Enabling Early Audio Event Detection with Neural Networks

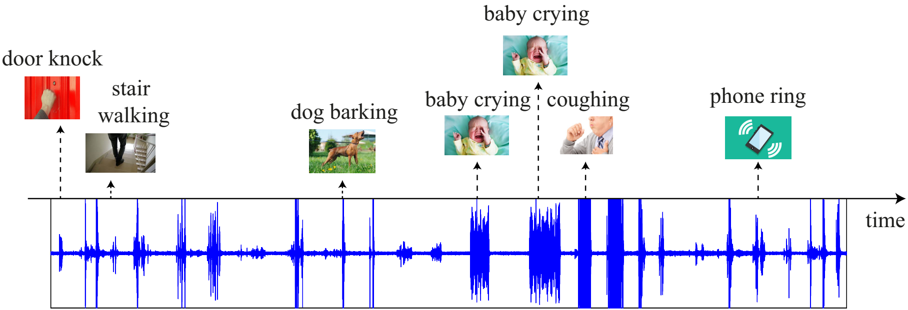
Huy Phan[⋆], Philipp Koch[†], Ian McLoughlin[‡] and Alfred Mertins[†]

[⋆]Department of Engineering Science, University of Oxford, UK
[‡]School of Computing, University of Kent, UK
[†]Institute for Signal Processing, University of Lübeck, Germany

ICASSP 2018, Calgary, Canada

Introduction

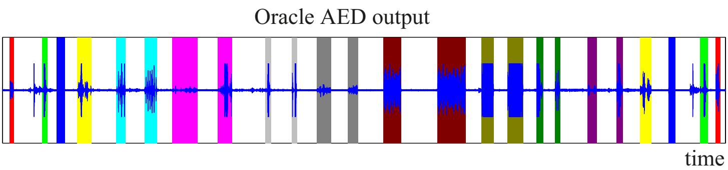
Introduction



- Hearing is for getting information from sound
- Environmental sound recognition is fundamental
- ‘Events’ are what we hear and notice
- What and when?

D. Ellis,“Recognizing and Classifying Environmental Sounds,” in *CHiME workshop*, 2013

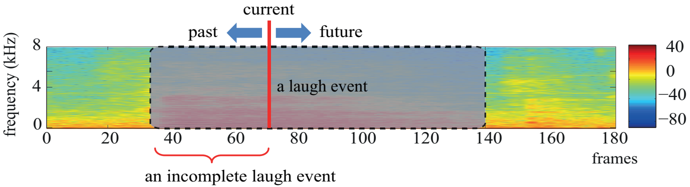
Audio Event Detection (AED)



- AED Task
  - What type of events? Where in time do they happen?
- Approaches
  - Detection-by-classification: DNN [McLoughlin et al., 2015], CNN [McLoughlin et al., 2017], RNN [Parascandolo et al., 2016], CRNN [Çakir et al. 2017], etc.
  - Joint detection and segmentation: GMM-HMM [Mesaros et al., 2010; Heittola et al. 2013], etc.
  - Onset and offset detection: Regression Forests [Phan et al., 2015], Classification-Regression Forests [Phan et al., 2016], etc.

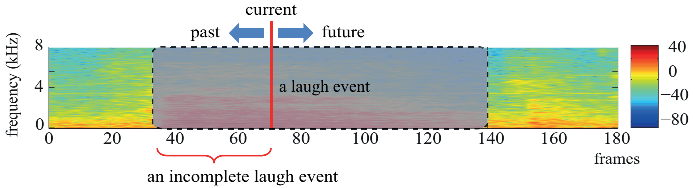
Many of previous works focused on detection of entire audio events

Early Event Detection in Audio Streams



H. Phan, M. Maass, R. Mazur, and A. Mertins,"Early Event Detection in Audio Streams," in *Proc. ICME*, pp. 1-6, 2015.

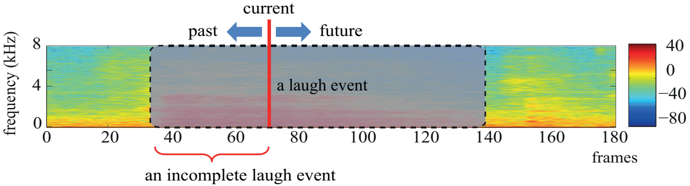
Early Event Detection in Audio Streams



- Early detection of ongoing events with their partial observation

H. Phan, M. Maass, R. Mazur, and A. Mertins,"Early Event Detection in Audio Streams," in *Proc. ICME*, pp. 1-6, 2015.

Early Event Detection in Audio Streams

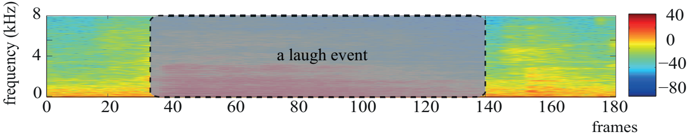


- Early detection of ongoing events with their partial observation
- Reliability: Early detection without losing detection performance
  $\Rightarrow$ Requiring the monotonicity of a detection function

---

H. Phan, M. Maass, R. Mazur, and A. Mertins,"Early Event Detection in Audio Streams," in *Proc. ICME*, pp. 1-6, 2015.

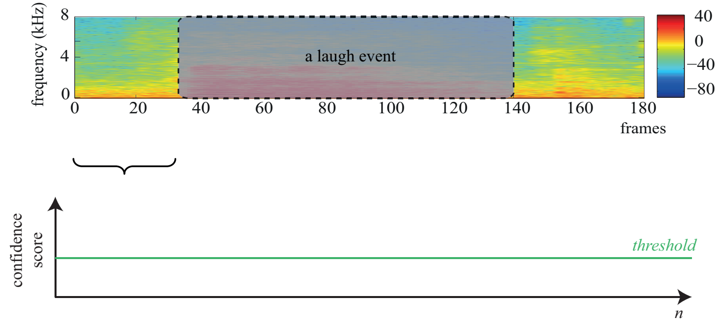
Early Event Detection in Audio Streams



- Early detection of ongoing events with their partial observation
- Reliability: Early detection without losing detection performance
  $\Rightarrow$ Requiring the monotonicity of a detection function

H. Phan, M. Maass, R. Mazur, and A. Mertins,"Early Event Detection in Audio Streams," in *Proc. ICME*, pp. 1-6, 2015.

Early Event Detection in Audio Streams

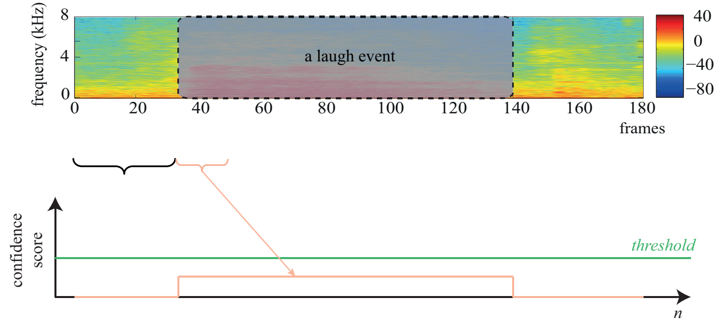


- Early detection of ongoing events with their partial observation
- Reliability: Early detection without losing detection performance
  $\Rightarrow$ Requiring the monotonicity of a detection function

---

H. Phan, M. Maass, R. Mazur, and A. Mertins,"Early Event Detection in Audio Streams," in *Proc. ICME*, pp. 1-6, 2015.

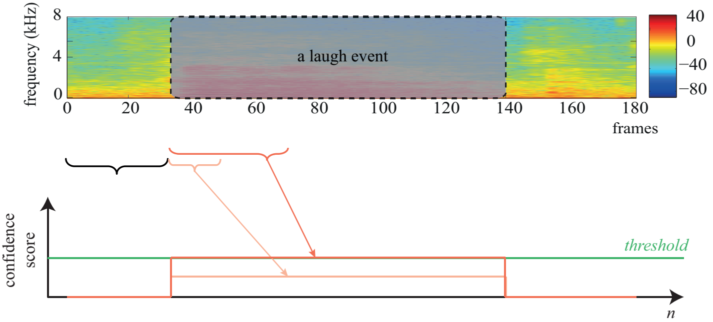
Early Event Detection in Audio Streams



- Early detection of ongoing events with their partial observation
- Reliability: Early detection without losing detection performance
  $\Rightarrow$ Requiring the monotonicity of a detection function

H. Phan, M. Maass, R. Mazur, and A. Mertins,"Early Event Detection in Audio Streams," in *Proc. ICME*, pp. 1-6, 2015.

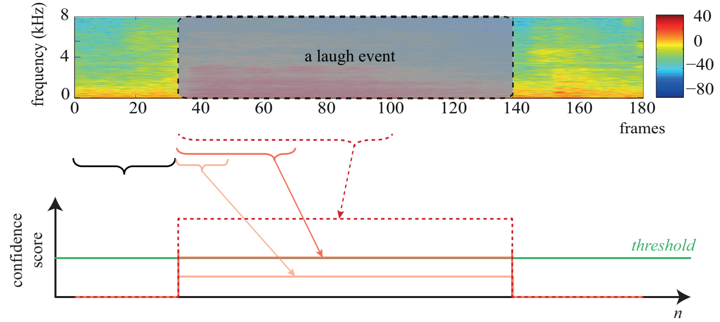
Early Event Detection in Audio Streams



- Early detection of ongoing events with their partial observation
- Reliability: Early detection without losing detection performance
  ⇒ Requiring the monotonicity of a detection function

H. Phan, M. Maass, R. Mazur, and A. Mertins,"Early Event Detection in Audio Streams," in *Proc. ICME*, pp. 1-6, 2015.

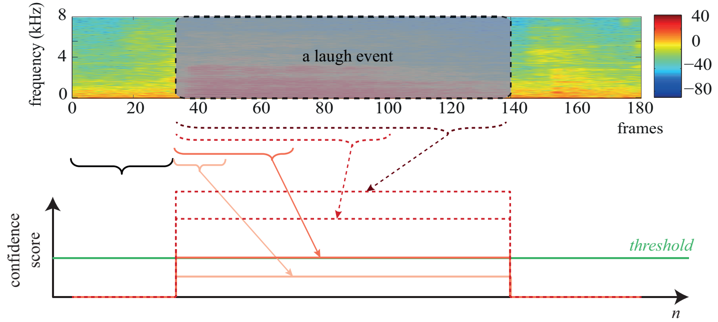
Early Event Detection in Audio Streams



- Early detection of ongoing events with their partial observation
- Reliability: Early detection without losing detection performance
  $\Rightarrow$ Requiring the monotonicity of a detection function

H. Phan, M. Maass, R. Mazur, and A. Mertins,"Early Event Detection in Audio Streams," in *Proc. ICME*, pp. 1-6, 2015.

Early Event Detection in Audio Streams

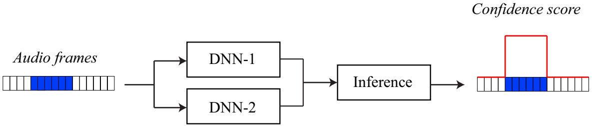


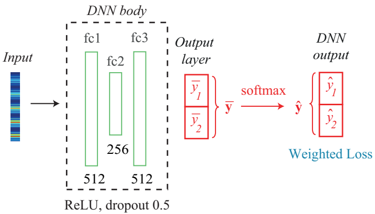
- Early detection of ongoing events with their partial observation
- Reliability: Early detection without losing detection performance
  $\Rightarrow$ Requiring the monotonicity of a detection function

H. Phan, M. Maass, R. Mazur, and A. Mertins,"Early Event Detection in Audio Streams," in *Proc. ICME*, pp. 1-6, 2015.

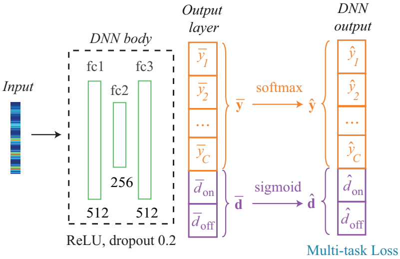
Dual-DNN System

Dual-DNN Detection System





DNN-1



DNN-2

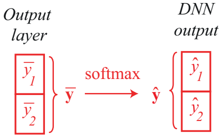
- Background/foreground classifier
- Joint event classification and event boundary estimation

DNN-1: Background/Foreground Classification

- Weighting loss:

$$E_{\mathrm{w}}(\boldsymbol{\theta}) = -\frac{1}{N}\left(\lambda_{\mathrm{fg}} \sum_{n=1}^{N} \mathbb{I}_{\mathrm{fg}}(\mathbf{x}_n)\mathbf{y}_n \log\left(\hat{\mathbf{y}}_n(\mathbf{x}_n, \boldsymbol{\theta})\right) + \lambda_{\mathrm{bg}} \sum_{n=1}^{N} \mathbb{I}_{\mathrm{bg}}(\mathbf{x}_n)\mathbf{y}_n \log\left(\hat{\mathbf{y}}_n(\mathbf{x}_n, \boldsymbol{\theta})\right)\right) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

  - $\mathbb{I}_{\mathrm{bg}}(\mathbf{x})$: Indicator function, 1 if $\mathbf{x}$ is background and 0 if not
  - $\mathbb{I}_{\mathrm{fg}}(\mathbf{x})$: Indicator function, 1 if $\mathbf{x}$ is foreground and 0 if not
  - λ_{fg}: Penalization weight for false negative errors
  - λ_{bg}: Penalization weight for false positive errors

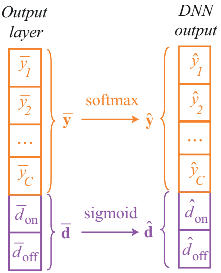


DNN-2: Joint Event Classification & Boundary Estimation

- Multi-task loss:

$$E_{\text{mt}}(\boldsymbol{\theta}) = \lambda_{\text{class}} E_{\text{class}}(\boldsymbol{\theta}) + \lambda_{\text{dist}} E_{\text{dist}}(\boldsymbol{\theta}) + \lambda_{\text{conf}} E_{\text{conf}}(\boldsymbol{\theta}) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2$$

  - E_{mt}: Total loss
  - E_{class}: Class loss
  - E_{dist}: Distance loss
  - E_{dist}: Confidence loss
  - λ_{class}, λ_{dist}, λ_{conf}: Weights of the individual losses

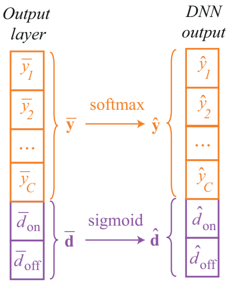


DNN-2: Joint Event Classification & Boundary Estimation

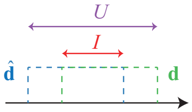
$$E_{\text{class}}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_n \log\left(\hat{\mathbf{y}}_n(\mathbf{x}_n\,,\boldsymbol{\theta})\right)$$

$$E_{\text{dist}}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N} \left\|\mathbf{d} - \hat{\mathbf{d}}_n\left(\mathbf{x}_n, \boldsymbol{\theta}\right)\right\|_2^2$$

$$E_{\text{conf}}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N} \left\|\mathbf{y}_n - \hat{\mathbf{y}}_n \frac{I\left(\mathbf{d}_n, \hat{\mathbf{d}}_n\left(\mathbf{x}_n, \boldsymbol{\theta}\right)\right)}{U\left(\mathbf{d}_n, \hat{\mathbf{d}}_n\left(\mathbf{x}_n, \boldsymbol{\theta}\right)\right)}\right\|_2^2$$



$$I\left(\mathbf{d}, \hat{\mathbf{d}}\right) = \min\left(d^+, \hat{d}^+\right) + \min\left(d^-, \hat{d}^-\right)$$
$$U\left(\mathbf{d}, \hat{\mathbf{d}}\right) = \max\left(d^+, \hat{d}^+\right) + \max\left(d^-, \hat{d}^-\right)$$



Inference

Inference

- Given an audio frame $\mathbf{x}_m$ at time index m, the confidence score that a target event class c occurs at n:

$$f_c(n \,|\, \mathbf{x}_m) = \begin{cases} P_1\,(1 \,|\, \mathbf{x}_m)\, P_2\,(c \,|\, \mathbf{x}_m) & \text{if} \quad n \in \text{ROI}, \\ 0 & \text{otherwise} \end{cases}$$

  - $P_1\,(1 \,|\, \mathbf{x}_m)$: Posterior prob. for $\mathbf{x}_m$ classified as foreground by DNN-1
  - $P_2\,(c \,|\, \mathbf{x}_m)$: Posterior prob. for $\mathbf{x}_m$ classified as class c by DNN-2
  - $n \in$ ROI if $m - \hat{d}_{\text{on}}(\mathbf{x}_m) \leq n \leq m + \hat{d}_{\text{off}}(\mathbf{x}_m)$

- The accumulated confidence score given all audio frames:

$$f_c(n) = \sum_m f_c(n \,|\, \mathbf{x}_m)$$

Inference

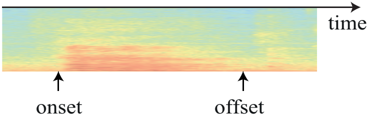
- Given an audio frame $\mathbf{x}_m$ at time index m, the confidence score that a target event class c occurs at n:

$$f_c(n \mid \mathbf{x}_m) = \begin{cases} P_1(1 \mid \mathbf{x}_m)\, P_2(c \mid \mathbf{x}_m) & \text{if} \quad n \in \text{ROI}, \\ 0 & \text{otherwise} \end{cases}$$

  - $P_1(1 \mid \mathbf{x}_m)$: Posterior prob. for $\mathbf{x}_m$ classified as foreground by DNN-1
  - $P_2(c \mid \mathbf{x}_m)$: Posterior prob. for $\mathbf{x}_m$ classified as class c by DNN-2
  - $n \in$ ROI if $m - \hat{d}_{\text{on}}(\mathbf{x}_m) \leq n \leq m + \hat{d}_{\text{off}}(\mathbf{x}_m)$

- The accumulated confidence score given all audio frames:

$$f_c(n) = \sum_m f_c(n \mid \mathbf{x}_m)$$



Inference

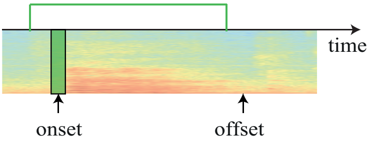
- Given an audio frame $\mathbf{x}_m$ at time index m, the confidence score that a target event class c occurs at n:

$$f_c(n \,|\, \mathbf{x}_m) = \begin{cases} P_1\left(1 \,|\, \mathbf{x}_m\right) P_2\left(c \,|\, \mathbf{x}_m\right) & \text{if} \quad n \in \text{ROI}, \\ 0 & \text{otherwise} \end{cases}$$

  - $P_1\left(1 \,|\, \mathbf{x}_m\right)$: Posterior prob. for $\mathbf{x}_m$ classified as foreground by DNN-1
  - $P_2\left(c \,|\, \mathbf{x}_m\right)$: Posterior prob. for $\mathbf{x}_m$ classified as class c by DNN-2
  - $n \in$ ROI if $m - \hat{d}_{\text{on}}(\mathbf{x}_m) \leq n \leq m + \hat{d}_{\text{off}}(\mathbf{x}_m)$

- The accumulated confidence score given all audio frames:

$$f_c(n) = \sum_m f_c(n \,|\, \mathbf{x}_m)$$



Inference

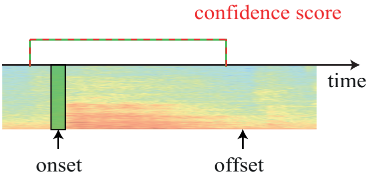
- Given an audio frame $\mathbf{x}_m$ at time index m, the confidence score that a target event class c occurs at n:

$$f_c(n \,|\, \mathbf{x}_m) = \begin{cases} P_1\,(1 \,|\, \mathbf{x}_m)\, P_2\,(c \,|\, \mathbf{x}_m) & \text{if} \quad n \in \text{ROI}, \\ 0 & \text{otherwise} \end{cases}$$

  - $P_1\,(1 \,|\, \mathbf{x}_m)$: Posterior prob. for $\mathbf{x}_m$ classified as foreground by DNN-1
  - $P_2\,(c \,|\, \mathbf{x}_m)$: Posterior prob. for $\mathbf{x}_m$ classified as class c by DNN-2
  - $n \in$ ROI if $m - \hat{d}_{\text{on}}(\mathbf{x}_m) \leq n \leq m + \hat{d}_{\text{off}}(\mathbf{x}_m)$

- The accumulated confidence score given all audio frames:

$$f_c(n) = \sum_m f_c(n \,|\, \mathbf{x}_m)$$



Inference

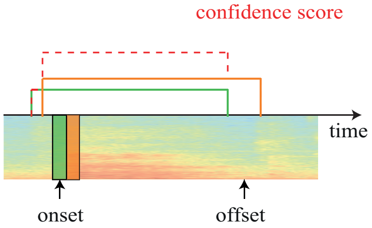
- Given an audio frame $\mathbf{x}_m$ at time index m, the confidence score that a target event class c occurs at n:

$$f_c(n \mid \mathbf{x}_m) = \begin{cases} P_1(1 \mid \mathbf{x}_m)\, P_2(c \mid \mathbf{x}_m) & \text{if} \quad n \in \text{ROI}, \\ 0 & \text{otherwise} \end{cases}$$

  - $P_1(1 \mid \mathbf{x}_m)$: Posterior prob. for $\mathbf{x}_m$ classified as foreground by DNN-1
  - $P_2(c \mid \mathbf{x}_m)$: Posterior prob. for $\mathbf{x}_m$ classified as class c by DNN-2
  - $n \in$ ROI if $m - \hat{d}_{\text{on}}(\mathbf{x}_m) \leq n \leq m + \hat{d}_{\text{off}}(\mathbf{x}_m)$

- The accumulated confidence score given all audio frames:

$$f_c(n) = \sum_m f_c(n \mid \mathbf{x}_m)$$



Inference

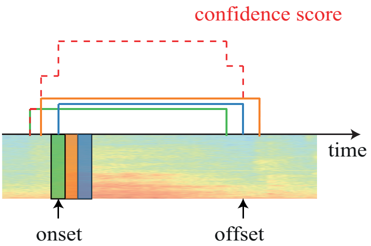
- Given an audio frame $\mathbf{x}_m$ at time index m, the confidence score that a target event class c occurs at n:

$$f_c(n \,|\, \mathbf{x}_m) = \begin{cases} P_1\left(1 \,|\, \mathbf{x}_m\right) P_2\left(c \,|\, \mathbf{x}_m\right) & \text{if} \quad n \in \text{ROI}, \\ 0 & \text{otherwise} \end{cases}$$

  - $P_1\left(1 \,|\, \mathbf{x}_m\right)$: Posterior prob. for $\mathbf{x}_m$ classified as foreground by DNN-1
  - $P_2\left(c \,|\, \mathbf{x}_m\right)$: Posterior prob. for $\mathbf{x}_m$ classified as class c by DNN-2
  - $n \in$ ROI if $m - \hat{d}_{\text{on}}(\mathbf{x}_m) \leq n \leq m + \hat{d}_{\text{off}}(\mathbf{x}_m)$

- The accumulated confidence score given all audio frames:

$$f_c(n) = \sum_m f_c(n \,|\, \mathbf{x}_m)$$



Detection Function's Monotonicity

- Assume the accumulated confidence score at index $n > 0$ given all frames up to index $\bar{m} > 0$:

$$f_{\bar{m}}(n) = \sum_{m=1}^{\bar{m}} f(n \,|\, \mathbf{x}_m)$$

- The updated confidence score when the new frame $\bar{m}+1$ is observed:

$$f_{\bar{m}+1}(n) = \sum_{m=1}^{\bar{m}+1} f(n \,|\, \mathbf{x}_m) = \sum_{m=1}^{\bar{m}} f(n \,|\, \mathbf{x}_m) + f(n \,|\, \mathbf{x}_{\bar{m}+1})$$

$$\geq \sum_{m=1}^{\bar{m}} f(n \,|\, \mathbf{x}_m) = f_{\bar{m}}(n)$$

due to $f(n \,|\, \mathbf{x}_m) \geq 0$, $\forall m > 0$

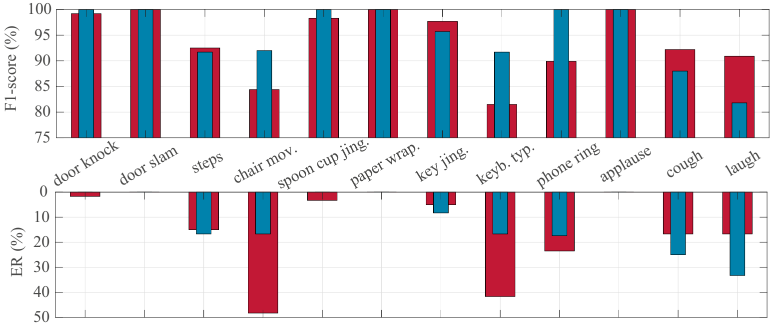
Experiments

Experimental Setup

- ITC-Irst database (CHIL/CLEAR 2006)
  - 1.7 hours in total
  - Single microphone out of 32 microphones used
  - Evaluating on 12 event categories (e.g. *door knock*, *coughing*)
- Feature extraction
  - 100 ms frame length and 90 ms overlap
  - 64 log Gammatone spectral coefficients in freq. range [50, 22050] Hz
- Baseline systems
  - **SVM**: Detection-by-classification with RBF-kernel SVMs
  - **GMM-HMM**: Joint detection and segmentation with GMM-HMM
  - **Reg. Forests:** Onset and offset detection with Regression Forests
- Evaluation metrics: Detection error rate (ER) and F1-score

Detection Performance (Offline)

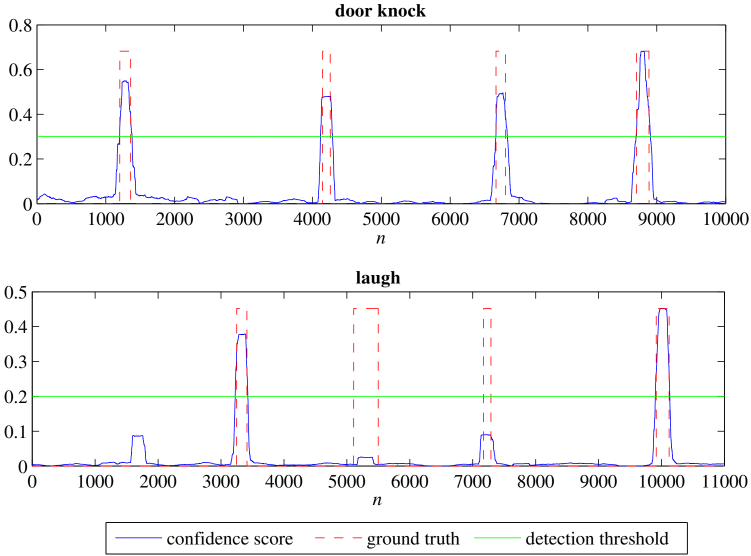
	SVM	GMM-HMM	Reg. Forests	Dual-DNN
ER (%)	30.8	39.0	15.1	**11.0** ($\downarrow$ 4.1)
F1-score (%)	83.7	84.4	93.1	**95.2** ($\uparrow$ 2.1)



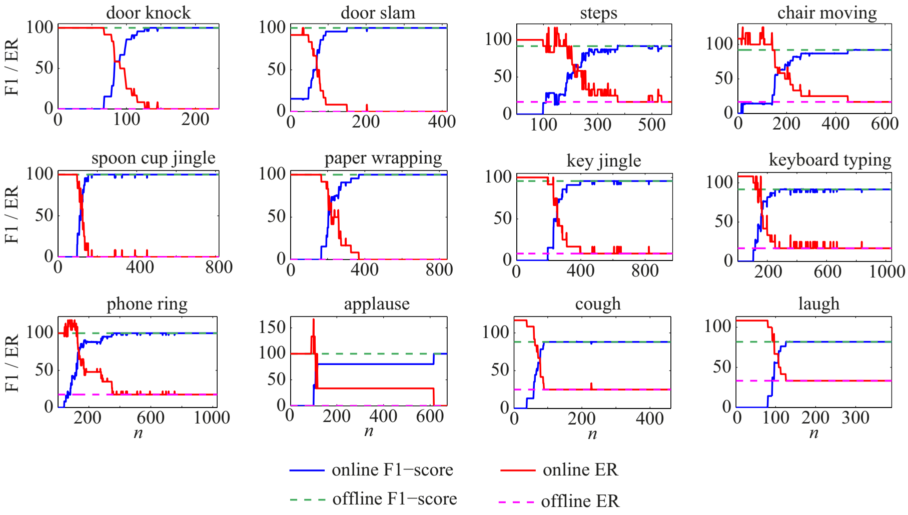
H. Phan, M. Maaß, R. Mazur, and A. Mertins,"Random Regression Forests for Acoustic Event Detection and Classification," *TASLP*, vol. 23, no. 1, pp. 20-31, 2015.

H. Phan, P. Koch, F. Katzberg, M. Maaß, R. Mazur, I. McLoughlin, and A. Mertins, "What Makes Audio Event Detection Harder than Classification," in *Proc. EUSIPCO*, pp. 2739-2743, 2017.

Good and Bad Cases



Early Detection Performance (Online)



Summary

Summary

- Addressing early audio event detection in audio streams
- Dual-DNN detection system with tailored loss functions
- Inference to reliably detect and anticipate ongoing events
- Good performance on the studied dataset
- Early event detection capability demonstrated

Thank you for your attention

Early Detection Performance with Reg. Forests