Bag-of-Features Acoustic Event Detection for Sensor Networks

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

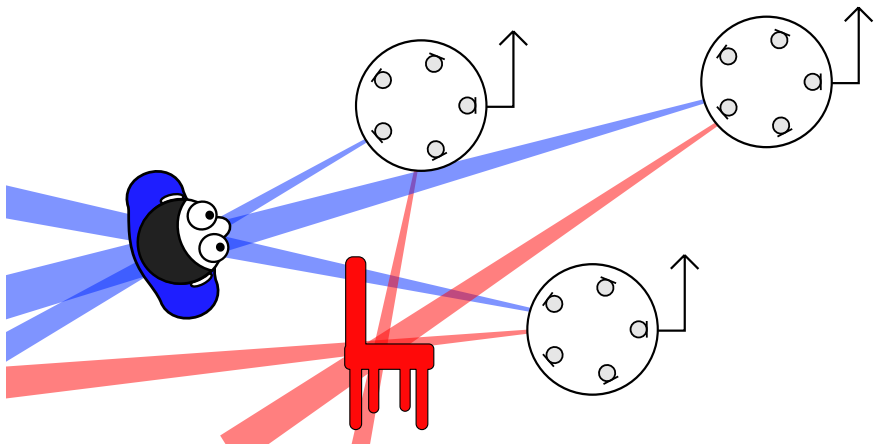
Pattern Recognition, Computer Science XII,
TU Dortmund University

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]

---

[1] 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. *IEEE Signal Process. Mag.*, 33(4):14–29, July 2016

[2] H. Phan, M. Maass, L. Hertel, R. Mazur, and A. Mertins. A multi-channel fusion framework for audio event detection. In *IEEE Workshop App. Signal Process. to Audio & Acoustics*, 2015

[3] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos. Multi-microphone fusion for detection of speech and acoustic events in smart spaces. In *European Signal Process. Conf.*, pages 2375–2379, Lisbon, Portugal, Sept. 2014

Method Overview

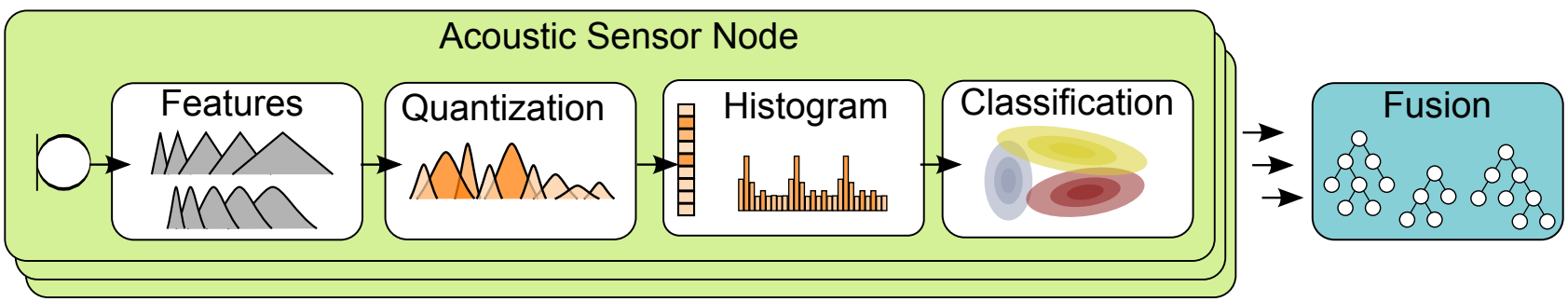
Bag-of-Features

- approach originating in text retrieval
- successful in AED [1]
- fast and online

Multi-channel fusion

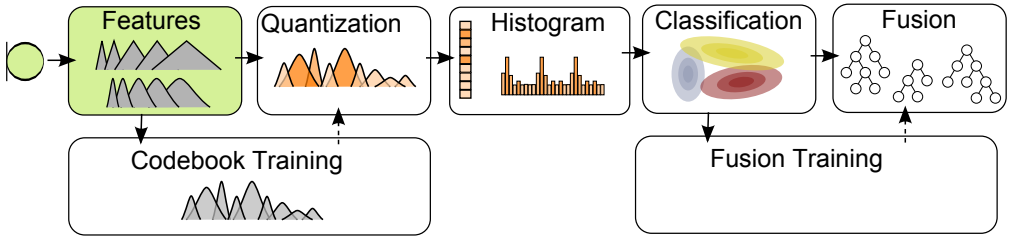
- individual microphones or arrays as sensor node
- heuristic fusion: vote, max, product, ...
- learning based fusion: classifier stacking

Processing pipeline

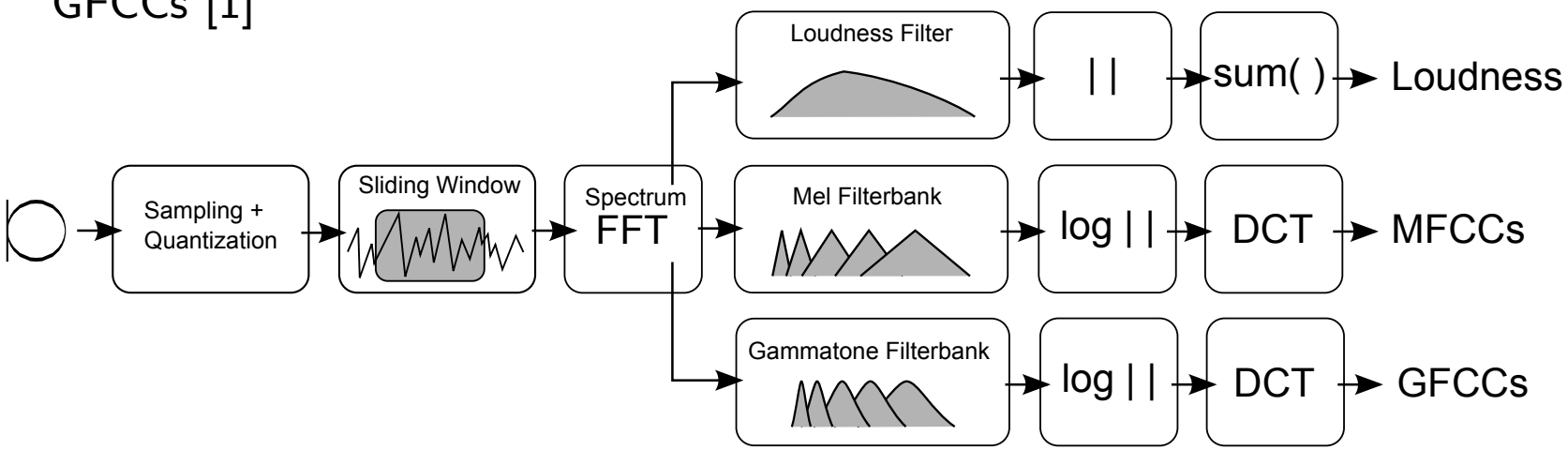


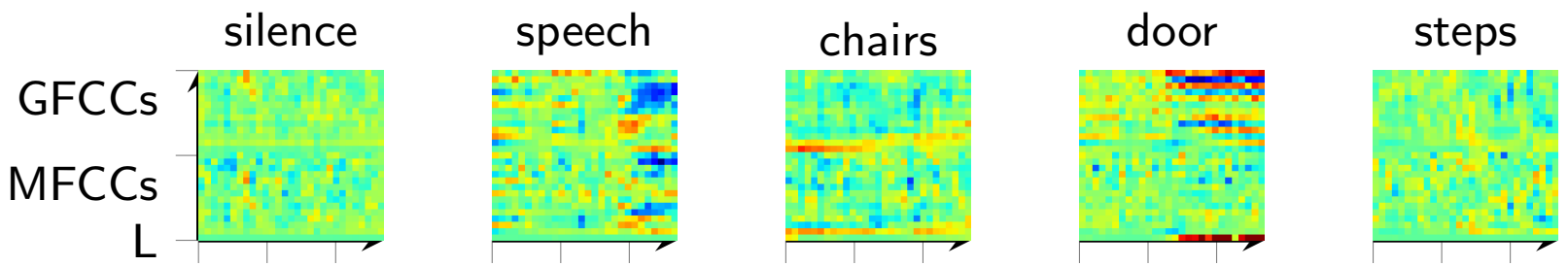
[1] A. Plinge, R. Grzeszick, and G. A. Fink. A bag-of-features approach to acoustic event detection. In *IEEE Int. Conf. Acoustics Speech & Signal Process.*, Florence, Italy, May 2014

Method (1/5) Features



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



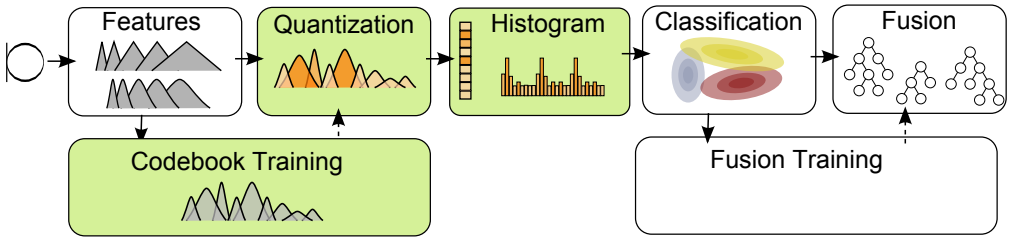


[1] X. Zhao, Y. Shao, and D. Wang. CASA-based robust speaker identification. *IEEE Trans. Audio, Speech, Language Process.*, 20(5):1608–1616, 2012

[2] A. Plinge, R. Grzeszick, and G. A. Fink. A bag-of-features approach to acoustic event detection. In *IEEE Int. Conf. Acoustics Speech & Signal Process.*, Florence, Italy, May 2014

[3] code at `http://patrec.cs.tu-dortmund.de/resources`

Method (2/5) Quantization



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

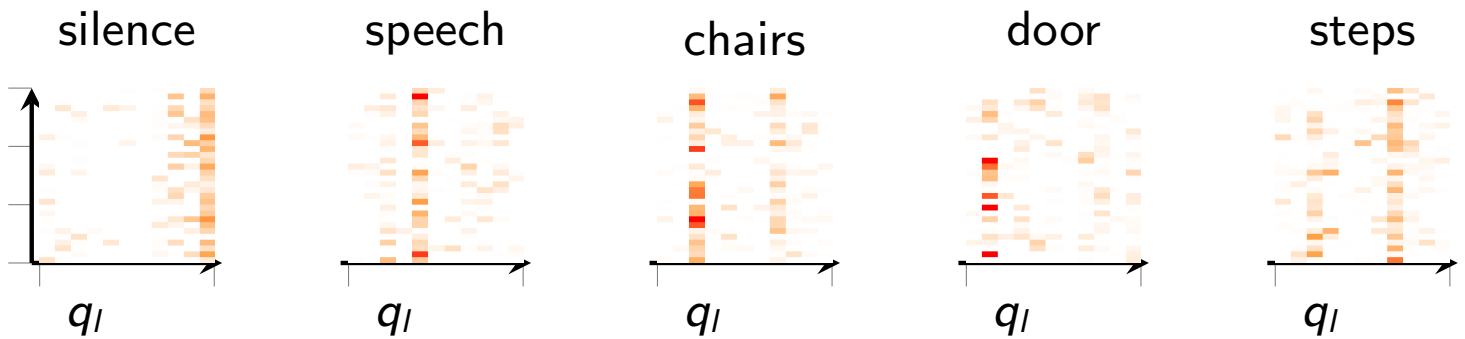
$$v_{l=(I\cdot c+i)} = (\mu_{i,c}, \sigma_{i,c})$$

- quantize each frame k by super-codebook

$$q_{k,l}(\boldsymbol{y}_k, v_l) = \mathcal{N}(\boldsymbol{y}_k | \mu_l, \sigma_l)$$

- histogram over a window of K frames

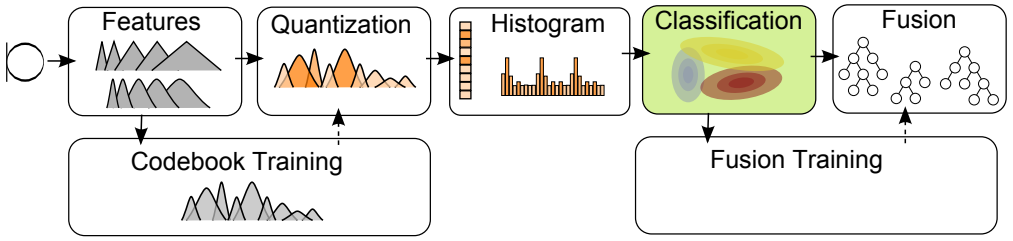
$$b_l(Y_n, v_l) = \frac{1}{K} \sum_{k=1}^{K} q_{k,l}(\boldsymbol{y}_k, v_l)$$



[1] A. Plinge, R. Grzeszick, and G. A. Fink. A bag-of-features approach to acoustic event detection. In *IEEE Int. Conf. Acoustics Speech & Signal Process.*, Florence, Italy, May 2014

[2] code at `http://patrec.cs.tu-dortmund.de/resources`

Method (3/5) Classification



Multinominal Bayes classification

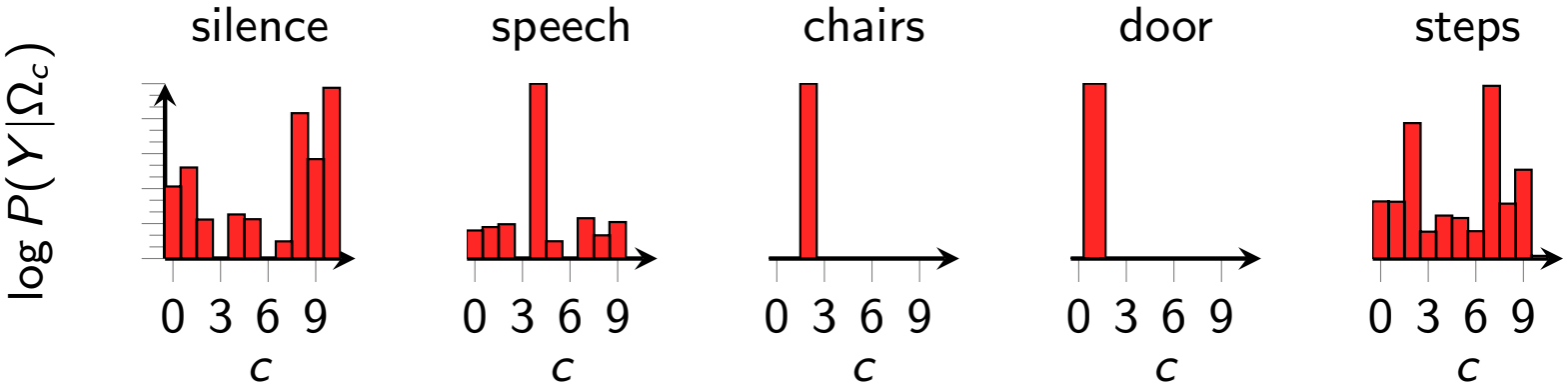
- train with Lidstone smoothing

$$P(v_l|\Omega_c) = \frac{\alpha + \sum_{Y_n \in \Omega_c} b_l(Y_n, v_l)}{\alpha L + \sum_{m=1}^{L} \sum_{Y_n \in \Omega_c} b_m(Y_n, v_m)}$$

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

→ maximum likelihood classification

$$P(Y_n|\Omega_c) = \prod_{v_l \in \mathbf{v}} P(v_l|\Omega_c)^{b_l(Y_n, v_l)}$$



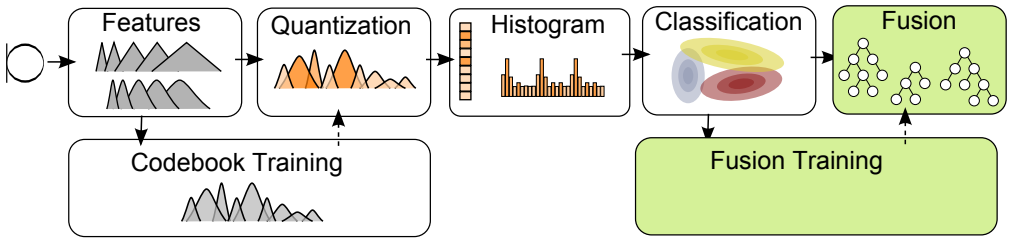
[1] A. Plinge, R. Grzeszick, and G. A. Fink. A bag-of-features approach to acoustic event detection. In *IEEE Int. Conf. Acoustics Speech & Signal Process.*, Florence, Italy, May 2014

[2] code at `http://patrec.cs.tu-dortmund.de/resources`

Method (4/5) Fusion

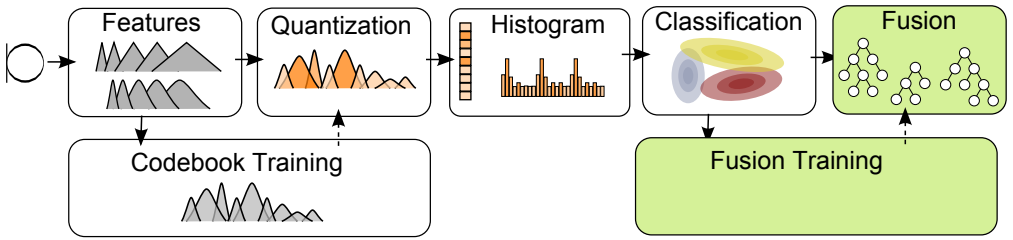
BoF Models

- per channel,
- per array, or
- global



[1] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos. Multi-microphone fusion for detection of speech and acoustic events in smart spaces. In *European Signal Process. Conf.*, pages 2375–2379, Lisbon, Portugal, Sept. 2014

Method (4/5) Fusion



BoF Models

- per channel,
- per array, or
- global

Heuristic fusion [1]

- majority voting

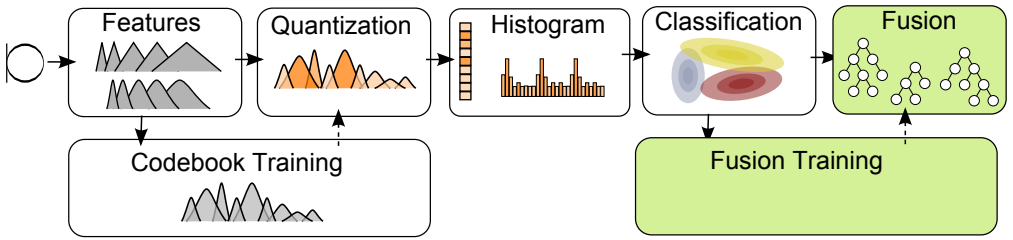
$$\hat{c}_{(m)} = \underset{c}{\operatorname{argmax}}\ P_m(\boldsymbol{Y}_{m,n}|\Omega_c)$$

$$\hat{c} = \operatorname{argmax}_{c'} |\{\hat{c}_{(m)} = c'\}|$$

$$\operatorname{argmax}_{c'} \left| \left\{ \begin{array}{ccc} P_1(\boldsymbol{Y}_{1,n}|\Omega_1) & \dots & P_1(\boldsymbol{Y}_{1,n}|\Omega_C) \\ P_1(\boldsymbol{Y}_{1,n}|\Omega_2) & \dots & P_M(\boldsymbol{Y}_{2,n}|\Omega_C) \\ \vdots & & \vdots \\ \underbrace{P_1(\boldsymbol{Y}_{1,n}|\Omega_C)}_{\operatorname{argmax}_c = c'} & \dots & \underbrace{P_M(\boldsymbol{Y}_{M,n}|\Omega_C)}_{\operatorname{argmax}_c = c'} \end{array} \right\} \right|$$

[1] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos. Multi-microphone fusion for detection of speech and acoustic events in smart spaces. In *European Signal Process. Conf.*, pages 2375–2379, Lisbon, Portugal, Sept. 2014

Method (4/5) Fusion



BoF Models

- per channel,
- per array, or
- global

Heuristic fusion [1]

- majority voting

$$\hat{c}_{(m)} = \underset{c}{\operatorname{argmax}}\ P_m(\boldsymbol{Y}_{m,n}|\Omega_c)$$

$$\hat{c} = \operatorname{argmax}_{c'} |\{\hat{c}_{(m)} = c'\}|$$

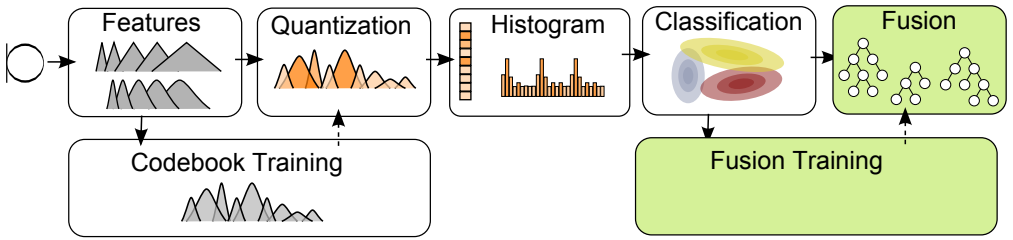
- maximum rule

$$\hat{c} = \underset{c}{\operatorname{argmax}}\ \max_{m}\ P_m(\boldsymbol{Y}_{m,n}|\Omega_c)$$

$$\operatorname{argmax}_c \left\{ \begin{array}{c} \max_m\{P_1(\boldsymbol{Y}_{1,n}|\Omega_1) \ldots P_M(\boldsymbol{Y}_{M,n}|\Omega_1)\} \\ \max_m\{P_1(\boldsymbol{Y}_{1,n}|\Omega_2) \ldots P_M(\boldsymbol{Y}_{M,n}|\Omega_2)\} \\ \ldots \\ \max_m\{P_1(\boldsymbol{Y}_{1,n}|\Omega_C) \ldots P_M(\boldsymbol{Y}_{M,n}|\Omega_C)\} \end{array} \right\}$$

[1] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos. Multi-microphone fusion for detection of speech and acoustic events in smart spaces. In *European Signal Process. Conf.*, pages 2375–2379, Lisbon, Portugal, Sept. 2014

Method (4/5) Fusion



BoF Models

- per channel,
- per array, or
- global

Heuristic fusion [1]

- majority voting

$$\hat{c}_{(m)} = \underset{c}{\operatorname{argmax}}\ P_m(\boldsymbol{Y}_{m,n}|\Omega_c)$$

$$\hat{c} = \operatorname{argmax}_{c'} |\{\hat{c}_{(m)} = c'\}|$$

- maximum rule

$$\hat{c} = \underset{c}{\operatorname{argmax}}\ \max_{m}\ P_m(\boldsymbol{Y}_{m,n}|\Omega_c)$$

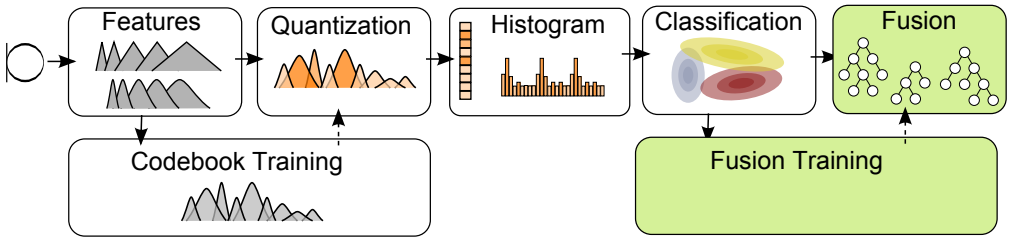
- product rule

$$\hat{c} = \underset{c}{\operatorname{argmax}} \prod_{m} P_m(\boldsymbol{Y}_{m,n}|\Omega_c)$$

$$\operatorname{argmax}_c \left\{ \begin{array}{c} P_1(\boldsymbol{Y}_{1,n}|\Omega_1) \cdot P_2(\boldsymbol{Y}_{2,n}|\Omega_1) \cdot \ldots P_M(\boldsymbol{Y}_{M,n}|\Omega_1) \\ P_1(\boldsymbol{Y}_{1,n}|\Omega_2) \cdot P_2(\boldsymbol{Y}_{2,n}|\Omega_2) \cdot \ldots P_M(\boldsymbol{Y}_{M,n}|\Omega_1) \\ \vdots \\ P_1(\boldsymbol{Y}_{1,n}|\Omega_C) \cdot P_2(\boldsymbol{Y}_{2,n}|\Omega_C) \cdot \ldots P_M(\boldsymbol{Y}_{M,n}|\Omega_1) \end{array} \right\}$$

[1] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos. Multi-microphone fusion for detection of speech and acoustic events in smart spaces. In *European Signal Process. Conf.*, pages 2375–2379, Lisbon, Portugal, Sept. 2014

Method (5/5) Fusion



Learned Fusion [1]

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

$$\hat{c} = \mathcal{F} \begin{pmatrix} P_1(\boldsymbol{Y}_{1,n}|\Omega_1) & \dots & P_M(\boldsymbol{Y}_{M,n}|\Omega_1) \\ P_1(\boldsymbol{Y}_{1,n}|\Omega_2) & \dots & P_M(\boldsymbol{Y}_{M,n}|\Omega_2) \\ & \ddots & \\ P_1(\boldsymbol{Y}_{1,n}|\Omega_C) & \dots & P_M(\boldsymbol{Y}_{M,n}|\Omega_C) \end{pmatrix}$$

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

$$\underset{m}{\operatorname{argsort}} \max_{c} \; P_m(\boldsymbol{Y}_{m,n}|\Omega_c)$$

---

[1] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink. Bag-of-features acoustic event detection for sensor networks. In *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Budapest, Hungary, Sept. 2016

Evaluation ITC: dataset

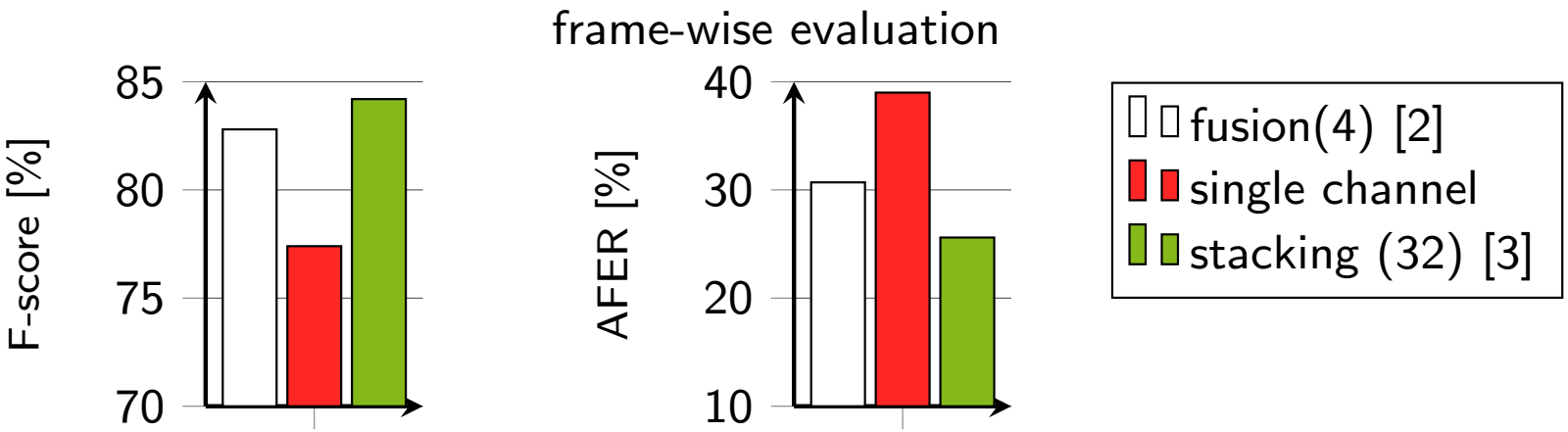
ITC-Irst dataset [1]

- 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*

[1] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo. Clear evaluation of acoustic event detection and classification systems. In R. Stiefelhagen and J. Garofolo, editors, *Multimodal Technologies for Perception of Humans*, volume 4122 of *Lecture Notes in Computer Science*, pages 311–322. Springer Berlin Heidelberg, 2007

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]



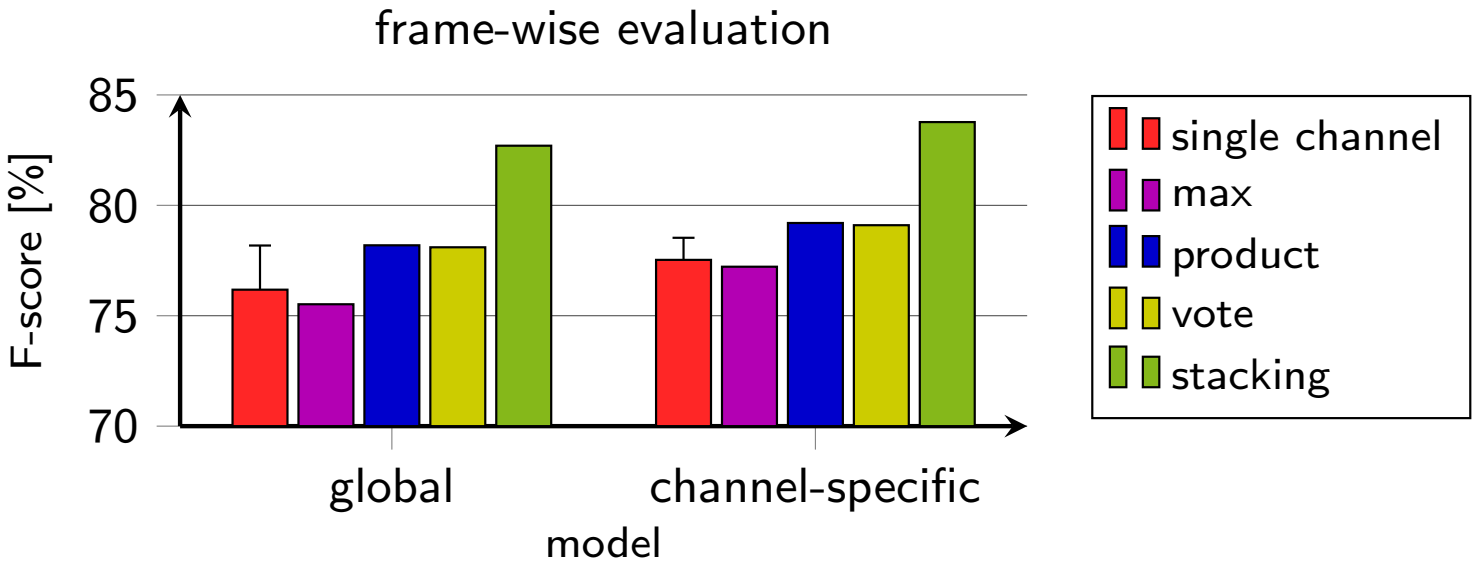
[1] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo. Clear evaluation of acoustic event detection and classification systems. In R. Stiefelhagen and J. Garofolo, editors, *Multimodal Technologies for Perception of Humans*, volume 4122 of *Lecture Notes in Computer Science*, pages 311–322. Springer Berlin Heidelberg, 2007

[2] H. Phan, M. Maass, L. Hertel, R. Mazur, and A. Mertins. A multi-channel fusion framework for audio event detection. In *IEEE Workshop App. Signal Process. to Audio & Acoustics*, 2015

[3] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink. Bag-of-features acoustic event detection for sensor networks. In *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Budapest, Hungary, Sept. 2016

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



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

[1] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink. Bag-of-features acoustic event detection for sensor networks. In *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Budapest, Hungary, Sept. 2016

Evaluation: FINCA dataset

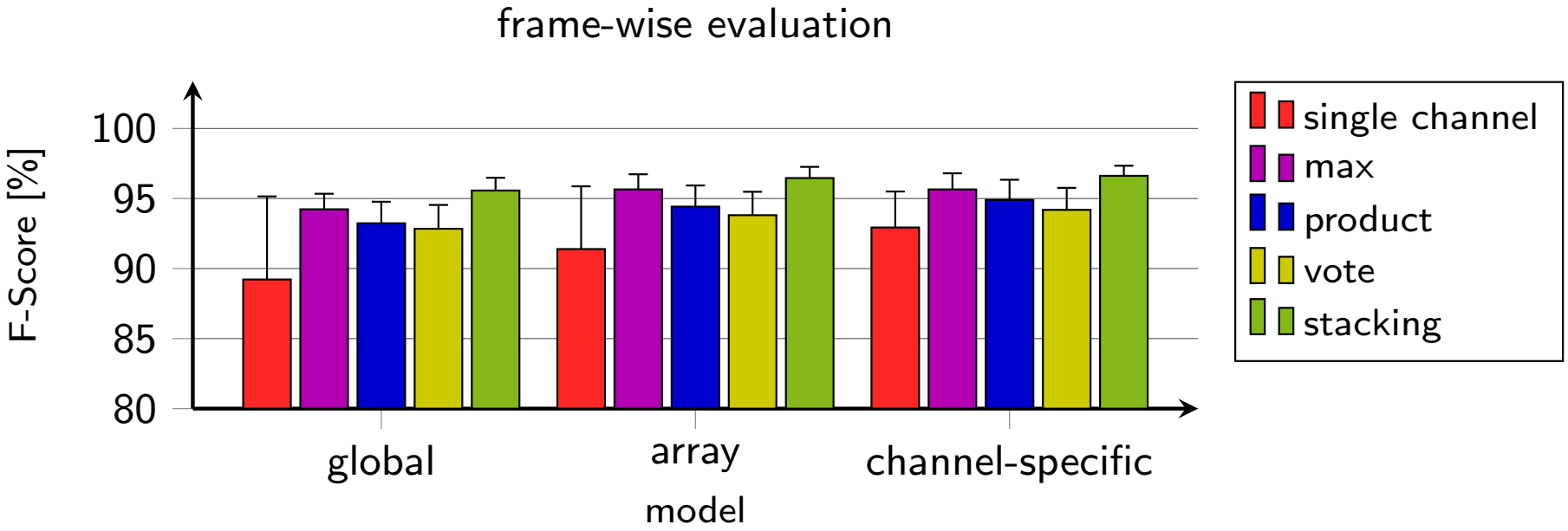
FINCA dataset [1]

- 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*.

[1] dataset available at `http://patrec.cs.tu-dortmund.de/resources`

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



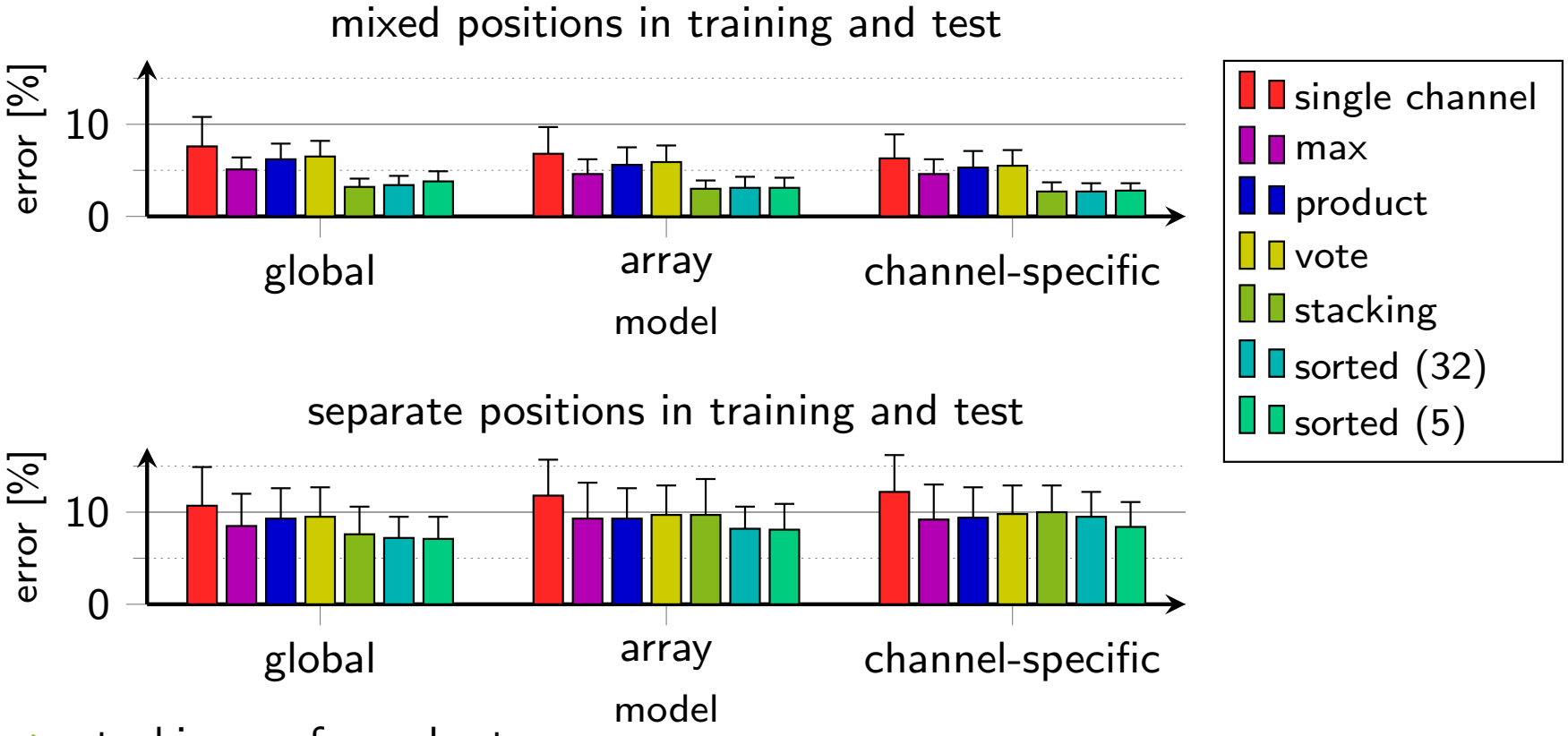
- channel-specific models perform better
- stacking better than heuristics

[1] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink. Bag-of-features acoustic event detection for sensor networks. In *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Budapest, Hungary, Sept. 2016

[2] dataset available at `http://patrec.cs.tu-dortmund.de/resources`

Evaluation FINCA: Position invariance

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



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

[1] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink. Bag-of-features acoustic event detection for sensor networks. In *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Budapest, Hungary, Sept. 2016

[2] dataset available at `http://patrec.cs.tu-dortmund.de/resources`

Conclusion

- acoustic sensor networks allow multi-channel AED
- extension [1] of Bag-of-Features online AED [2]
- multi-channel fusion improves the results
- classifier stacking outperforms heuristic strategies
- channel re-ordering by sorting can improve position invariance

[1] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink. Bag-of-features acoustic event detection for sensor networks. In *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Budapest, Hungary, Sept. 2016

[2] R. Grzeszick, A. Plinge, and G. A. Fink. Temporal acoustic words for online acoustic event detection. In *Proc. 37th German Conf. Pattern Recognition*, Aachen, Germany, 2015

[3] `http://patrec.cs.tu-dortmund.de/resources`

References

P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos.
Multi-microphone fusion for detection of speech and acoustic events in smart spaces.
In *European Signal Process. Conf.*, pages 2375–2379, Lisbon, Portugal, Sept. 2014.

R. Grzeszick, A. Plinge, and G. A. Fink.
Temporal acoustic words for online acoustic event detection.
In *Proc. 37th German Conf. Pattern Recognition*, Aachen, Germany, 2015.

J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink.
Bag-of-features acoustic event detection for sensor networks.
In *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Budapest, Hungary, Sept. 2016.

H. Phan, M. Maass, L. Hertel, R. Mazur, and A. Mertins.
A multi-channel fusion framework for audio event detection.
In *IEEE Workshop App. Signal Process. to Audio & Acoustics*, 2015.

A. Plinge and G. A. Fink.
Multi-speaker tracking using multiple distributed microphone arrays.
In *IEEE Int. Conf. Acoustics Speech & Signal Process.*, Florence, Italy, May 2014.

A. Plinge and S. Gannot.
Multi-microphone speech enhancement informed by auditory scene analysis.
In *Sensor Array and Multichannel Signal Process. Workshop*, Rio de Janeiro, Brazil, July 2016.

A. Plinge, R. Grzeszick, and G. A. Fink.
A bag-of-features approach to acoustic event detection.
In *IEEE Int. Conf. Acoustics Speech & Signal Process.*, Florence, Italy, May 2014.

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.
IEEE Signal Process. Mag., 33(4):14–29, July 2016.

A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo.
Clear evaluation of acoustic event detection and classification systems.
In R. Stiefelhagen and J. Garofolo, editors, *Multimodal Technologies for Perception of Humans*, volume 4122 of *Lecture Notes in Computer Science*, pages 311–322. Springer Berlin Heidelberg, 2007.

X. Zhao, Y. Shao, and D. Wang.
CASA-based robust speaker identification.
IEEE Trans. Audio, Speech, Language Process., 20(5):1608–1616, 2012.