AAC Encoding Detection and Bitrate Estimation using a Convolutional Neural Network

Daniel Seichter #*, Luca Cuccovillo #, Patrick Aichroth #

Fraunhofer Institute for Digital Media Technology, * Ilmenau University of Technology

Overview

AAC encoding detection and bitrate estimation

- Blind analysis of PCM material
- Based on a Convolutional Neural Network (CNN)
- Accuracy of 94.56% by analysis of only 116.10 ms of content

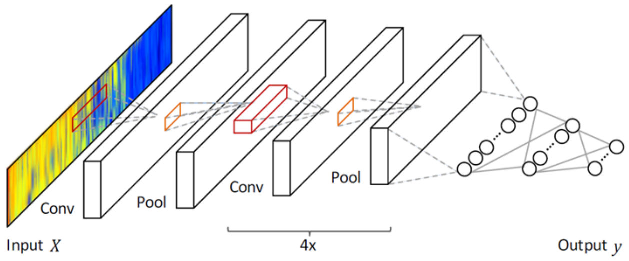


Figure 1 – CNN for AAC encoding detection

Experimental Setup

Content preparation

- Training, validation and test set are *completely disjoint*
- *Full range* of available bitrates was covered
- 50 files with varying content, unrelated to each other
- Elementary *test examples* consist of 4 overlapping AAC frames

Target Set	Amount per class (#)			
	Files	Segments	Frames	Examples
Training	20	920	77280	19320
Validation	10	460	38620	9660
Test	20	920	77280	19320

Figure 4 – Content setup for CNN training, validation and testing

Robust algorithm for AAC detection

Which input features?

- MDCT coefficients hold important encoding traces
- Must be extracted using the *correct offset* and *window shape*
- Both the *evolution in time* and in the *frequency domain* are relevant

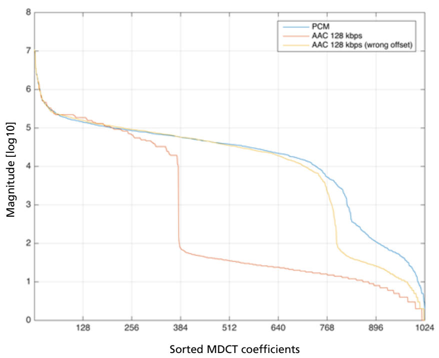
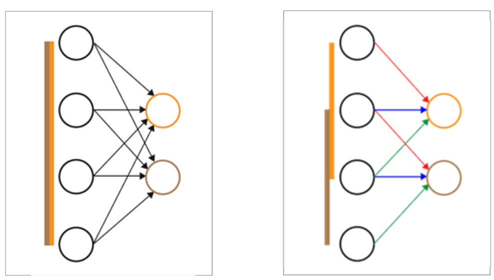


Figure 2 – Input features for AAC detection

Which classifier?

- Deep Networks can handle high input variability
- Custom features too sensitive to the specific testing setup
- Local connectivity of CNNs is able to correctly handle and describe both time and frequency domain



Feed-Forward Neural Network — Convolutional Neural Network

Figure 3 – Local connectivity of CNNs

Result Analysis

Direct application of the CNN

- Uses 4 AAC overlapping frames to create an example 16.10 ms long
- Output class directly related to the highest output of the CNN
- Average accuracy of 94.65%

	PCM	32	48	64	96	128	192	256	320
PCM	94.7	0.1	0.1	0.9	2.4	0.7	0.4	0.5	0.2
32	0.0	96.9	3.0	0.1	0.0	0.0	0.0	0.0	0.0
48	0.0	5.9	91.0	3.1	0.0	0.0	0.0	0.0	0.0
64	0.0	0.1	1.2	97.7	1.0	0.0	0.0	0.0	0.0
96	0.0	0.0	0.0	0.6	98.8	0.5	0.1	0.0	0.0
128	0.1	0.0	0.0	0.1	3.4	95.7	0.6	0.1	0.0
192	0.0	0.0	0.0	0.0	0.6	0.8	94.5	3.9	0.2
256	0.2	0.1	0.0	0.1	0.3	0.1	8.3	90.6	0.3
320	0.7	0.1	0.0	0.2	0.3	0.1	1.2	5.4	92.0

Figure 5 – Confusion matrix with 116.10 ms of content

Score-based fusion of the CNN output

- Uses 21 network examples to create a segment of 2 s duration
- Output class related to the highest output of the CNN after fusion
- Average accuracy of 97.9%

	PCM	32	48	64	96	128	192	256	320
PCM	96.9	0.0	0.0	0.2	2.6	0.3	0.0	0.0	0.0
32	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
48	0.0	1.5	98.1	0.4	0.0	0.0	0.0	0.0	0.0
64	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
96	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
128	0.0	0.0	0.0	0.0	2.2	97.8	0.0	0.0	0.0
192	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
256	0.0	0.0	0.0	0.0	0.0	0.0	6.1	93.9	0.0
320	0.1	0.0	0.0	0.0	0.1	0.0	0.8	4.5	94.5

Figure 6 – Confusion matrix with 2 s of content

QR-code to the project website:
http://s.fhg.de/idmt-audioforensics

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT