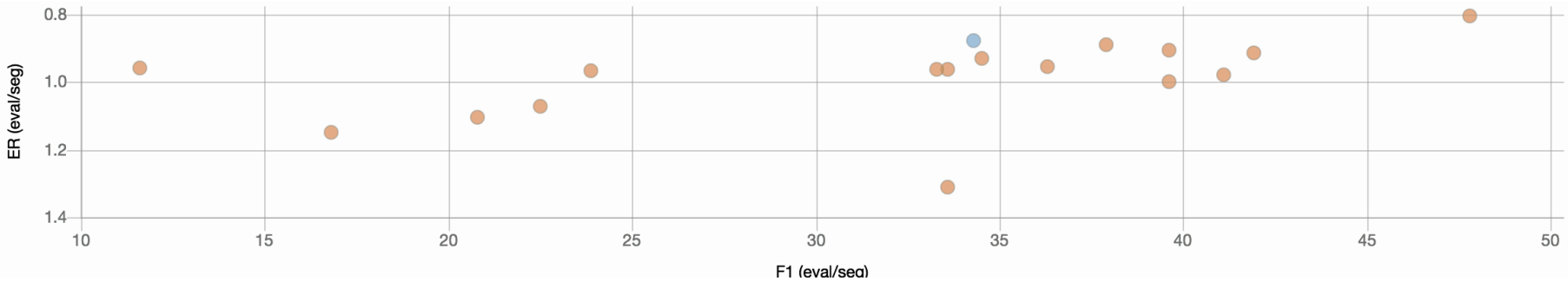
- Generalisation capabilities
  - Most systems report a significant drop in performance (10-30%) compared with results from the development dataset
- Results on sound classes differ: system by Komatsu et al reports F-score 90.7% on door knock, 37.7% on door slam

# Task 3 results

- 16 submissions / 12 teams / 45 authors



# Task 3 results



# Task 3 analysis of results

- Acoustic features
  - 9 systems using MFCCs
  - 4 systems use mel energies

  -> provide a reasonably good representation
  - Possible to obtain improvements by other features (e.g. Gabor filterbank, spatial features)

# Task 3 analysis of results

- Classifiers:
  - 7 DNN-based methods
  - 5 random forest based methods
  - 2 ensemble classifiers
- Top 7 submitted systems based on DNN
  - Easy way to do multilabel classification
- Second best system is the GMM baseline
  - Was extended in various ways (GMM-HMMs, tandem DNN-GMM)
- GMMs and DNNs perform better than NMF
- Temporal models effective: HMMs, LSTMs, CNNs

# Task 3 analysis of results

- Several submitted results where ER > 1
  - Did the participants optimize their systems for the F-score / not optimization of all system parameters?
- Residential Area context easier (ER 0.78) than the Home context (ER 0.91)
  - Resid. area classes clearly distinct (bird / car / children…)
  - Home classes more similar to each other
- Manual annotations are subjective and there is a degree of uncertainty
  - Affects evaluation scores and training of methods

# Task 3 analysis of results

- Top system (Adavanne) practically detects only most frequent classes
  - Home context 76% F-score on water tap and 16.5 % on washing dishes
  - Resid. area 62% F-score on bird singing, 76.7% on car passing by, 32% on wind blowing, other classes 0%
- Amount of sound events is unbalanced
  - Small number of instances is a problem for machine learning, especially for deep learning
  - Small classes are undetected by most systems

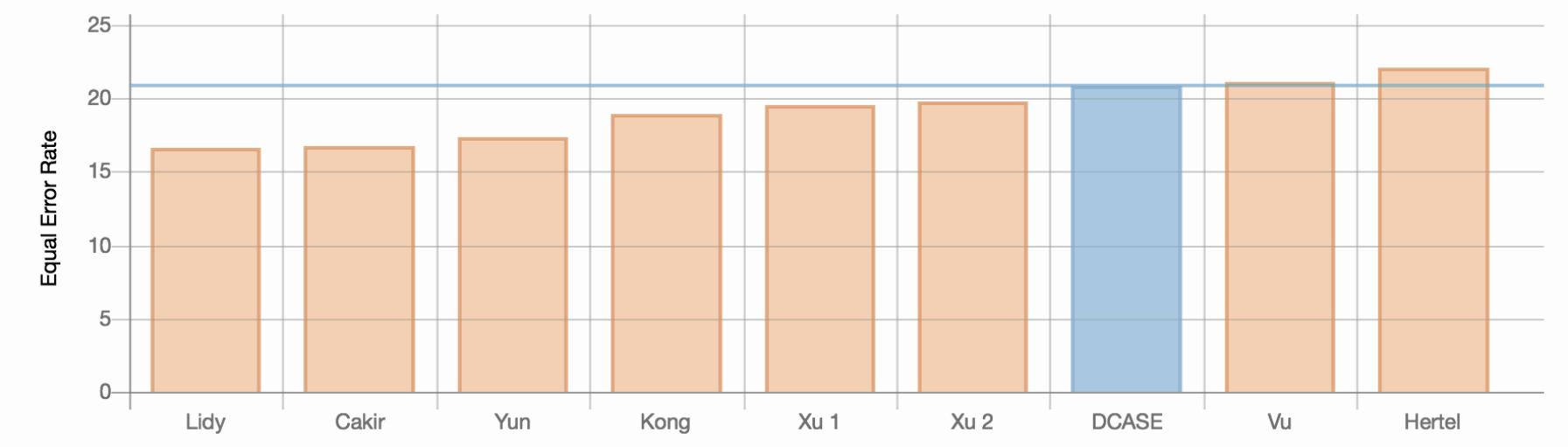
# Synthetic vs. real data

- Tasks 2 and 3 address the same task and use the same metrics, but use different material (synthetic vs. real)
- Large difference in results

| | Error rate | F-score |
|---|---|---|
| Task 2 (synthetic) | 0.33 | 80.2 % |
| Task 3 (real) | 0.81 | 47.8 % |

# Task 4 results

- 8 submissions / 7 teams / 23 authors



# Task 4 analysis of results

- 3 best-performing systems respectively use CQT features, Mel spectra, MFCCs
- Classifiers: 3 CNNs, 3 FNNs, 1 RNN, 1 GMM
- Both CNN- and GMM-based systems rank above alternative FNN-based systems

# Task 4 analysis of results

- Best-performing system (Lidy) outperforms baseline by 21%; 4.3 percentage points
- Averaging performance across systems reveals:
  - Least challenging label classes: Video Game/TV (6.1%), Broadband Noise (8.4%), Child Speech (20.5%)
  - Most challenging label classes: Other Identifiable Sounds (27.1%), Adult Male Speech (26.7%), Adult Female Speech (24.1%)

# General trends

- Emergence of deep neural network based methods
  - DCASE 2013: no DNN-based methods
  - 2016: majority of methods involve DNNs

-> Data-driven approaches replace manual design

-> Development of methods requires for more data

# Discussion