"What You Hear is What You Get"

Audio Concepts for Video Event Detection on User-Generated Content

[1]Benjamin Elizalde, [2]Mirco Ravanelli, [1]Gerald Friedland

Context

-When performing video event detection on user-generated content (UGC) different events are better described by different concepts such as music, laughter or clapping.

Problem

-Low level features do not provide humanly understandable evidence of why videos belong to a specific event.

-Ad-hoc annotations ignore the complex characteristics of UGC audio such as concept ambiguities, overlap and duration.

Our Approach

-Classify audio concepts and used them for video event detection.

Example of Events

E001	Attempting a board trick
E011	Making a sandwich
…	
E025	Marriage proposal
E029	Winning a race without a vehicle
E030	Working on a metal crafts project

Example of Audio Concepts

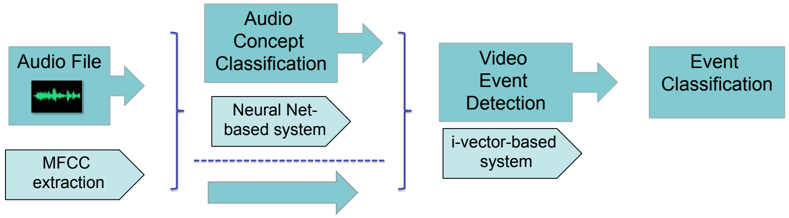
Music	Speech	Engine Light	Clinking	Radio
Crowd	Children voices	Wash-board	Drums	Singing
Rolling	Cheer	Water	Bird	Rustle
Engine heavy	Scratch	Mumble	Power Tool	Scream
Clatter	Beep	Cat	Ground Traffic	Horse

[1]International Computer Science Institute

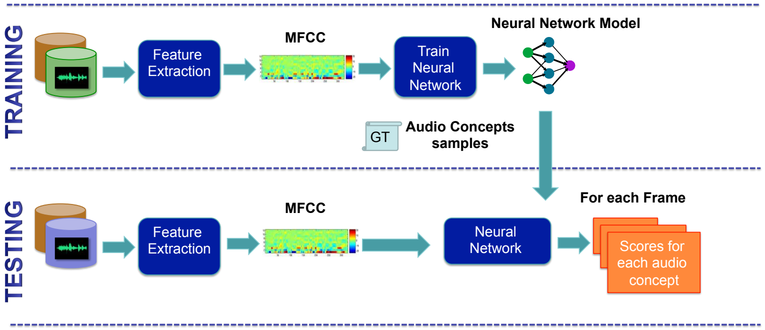
[2]Fondazione Bruno Kessler

2013

Video Event Detection Systems

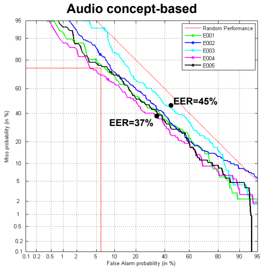


Audio Concept Classification System; towards Deep Learning

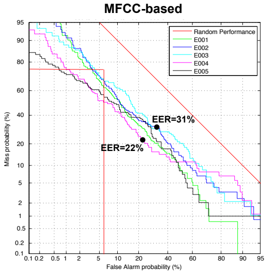


Audio Concept Results and Analysis (SRI-Sarnoff Annotation Set)

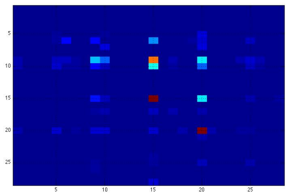
Audio concept-based



MFCC-based



Confusion Matrix: Ambiguous concepts

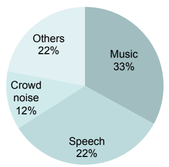


Accuracies varies in a wide range

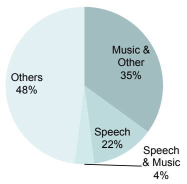
Music, Speech > 80% accuracy

Beep, Blowing <1% accuracy

Audio concepts have unbalance distributions



38% of audio concepts overlaps with one or more concepts



Average duration of the concepts is 1 second

Speech <= 1sec 38%

Crowd 380 seconds

Speech <= 2 sec >= 1sec 30%

Beep 0.5 seconds

Technicalities

TRECVID MED 2012
- 2,000 train, 12,000 test
- 15 events

Audio concepts annotation sets have unique characteristics

-SRI: 28 concepts, 11.8 hours
-CMU: 41 concepts, 13.6 hours
-Stanford: 20 concepts, 11.8 hours
-Gatech: 39 concepts, 4.3 hours

Concepts

-Two hidden layers with 1,000 neurons each
-Random initialization
-Context window of 9 frames

-MFCC, 12 coefficients + energy
-25 ms window every 10 ms

Conclusions and Ongoing Research

- Audio concept classification provides humanly understandable evidence of why videos belong to a specific event.
- Ad-hoc audio concept annotations alone does not provide reliable high-accuracy evidence nor efficient video event detection.
- So far, low level-features work better than audio-concept-features.
- Improve Audio Concept Classification with Deep Learning.

Acknowledgments

Adam Janin, from ICSI. Hui Cheng, Ajay Divakaran and Qian Yu from SRI-Sarnoff.

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20066.
The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsement, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.