Bag-of-Features Acoustic Event Detection for Sensor Networks

Julian Kürby, René Grzeszick, Axel Plinge, and Gernot A. Fink

Pattern Recognition, Computer Science XII,
TU Dortmund University

September 3, 2016
DCASE Workshop
Budapest, Hungary
Motivation

Acoustic Sensor Networks (ASNs)

- are increasingly available: smartphones, laptops, hearing aids, ...
- offer the possibility of collaborative processing

Acoustic Event Detection (AED)

- useful for ASN applications [1]
- distributed sensors can improve performance [2]
- can we do better than heuristics? [3]

Method Overview

**Bag-of-Features**
- approach originating in text retrieval
- successful in AED [1]
- fast and online

**Multi-channel fusion**
- individual microphones or arrays as sensor node
- heuristic fusion: vote, max, product, ...
- learning based fusion: classifier stacking

**Processing pipeline**

---

Method (1/5) Features

- sliding window
- for each frame $k$, compute $y_k$
  perceptual loudness, MFCCs, and GFCCs [1]

---


Method (2/5) Quantization

- compute class-wise GMM by EM
- concatenate to super-codebook

\[ v_l = (l \cdot c + i) = (\mu_i, c, \sigma_i, c) \]

- quantize each frame \( k \) by super-codebook

\[ q_{k,l}(y_k, v_l) = \mathcal{N}(y_k | \mu_l, \sigma_l) \]

- histogram over a window of \( K \) frames

\[ b_l(Y_n, v_l) = \frac{1}{K} \sum_{k=1}^{K} q_{k,l}(y_k, v_l) \]


Method (3/5) Classification

Multinominal Bayes classification

- train with Lidstone smoothing
  \[
P(v_l | \Omega_c) = \frac{\alpha + \sum_{Y_n \in \Omega_c} b_l(Y_n, v_l)}{\alpha L + \sum_{m=1}^{L} \sum_{Y_n \in \Omega_c} b_m(Y_n, v_m)}
  \]

- all classes equally likely, i.e., have the same prior

- maximum likelihood classification
  \[
P(Y_n | \Omega_c) = \prod_{v_l \in v} P(v_l | \Omega_c)^{b_l(Y_n, v_l)}
  \]


Method (4/5) Fusion

BoF Models

- per channel,
- per array, or
- global

Method (4/5) Fusion

BoF Models
- per channel,
- per array, or
- global

Heuristic fusion [1]
- majority voting
  \[ \hat{c}(m) = \arg\max_c P_m(Y_{m,n}|\Omega_c) \]
  \[ \hat{c} = \arg\max_{c'} | \{ \hat{c}(m) = c' \} | \]

\[ \begin{aligned}
    & P_1(Y_{1,n}|\Omega_1) \ldots P_1(Y_{1,n}|\Omega_C) \\
    & P_1(Y_{1,n}|\Omega_2) \ldots P_M(Y_{2,n}|\Omega_C) \\
    & \vdots \\
    & P_1(Y_{1,n}|\Omega_C) \ldots P_M(Y_{M,n}|\Omega_C) \\
    & \arg\max_{c'} = c' \quad \arg\max_{c'} = c'
\end{aligned} \]

Method (4/5) Fusion

BoF Models
- per channel,
- per array, or
- global

Heuristic fusion [1]
- majority voting
  \[ \hat{c}_{(m)} = \arg\max_{c} P_m(Y_{m,n} | \Omega_c) \]
  \[ \hat{c} = \arg\max_{c'} \left| \{ \hat{c}_{(m)} = c' \} \right| \]
- maximum rule
  \[ \hat{c} = \arg\max_{c} \max_{m} P_m(Y_{m,n} | \Omega_c) \]

\[ \begin{aligned}
\arg\max_{c} \left\{ \max_{m} \{ P_1(Y_{1,n} | \Omega_1) \ldots P_M(Y_{M,n} | \Omega_1) \} \right. \\
\max_{m} \{ P_1(Y_{1,n} | \Omega_2) \ldots P_M(Y_{M,n} | \Omega_2) \} \\
\ldots \\
\left. \max_{m} \{ P_1(Y_{1,n} | \Omega_C) \ldots P_M(Y_{M,n} | \Omega_C) \} \right\} 
\end{aligned} \]

Method (4/5) Fusion

**BoF Models**
- per channel,
- per array, or
- global

**Heuristic fusion** [1]
- majority voting
  \[ \hat{c}(m) = \arg\max_c P_m(Y_{m,n}|\Omega_c) \]
  \[ \hat{c} = \arg\max_{c'} |\{ \hat{c}(m) = c' \} | \]
- maximum rule
  \[ \hat{c} = \arg\max_c \max_m P_m(Y_{m,n}|\Omega_c) \]
- product rule
  \[ \hat{c} = \arg\max_c \prod_m P_m(Y_{m,n}|\Omega_c) \]

---

Learned Fusion [1]

- classifier stacking – use a meta-learner instead of heuristics
- classification of the class-channel matrix

\[
\hat{c} = \mathcal{F}
\begin{pmatrix}
P_1(\mathbf{Y}_{1,n}|\Omega_1) & \ldots & P_M(\mathbf{Y}_{M,n}|\Omega_1) \\
P_1(\mathbf{Y}_{1,n}|\Omega_2) & \ldots & P_M(\mathbf{Y}_{M,n}|\Omega_2) \\
\vdots & & \vdots \\
P_1(\mathbf{Y}_{1,n}|\Omega_C) & \ldots & P_M(\mathbf{Y}_{M,n}|\Omega_C)
\end{pmatrix}
\]

- train a random forest classifier \( \mathcal{F} \)
  using data not used for training the models
- invariance through channel-sorting

\[
\text{argsort} \max_m \ P_m(\mathbf{Y}_{m,n}|\Omega_c)
\]

**ITC-Irst dataset** [1]

- smart conference room
- seven t-shaped arrays at the walls
- four microphones on the table
- *door knock, door slam, steps, chair moving, spoon (cup jingle), paper wrapping, key jingle, keyboard typing, phone ring, applause, cough, laugh, door open, phone vibration, mimo pen buzz, falling object, and unknown/background*
Evaluation ITC: Literature Comparison

- three training session days with events occurring at different positions
- third session used for training the stacking classifier
- forth session for test
- 12 first classes as foreground [1]

---


Evaluation ITC: Fusion strategies

- three training session days with events occurring at different positions
- third session used for training the stacking classifier
- forth session for test

frame-wise evaluation

- channel-specific models perform better
- stacking better than heuristics

---

FINCA dataset [1]

- new real-world recordings
- smart conference room
- two microphone arrays at the ceiling and two in the table
- circular, 8 mic, 10cm diameter
- applause, chairs, cups, door, doorbell, doorknock, keyboard, knock, music, paper, phonering, phonevibration, pouring, screen, speech, steps, streetnoise, touching, ventilator, and silence.

Evaluation FINCA: Fusion strategies

- five $2/3 - 1/3$ splits for training and test
- $1/3$ of training used for the stacking classifier
- silence as background

frame-wise evaluation

- channel-specific models perform better
- stacking better than heuristics

---


Evaluation FINCA: Position invariance

- classification of nine classes occurring at different positions in the room

mixed positions in training and test

![Graph showing error percentages for different models and methods.]

- stacking performs best
- sorting mitigates effect of unseen positions
- global models better for unseen positions


Conclusion

- acoustic sensor networks allow multi-channel AED
- multi-channel fusion improves the results
- classifier stacking outperforms heuristic strategies
- channel re-ordering by sorting can improve position invariance


References

Multi-microphone fusion for detection of speech and acoustic events in smart spaces.

Temporal acoustic words for online acoustic event detection.
In Proc. 37th German Conf. Pattern Recognition, Aachen, Germany, 2015.

Bag-of-features acoustic event detection for sensor networks.

H. Phan, M. Maass, L. Hertel, R. Mazur, and A. Mertins.
A multi-channel fusion framework for audio event detection.

A. Plinge and G. A. Fink.
Multi-speaker tracking using multiple distributed microphone arrays.

A. Plinge and S. Gannot.
Multi-microphone speech enhancement informed by auditory scene analysis.
A bag-of-features approach to acoustic event detection.

A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink.
Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms.

Clear evaluation of acoustic event detection and classification systems.

X. Zhao, Y. Shao, and D. Wang.
CASA-based robust speaker identification.