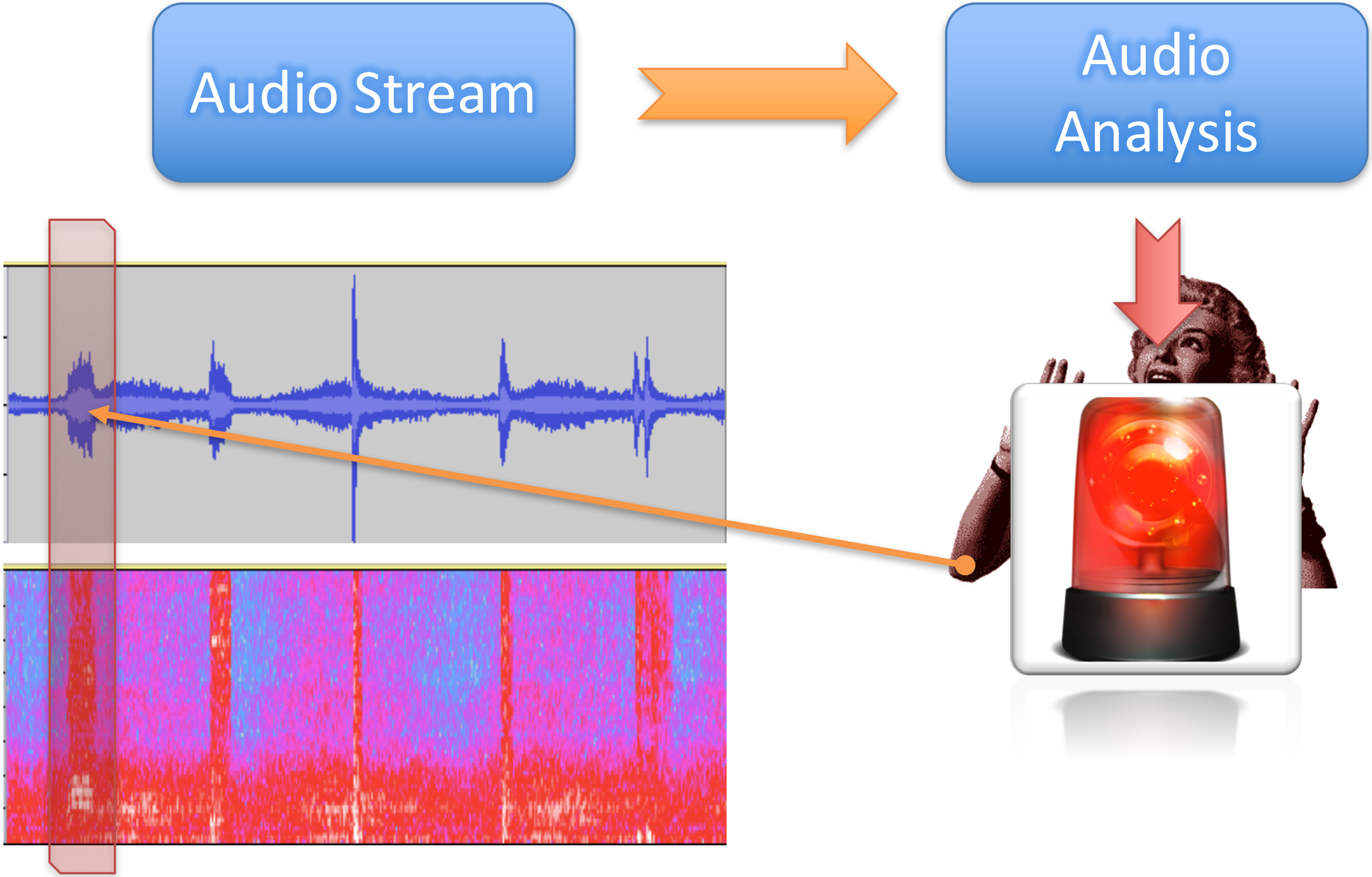
Audio event detection for audio surveillance: bag of words approach

Pattern Recognition 2016/2017

Dr. Nicola Strisciuglio

P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, M. Vento, *"Reliable detection of audio events in highly noisy environments,"* Pattern Recognition Letters, 2015
P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, M. Vento, *"Audio Surveillance of Roads: A System for Detecting Anomalous Sounds,"* in Intelligent Transportation Systems, IEEE Transactions on, 2015

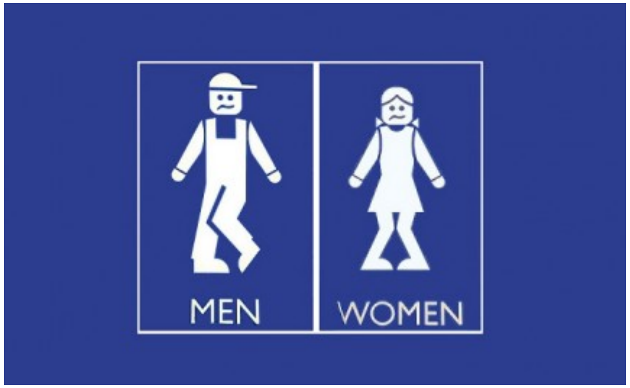
Audio surveillance (event detection)



Audio Stream

Audio Analysis

Use cases



Data sets

- Experiments of two data sets

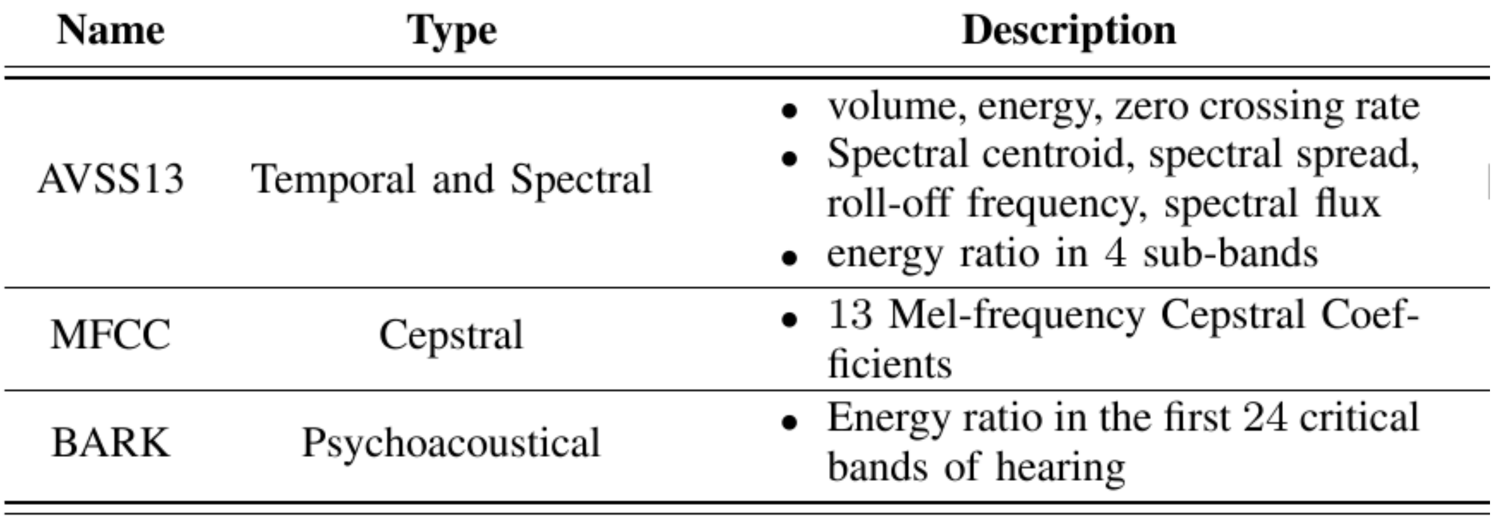
1. * MIVIA audio events
   - glass breakings, gun shots, screams
   - 6000 events per class (8000 in version 2)
   - 6 levels of SNR (8 levels in version 2, including 0*dB* and -5*dB*)
2. ** MIVIA road events
   - 400 events for roads monitoring

- Available for research purpose at http://mivia.unisa.it

Hypothesis

- The sound is composed of atomic, small audio units (like a text is composed of words)
- The occurrence of specific audio units is distinctive for a particular class of sounds
- Bag of audio words representation is suitable

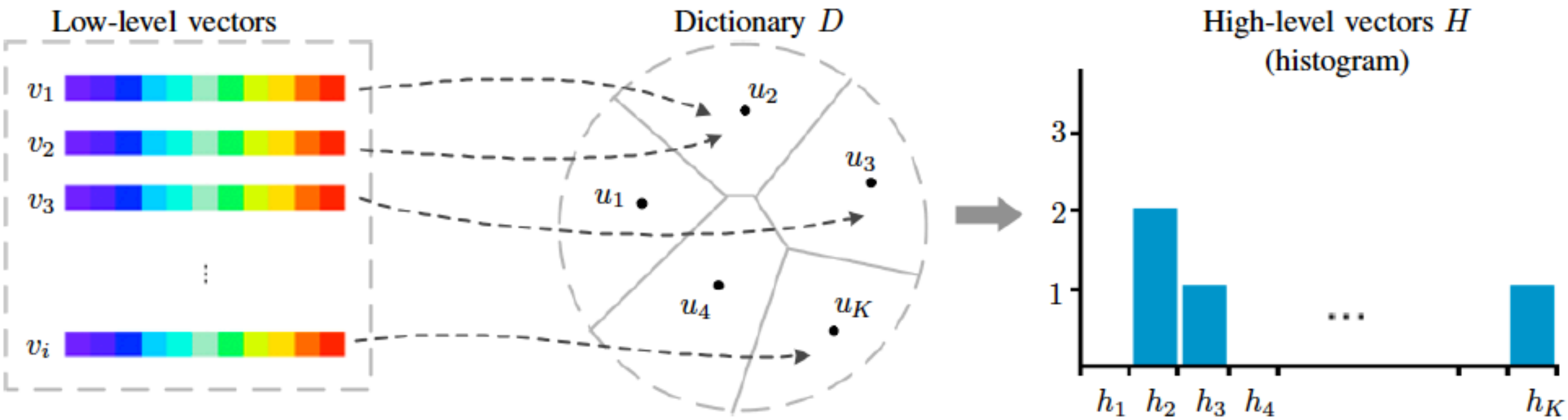
Short-time analysis



Name	Type	Description
AVSS13	Temporal and Spectral	• volume, energy, zero crossing rate • Spectral centroid, spectral spread, roll-off frequency, spectral flux • energy ratio in 4 sub-bands
MFCC	Cepstral	• 13 Mel-frequency Cepstral Coefficients
BARK	Psychoacoustical	• Energy ratio in the first 24 critical bands of hearing

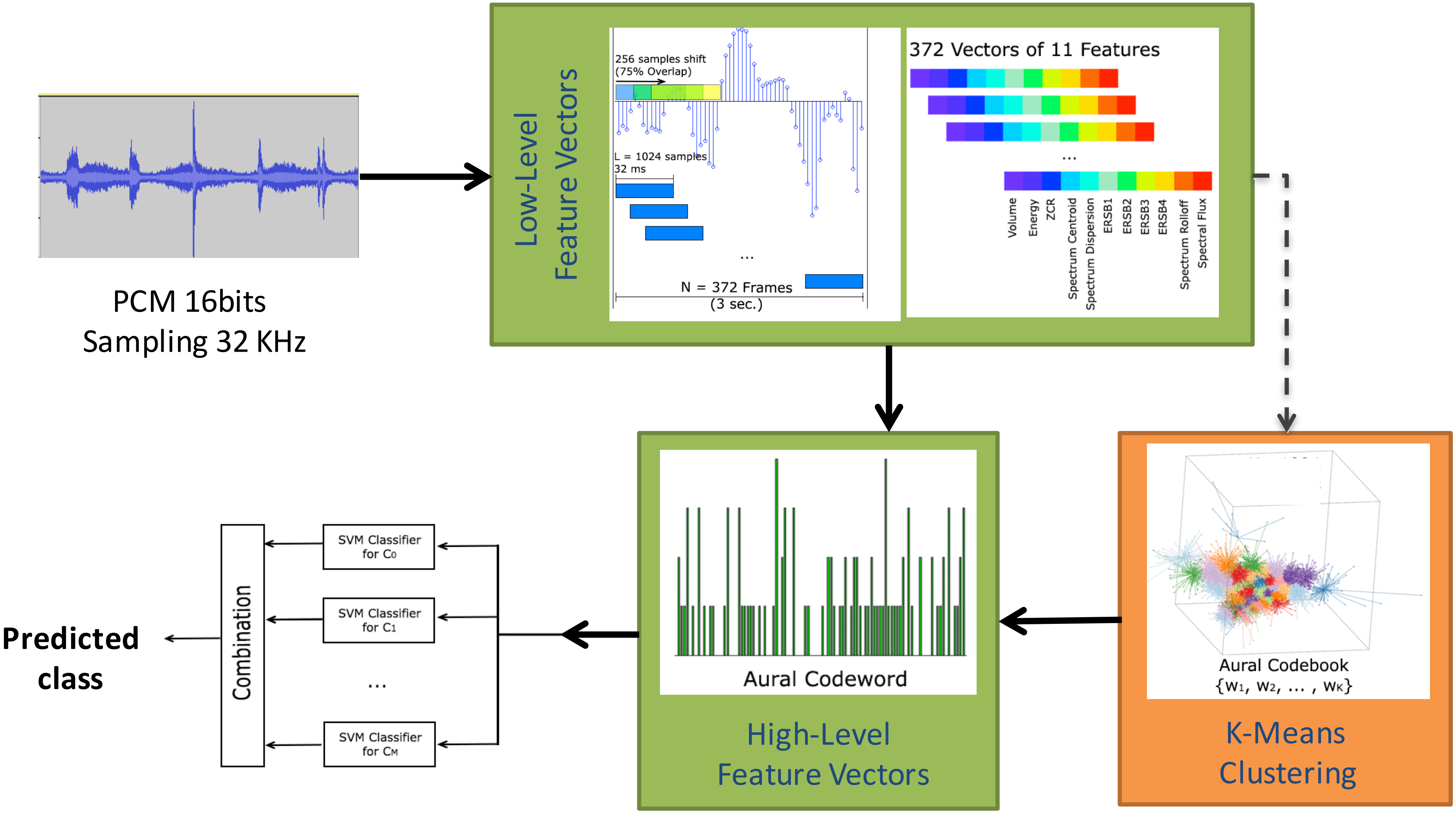
- Audio signals can vary within few milliseconds
- Capture short-time properties of the audio signal
- Overlap allows continuity of analysis

Long-time analysis



- ✓ Quantization of the vector space (Training phase).
- ✓ Histogram of the occurrences of the audio words.
- ✓ The presence of certain audio words is discriminant for specific events of interest.

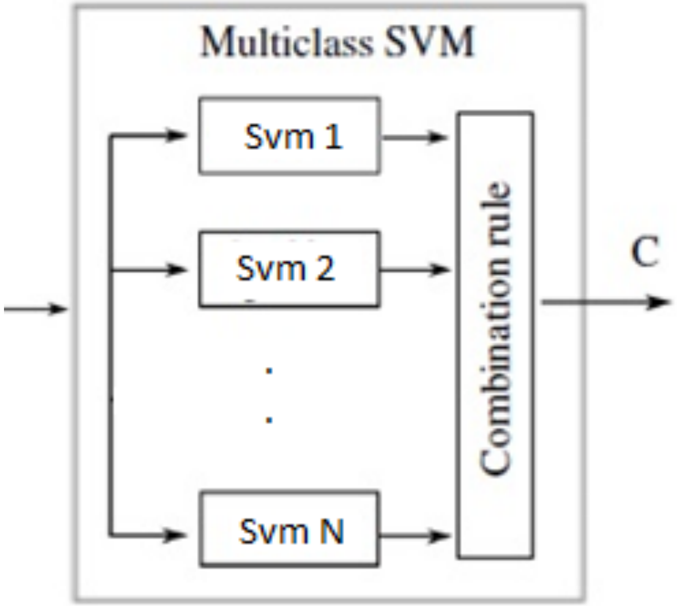
Bag of audio words



Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, Mario Vento, *"Reliable detection of audio events in highly noisy environments,"* Pattern Recognition Letters, 2015

Classification

- A pool of N one-vs-all SVM classifiers
- Each SVM is able to learn which high-level features are discriminant for the classes of interest.
- Final decision:



$$C = \begin{cases} C_0 & if\ S_i < \tau,\ \forall\, i = 0, \dots, N \\ argmax\ S_i & else \end{cases}$$

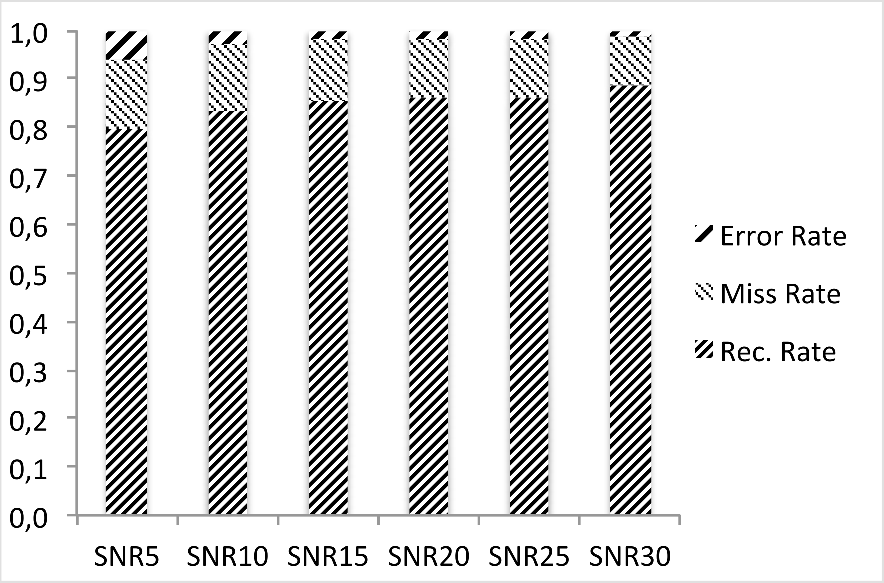
Experimental Evaluation

- Sliding window evaluation
  - An event of interest is correctly detect if it is detected in at least one of the time windows that overlap with it
- Evaluation metrics
  - Recognition Rate
  - False Positive Rate
  - Miss Rate
  - Error Rate

Results (1)

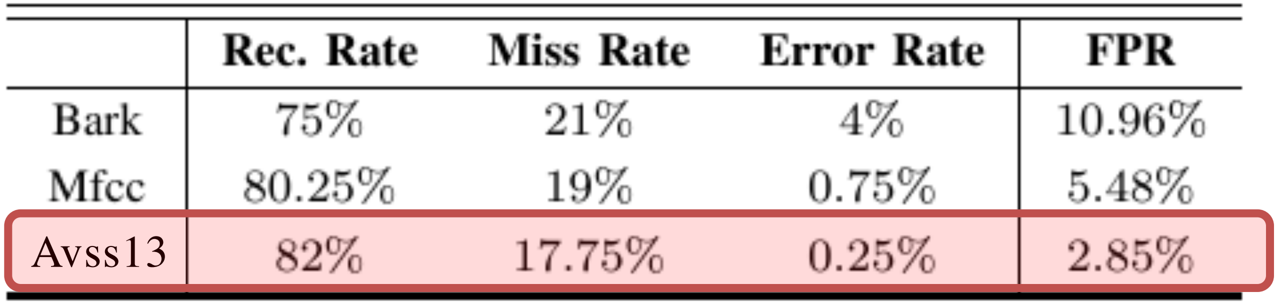
- Definition of a procedure to simulate different environments combining background sounds
- Target sound events at different SNR (MIVIA audio events)

- Recognition Rate: 84.8%
- False Positive Rate: 2.1%
- Error Rate: 2.7%
- Miss Rate: 12.5 %



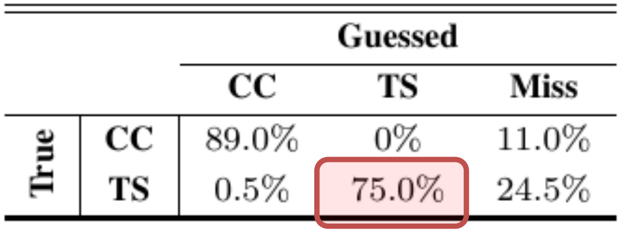
Results (2)

- Overall results (K = 64 clusters)



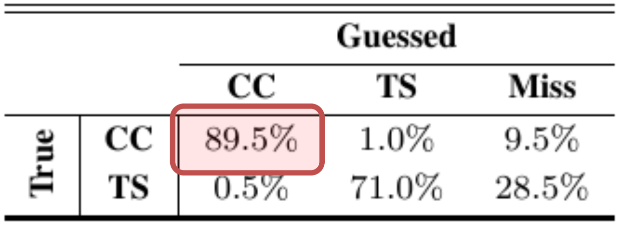
	Rec. Rate	Miss Rate	Error Rate	FPR
Bark	75%	21%	4%	10.96%
Mfcc	80.25%	19%	0.75%	5.48%
Avss13	82%	17.75%	0.25%	2.85%

- Classification matrices



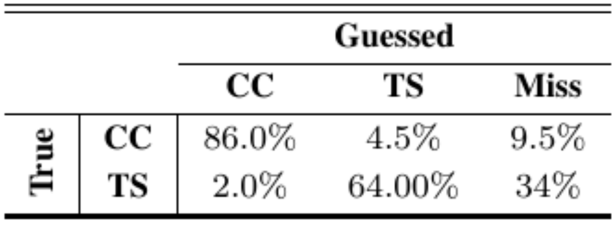
		Guessed		
		CC	TS	Miss
True	CC	89.0%	0%	11.0%
	TS	0.5%	75.0%	24.5%

AVSS13



		Guessed		
		CC	TS	Miss
True	CC	89.5%	1.0%	9.5%
	TS	0.5%	71.0%	28.5%

MFCC

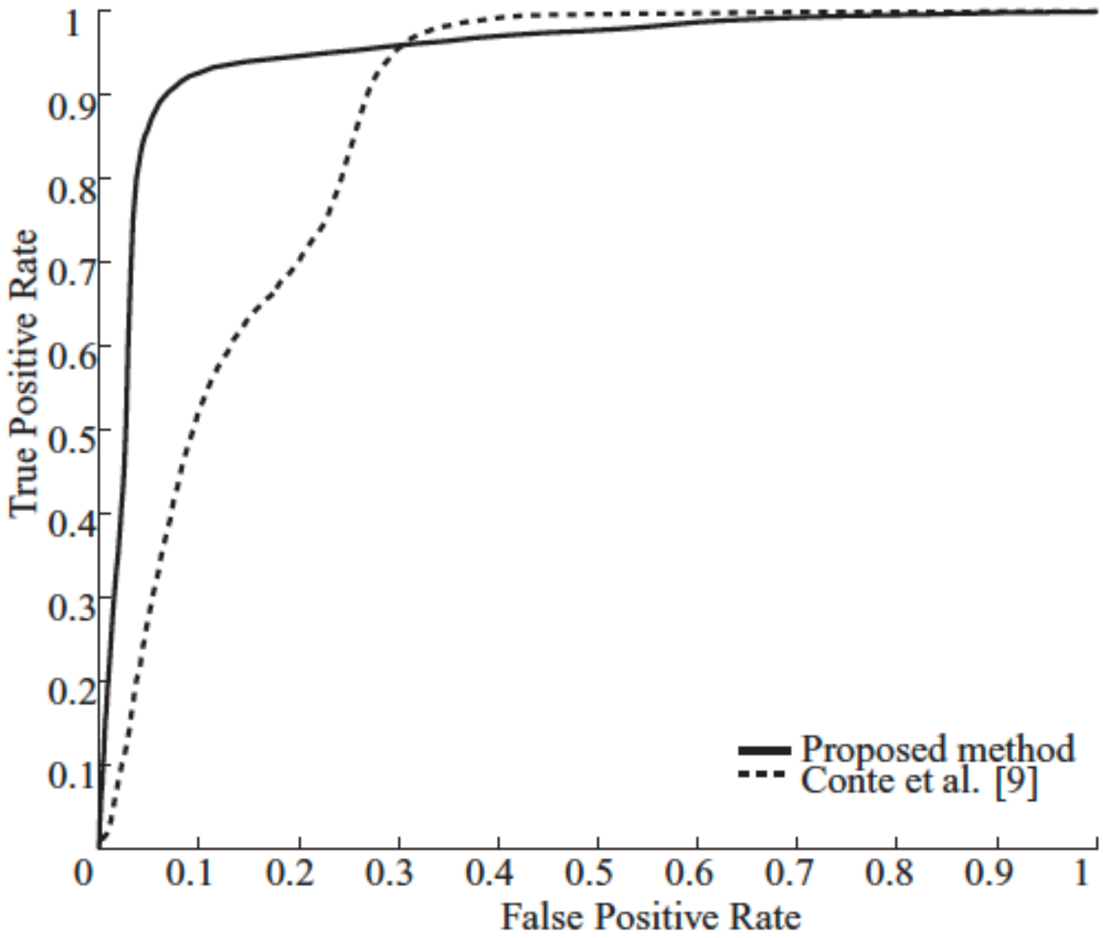


		Guessed		
		CC	TS	Miss
True	CC	86.0%	4.5%	9.5%
	TS	2.0%	64.00%	34%

Bark

Results (3)

ROC – MIVIA audio events



Strisciuglio et al. 2015