TRECVID-2009 Content-based Copy Detection task Overview

Wessel Kraaij
TNO // Radboud University

George Awad, Paul Over
NIST

Outline

- Task overview
- Dataset and queries
- Transformations
- Evaluation metrics
- Participants
- Results
- Global Observations
- Issues

Task design considerations

- Copy detection is applied in several real-word tasks:
  - television advertisement monitoring
  - detection of copyright infringement
  - detection of known (illegal) content
- 2009: first year after pilot task.
- Task has both a detection and localization component.
- Detection measure based on error rates.
- Weighted trade-off of type I and type II errors. (false alarms vs. misses)
- Computation of optimal operating point by NIST.
- *Comparison of performance @ operating point submitted by participants (actual) with performance @ optimal operating point.*

CBCD task overview

- Goal:
  - Build a benchmark collection for video copy detection methods
- Task:
  - Given a set of reference (test) video collection and a set of 1407 queries,
  - determine for each query if it contains a copy, with possible transformations, of video from the reference collection,
  - and if so, from where in the reference collection the copy comes
- For 2009 three main task types were derived:
  - Copy detection of video-only queries (required)
  - Copy detection of audio-only queries (optional)
  - Copy detection of video + audio queries (required)
- *At least 2 runs (for each of the two required tasks) are required representing two application profiles ("no false alarms", "balanced").*
- Some groups submitted "video-only" runs but ignored the required "video + audio" task!!

INRIA query creation framework

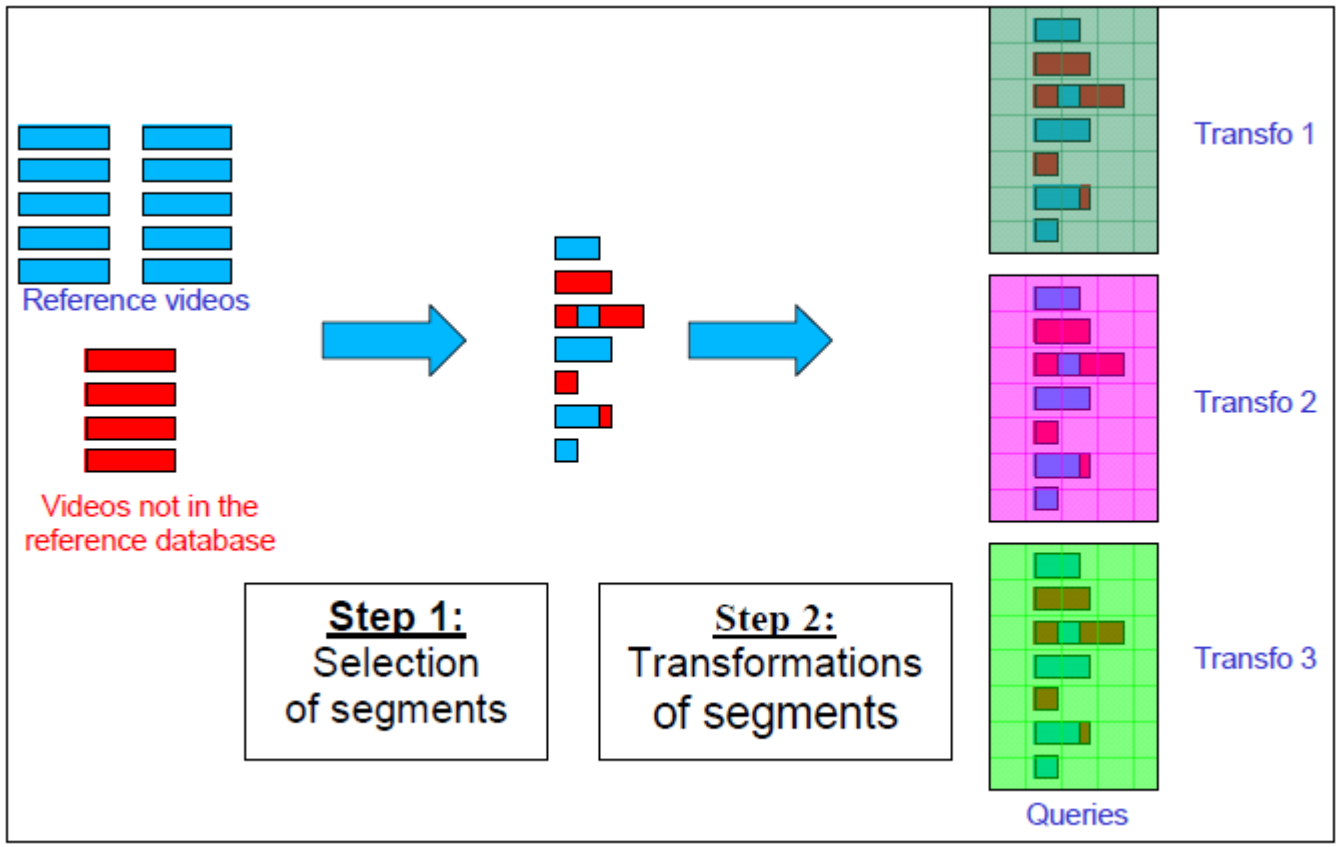


figure 1: framework for building video queries

Hard cuts, mean length ref: 32s, mean length nonref: 105s, mean query length ~ 91 s

Datasets and queries

- Dataset:
  - Reference video collection:
    - Testing data: TV2009 (180 hr) and TV2007 and TV2008 S&V data (200 hr)
    - Development data : TV2007 and TV2008 S&V data (200 hr)
  - Non-reference video collection :
    - Testing data: TV2009 BBC rushes data (30 hrs)
    - Development data: TV2007 and TV2008 BBC rushes data (53 hrs)
- Query types: (Developed by INRIA-IMEDIA software run at NIST)

Copies {
  - Type 1: composed of a reference video only. (1/3)
  - Type 2: composed of a reference video embedded in a non-reference video. (1/3)
  - Type 3: composed of a non-reference video only. (1/3)

- Number of queries:
  - 201 total original queries. 67 queries for each type.
- After creating the queries, each was transformed.
  - 7 video transformations by NIST (using a tool created by INRIA-IMEDIA)
  - 7 audio transformations by Dan Ellis at Columbia University
- Yielding… 7 * 201 = 1407 video queries , 7 * 201 = 1407 audio queries
  and 7 * 7 * 201 = 9849 audio+video queries

Video transformations

- As requested in Tv2008, some transformations were not realistic and extreme (T7 and T9). This year 3 transformations were dropped:
  - T1 (camcording) , T7 and T9.
- 7 Transformations were selected:
  - Picture in picture (T2)
  - Insertions of pattern (T3)
  - Strong re-encoding (T4)
  - Change of gamma (T5)
  - Frame dropping (T6)
  - Post production (T8) – by introducing 3 randomly selected combination of *Crop, Shift, Contrast, Text insertion, Vertical mirroring, Insertion of pattern, Picture in picture,*
  - Combination of 3 randomly selected transformations (T10) chosen from T2-T5, one transformation from *Blur, Gamma, Frame dropping, Contrast, Compression, Ratio, White noise and* T8.

Video transformations examples

Picture in Picture

Blur

Insertion of pattern

Strong re-encoding

Noise

Contrast

Change in gamma

Mirroring

Ratio

Crop

Shift

Text insertion

Audio transformations

- T1: nothing
- T2: mp3 compression
- T3: mp3 compression and multiband companding
- T4: bandwidth limit and single-band companding
- T5: mix with speech
- T6: mix with speech, then multiband compress
- T7: bandpass filter, mix with speech, compress

Some important task details/assumptions

- Detection systems submit a run threshold, which defines the system's operating point.
- Systems are asked to output a list of possible copies (each associated with a decision score).
- The run threshold is used to determine the asserted copies.
- A query can yield just one true positive
- A query can give rise to many false alarms (even within one reference video)
- Consequence:
  - Type I error modeled as *false alarm rate*
  - Type II error modeled as *Pmiss*

Evaluation metrics

- Three main metrics were adopted:
  1. Normalized Detection Cost Rate (NDCR)
     - measures error rates/probabilities on the test set:
       - Pmiss (probability of a missed copy)
       - Rfa (false alarm rate)
     - combines them using assumptions about two possible realistic scenarios:

     1 - No False Alarm profile:
       - *Copy target rate (Rtarget)* = 0.5/hr
       - *Cost of a miss (CMiss)* = 1
       - *Cost of a false alarm (CFA)* = 1000

     2 – Balanced profile:
       - *Copy target rate (Rtarget) = 0.5/hr*
       - *Cost of a miss (CMiss) = 1*
       - *Cost of a false alarm (CFA) = 1*
  2. F_1 (how accurately the copy is located, harmonic mean of P and R)
  3. Mean processing time per query

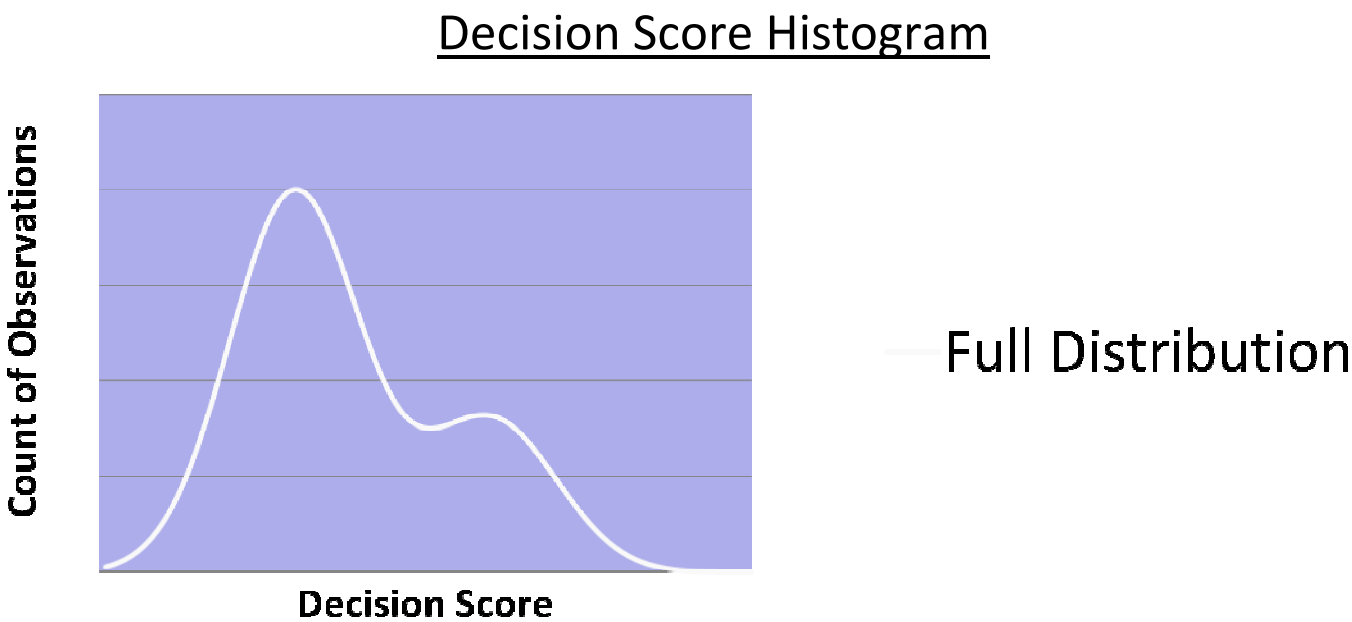
[Kraaij, Over, Fiscus, Joly,2009] Final CBCD Evaluation plan TRECVID 2009 v1.3

Evaluation metrics (2)

General rules:

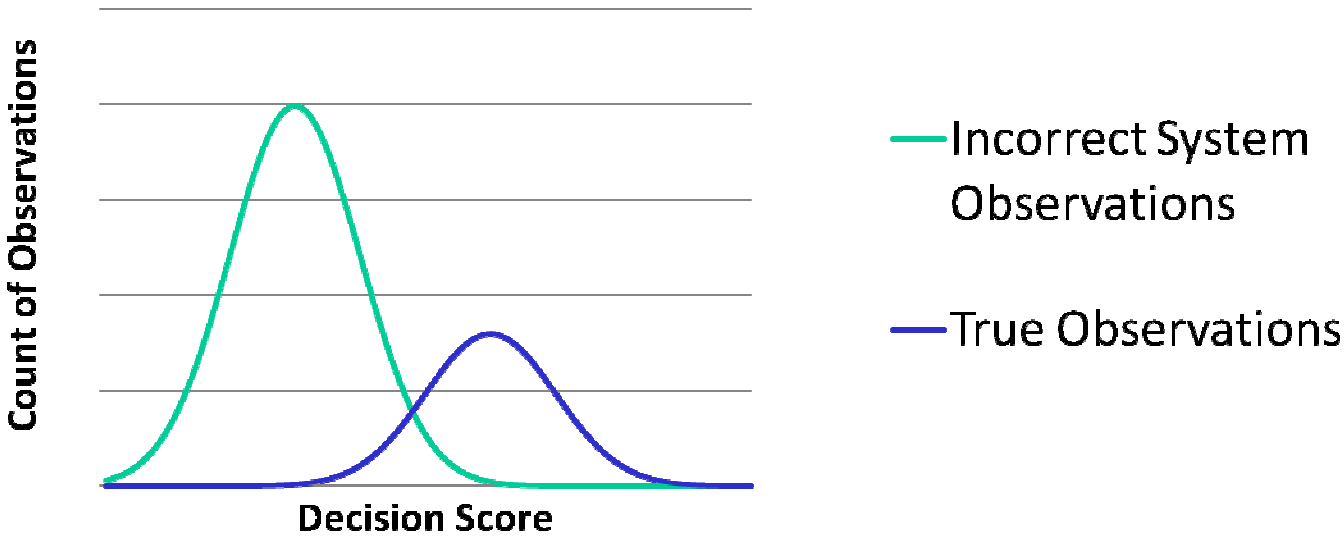
- No two query result items for a given video can overlap.
- For multiple result items per query, one mapping of submitted extents to ref extents is determined based on a combination of F1-score and the decision score (using the Hungarian solution to the Bipartite Graph matching problem).
- The reference data has been found if and only if:
  The asserted test video ID is correct AND asserted copy and ref. video overlap.

Decision Error Tradeoff Curves $Prob_{Miss}$ vs. $Rate_{FA}$



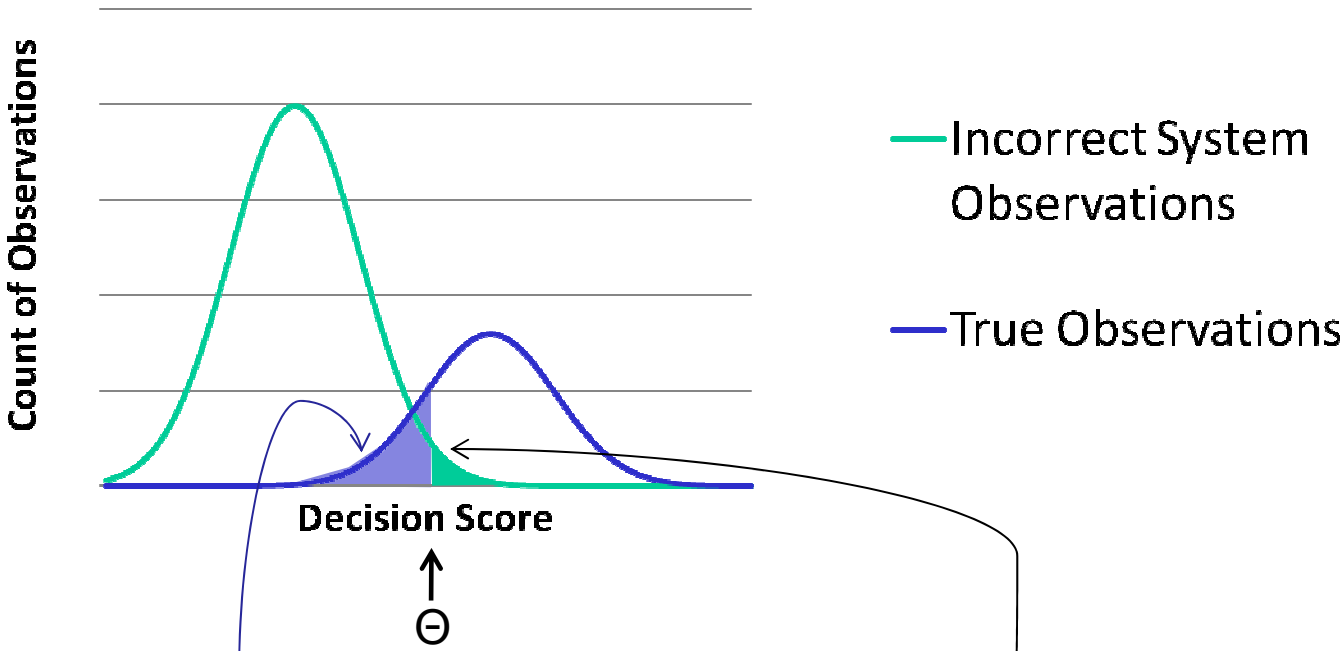
Decision Error Tradeoff Curves $Prob_{Miss}$ vs. $Rate_{FA}$

Decision Score Histogram Separated wrt. Reference Annotation s



Decision Error Tradeoff Curves $Prob_{Miss}$ vs. $Rate_{FA}$

Decision Score Histogram Separated wrt. Reference Annotation s

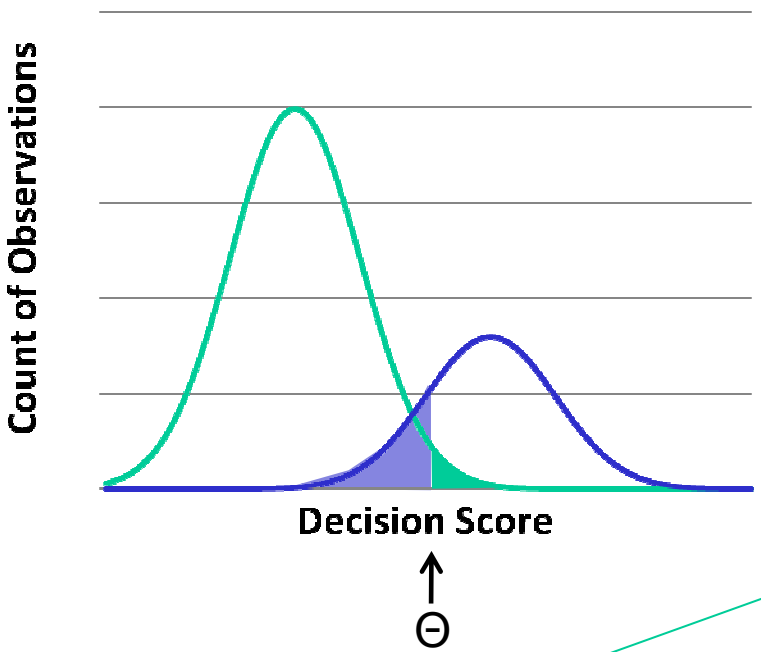


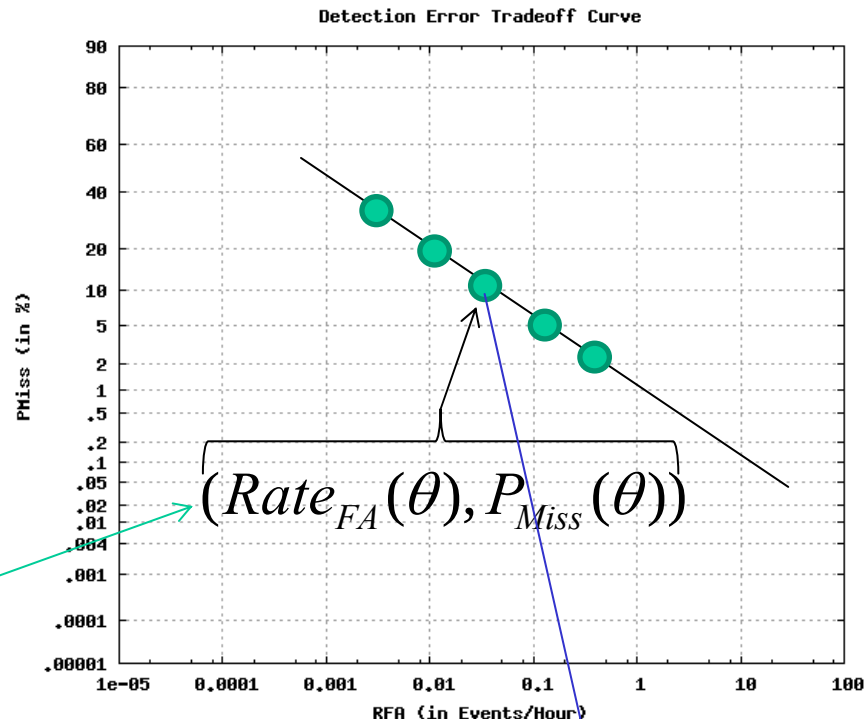
$$P_{Miss}(\theta) = \frac{\#MissedObs}{\#TrueObs} \qquad Rate_{FA}(\theta) = \frac{\#FalseAlarms}{SignalDuration}$$

signal: query

Decision Error Tradeoff Curves $Prob_{Miss}$ vs. $Rate_{FA}$

Compute $Rate_{FA}$ and P_{Miss} for all Θ



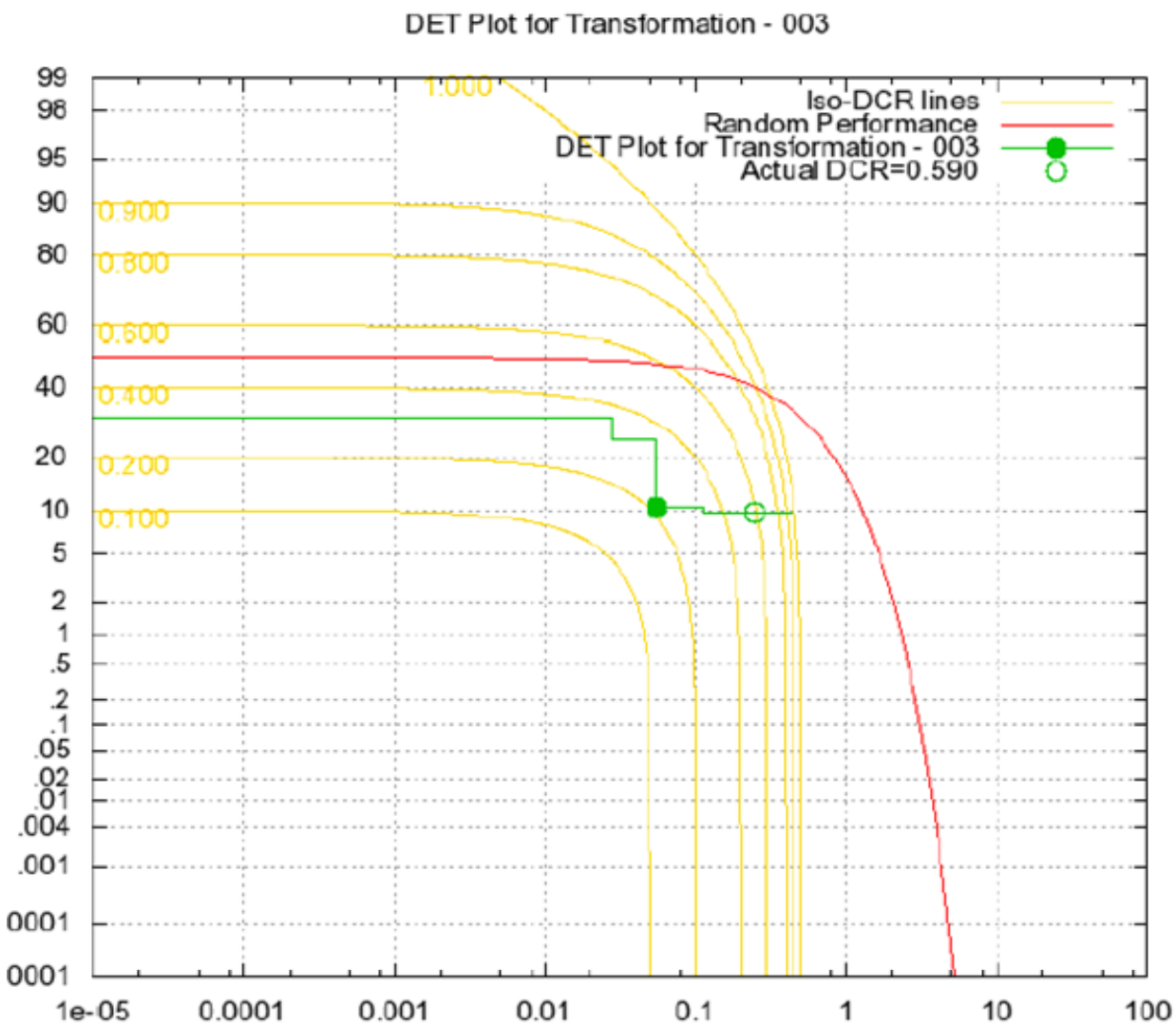


leads to: $NDCR = Pmiss + \beta\, Rfa$

The minimal β *defined by task characteristics*

Optimal threshold determined by NIST

Example det curve: optimal vs actual NDCR



20 Participants (finishers) (2008: 22)

Asahikasei Co.	--	FE	--	CD
AT&T Labs - Research	--	--	--	CD
Beijing University of Posts and Telecom.-MCPRL	ED	FE	SE	CD
Computer Research Institute of Montreal	--	--	--	CD
Fudan University	--	FE	--	CD
IBM Watson Research Center	ED	FE	SE	CD
Tsinghua University-IMG	ED	FE	SE	CD
Istanbul Technical University	--	--	--	CD
JOANNEUM RESEARCH Forschungsgesellschaft mbH-JRS	ED	**	--	CD
Chinese Academy of Sciences-MCG-ICT-CAS	--	--	SE	CD
Telefonica I+D	--	--	--	CD
Tsinghua University-MPAM	--	FE	--	CD
National Institute of Informatics	ED	FE	SE	CD
Nanjing University	--	--	--	CD
TNO	--	--	--	CD
TUBITAK UZAY	ED	FE	--	CD
University of Brescia	--	--	--	CD
City University of Hong Kong	ED	FE	SE	CD
University of Ottawa	ED	--	--	CD
Xi'an Jiaotong University	--	FE	SE	CD

-- : group didn't participate, blue: new participante

Submission types and counts

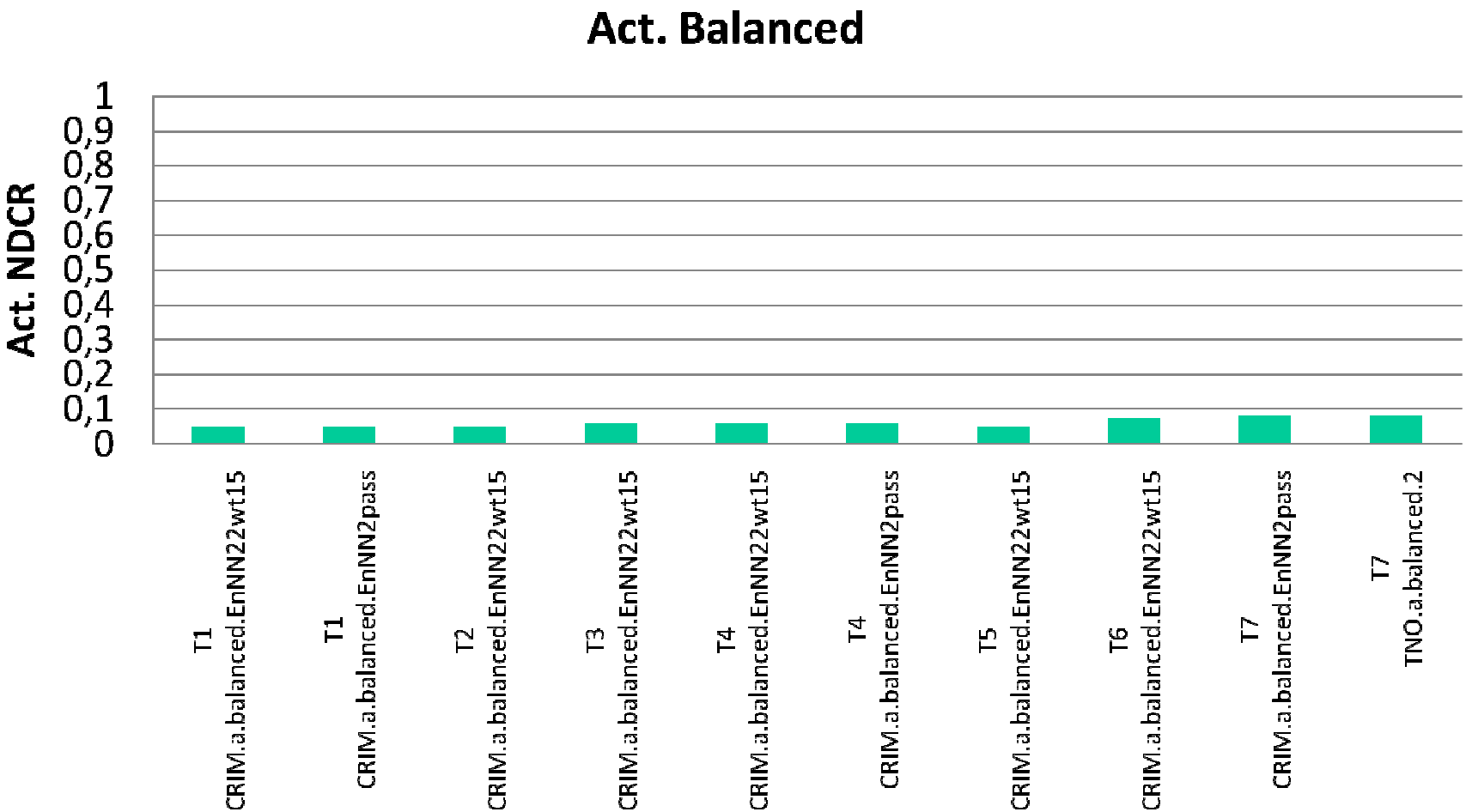
Run type	2008	2009
V (video only)	48	53
A (audio only)	1	12
M (video + audio)	6	42
Total runs	55	107

Good increase in a & m participation

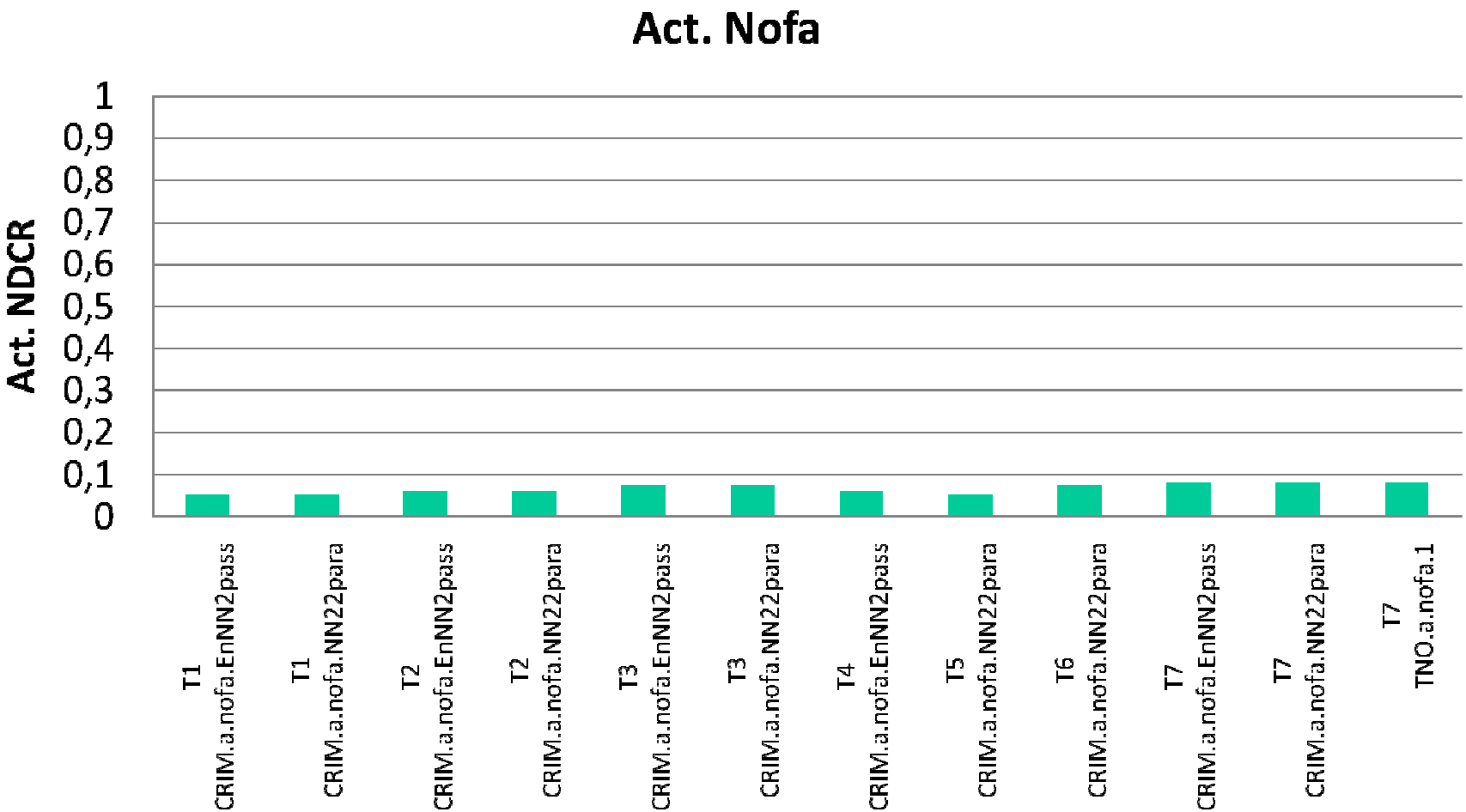
Video-only (balanced)	Video-only (nofa)	Audio-only (balanced)	Audio-only (nofa)	Video+Audio (balanced)	Video+Audio (nofa)
29	24	6	6	22	20

Balanced submissions between the two application profiles

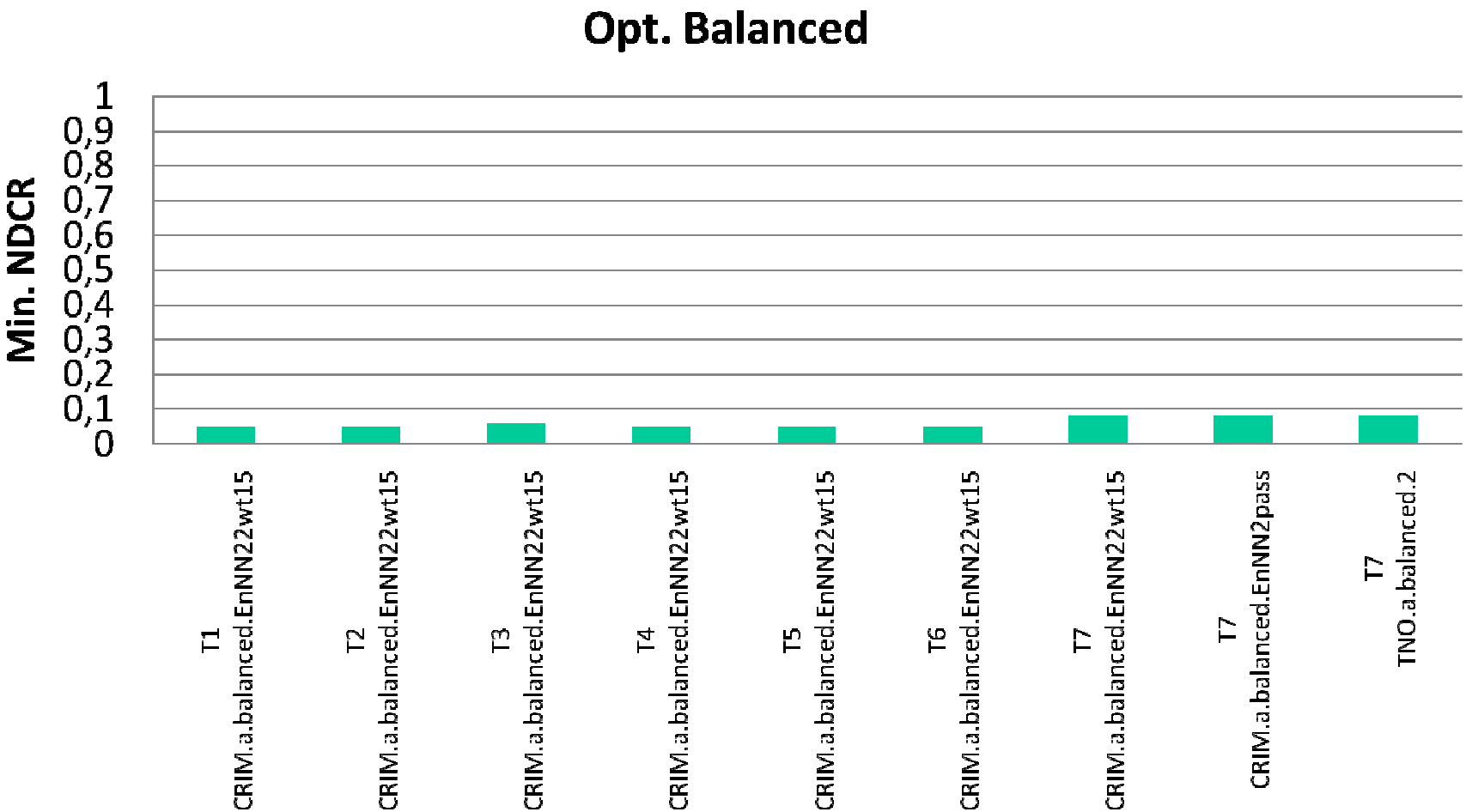
Top “audio-only” runs



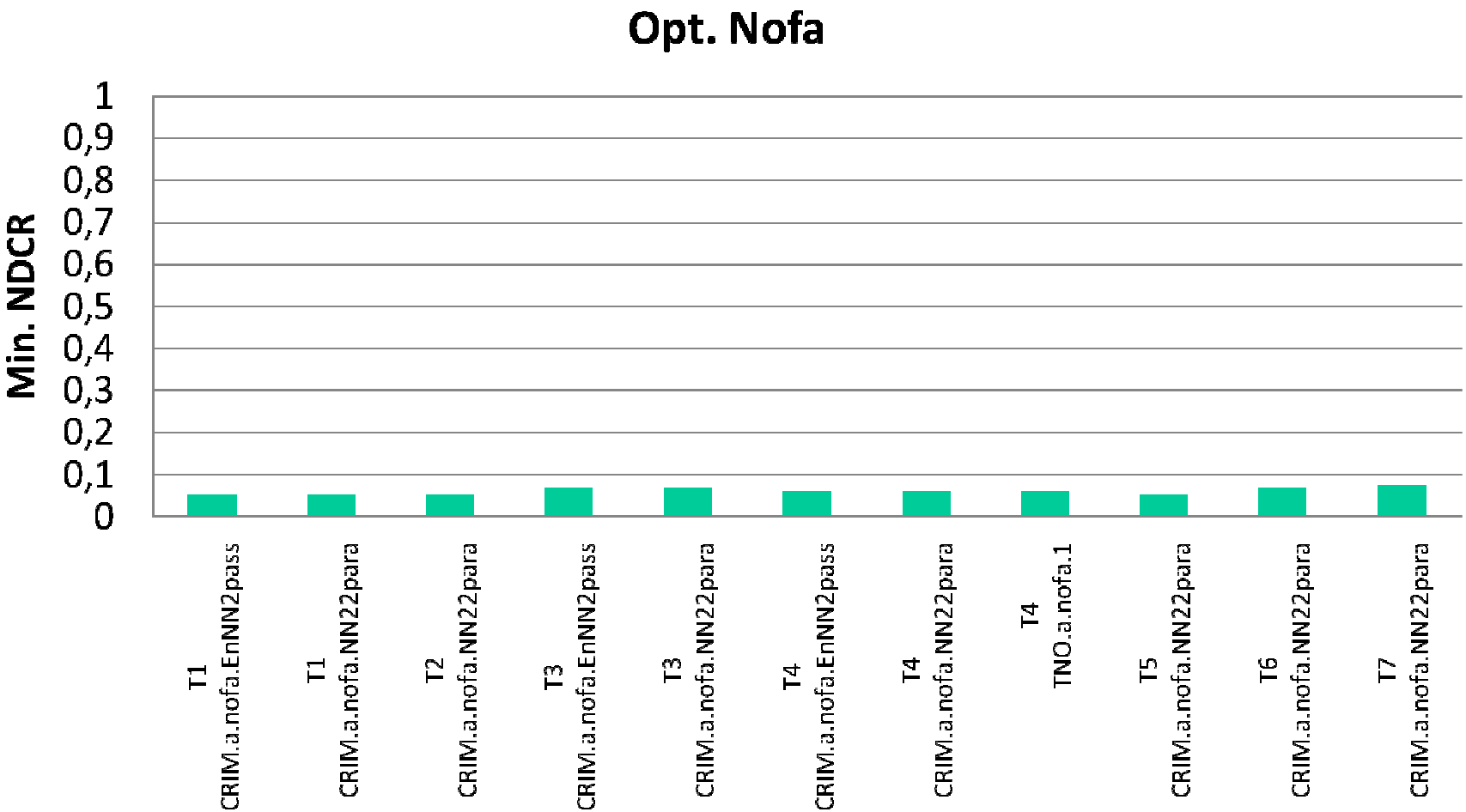
Top “audio-only” runs



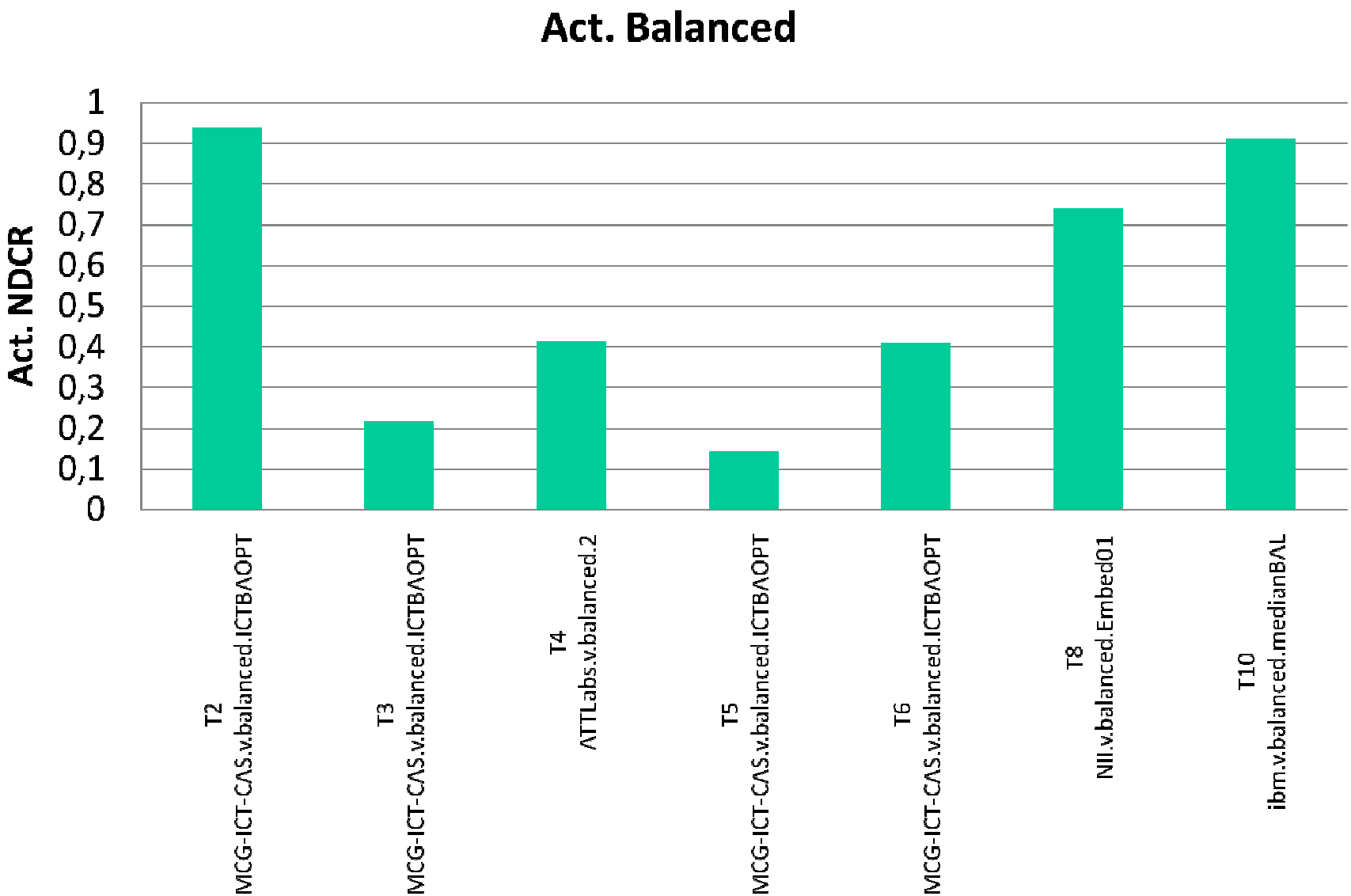
Top “audio-only” runs



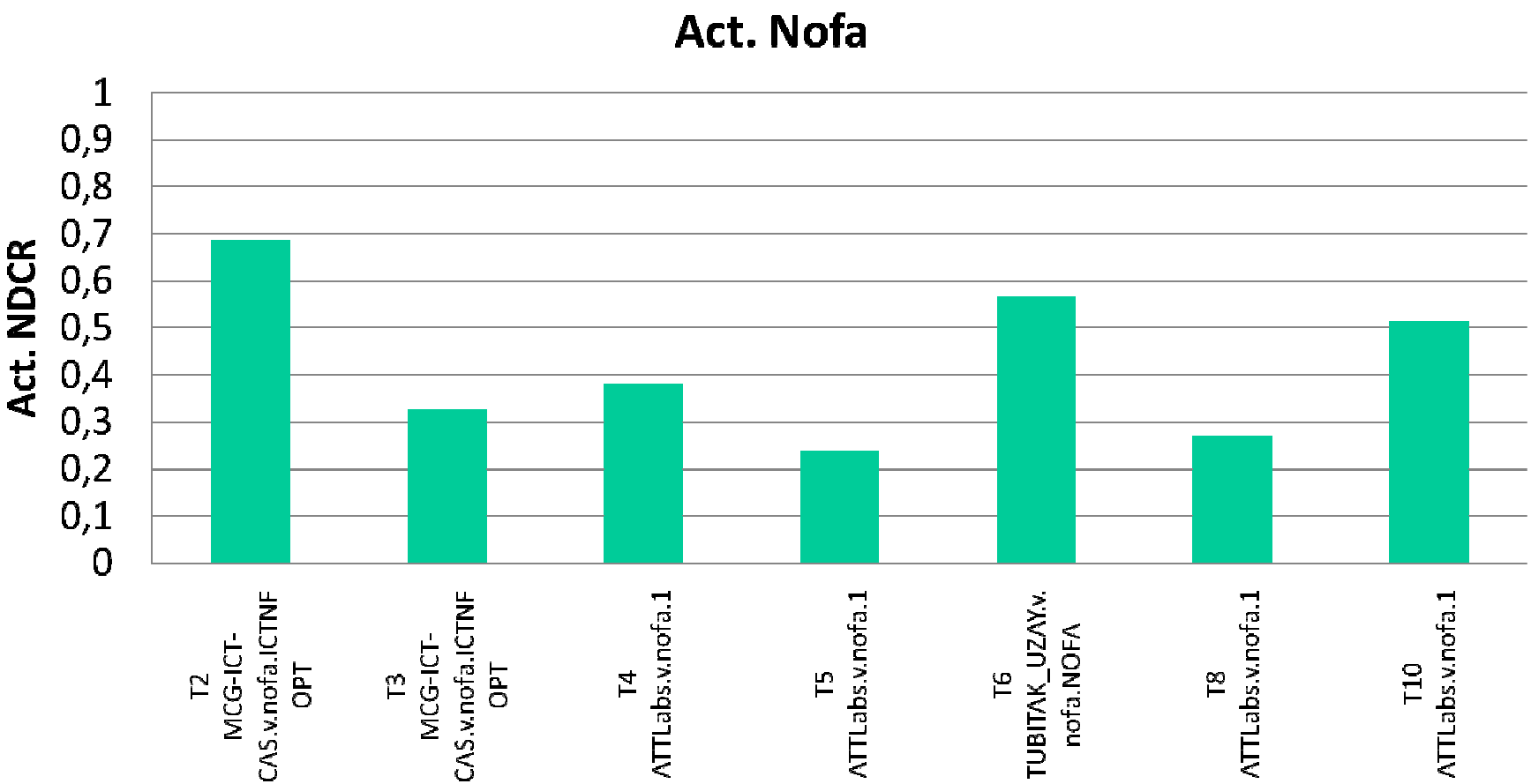
Top “audio-only” runs



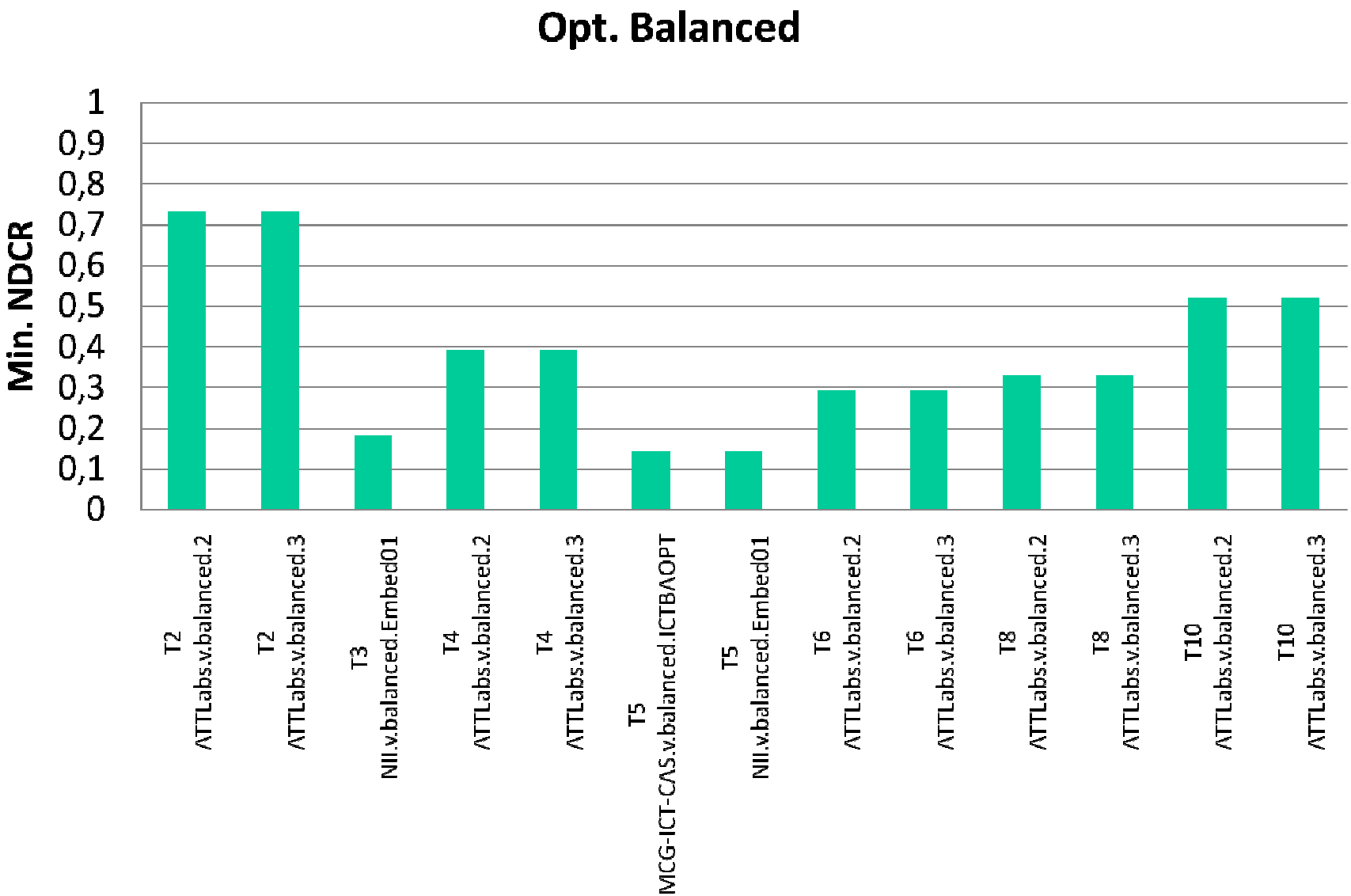
Top “video-only” runs



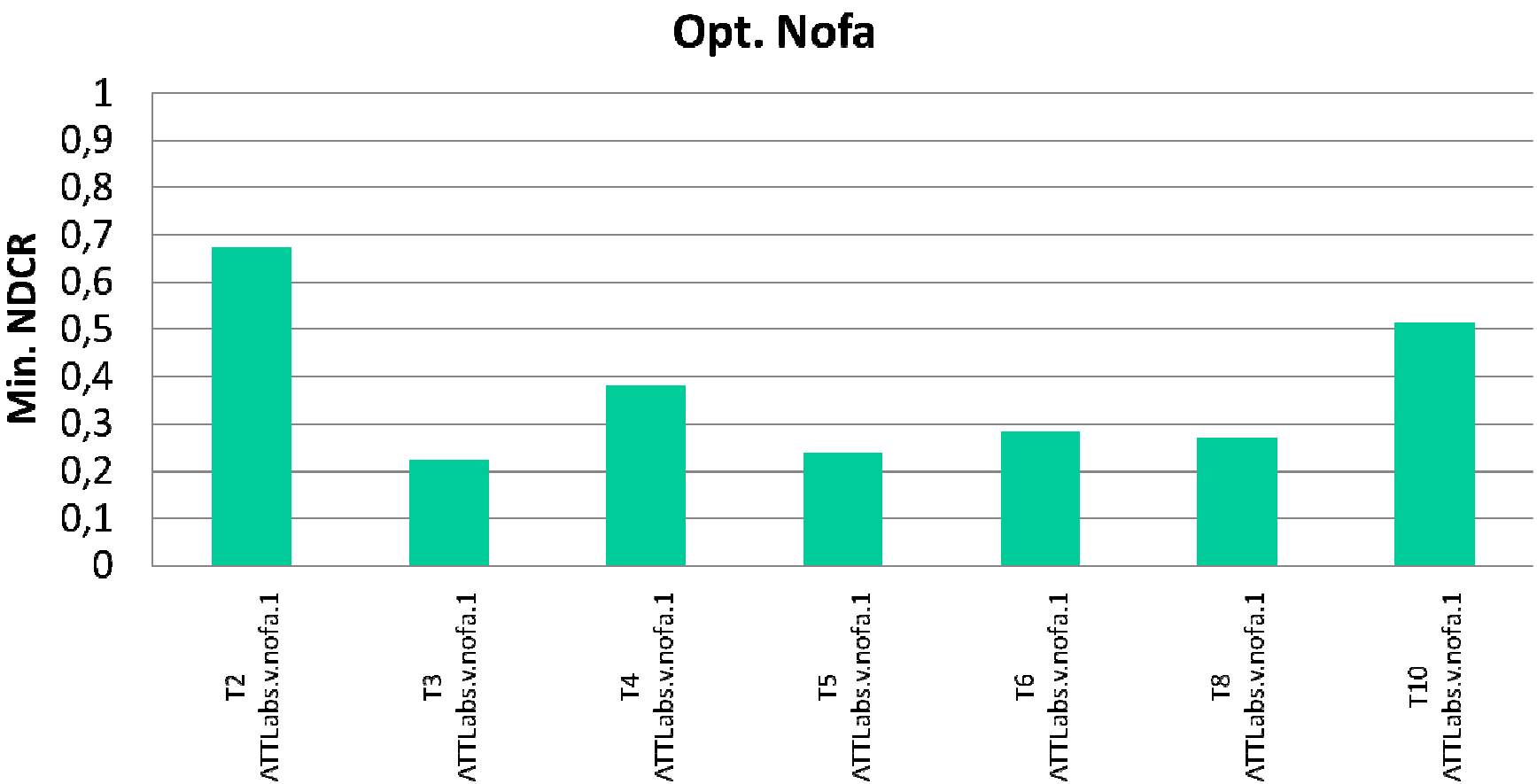
Top “video-only” runs



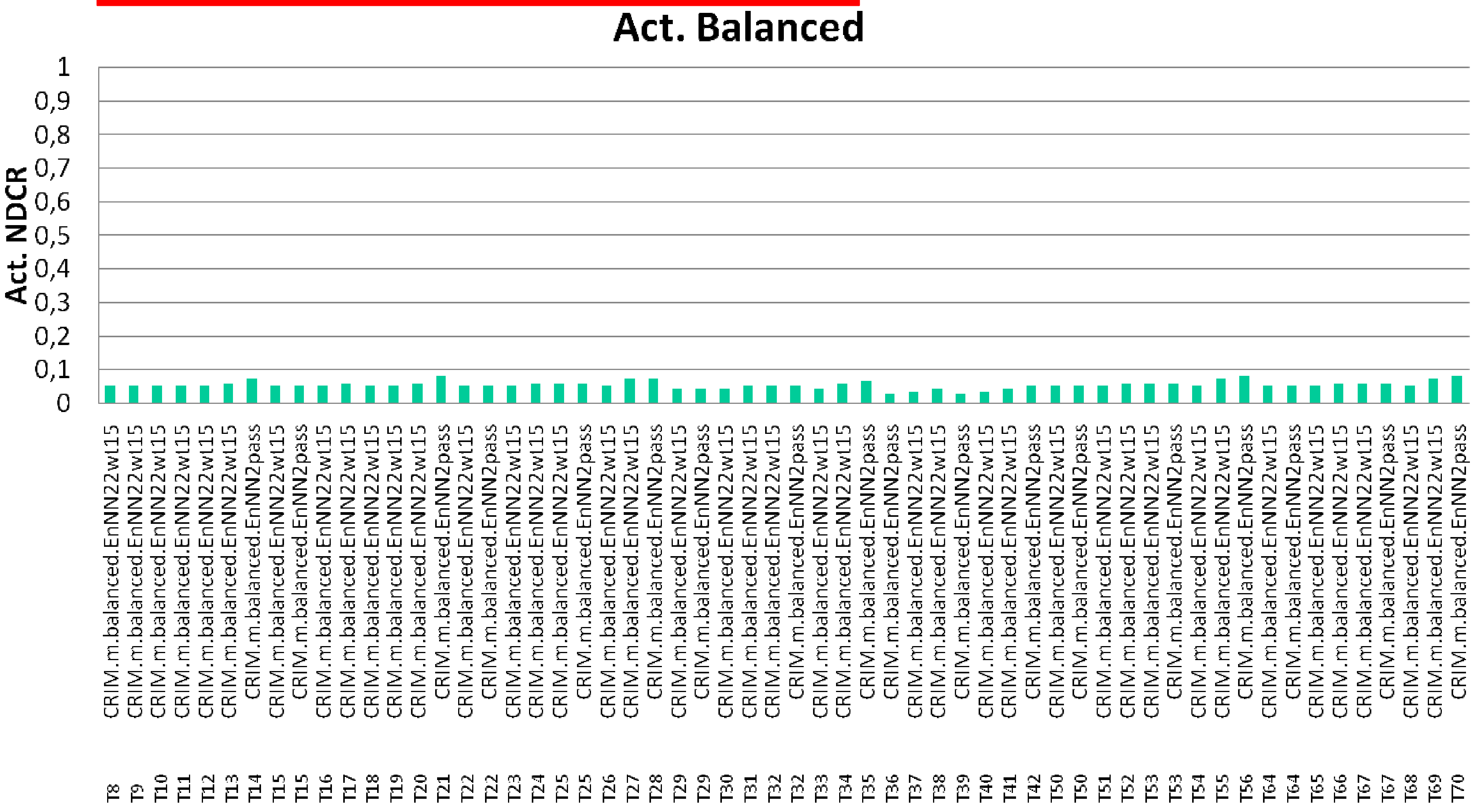
Top “video-only” runs



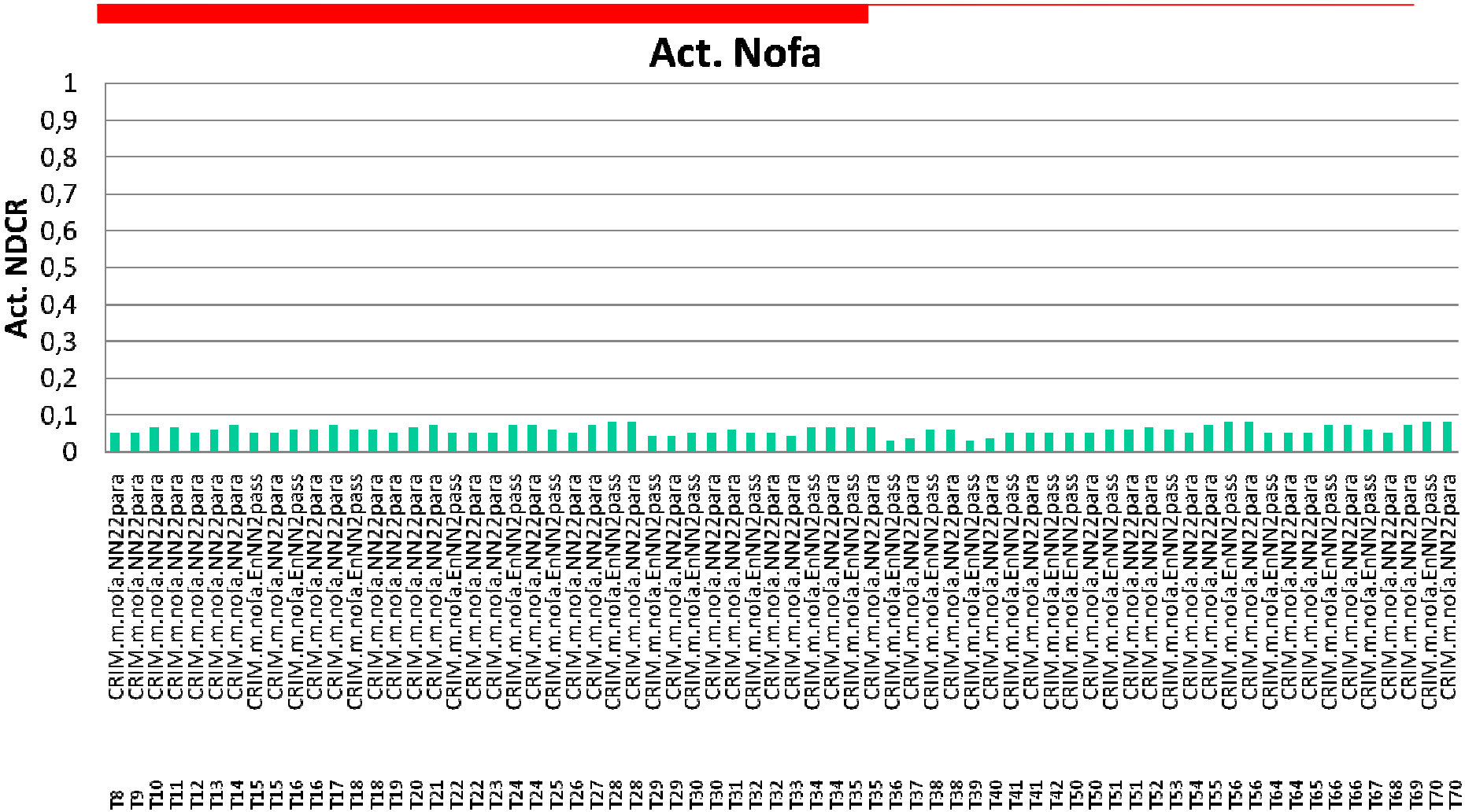
Top “video-only” runs



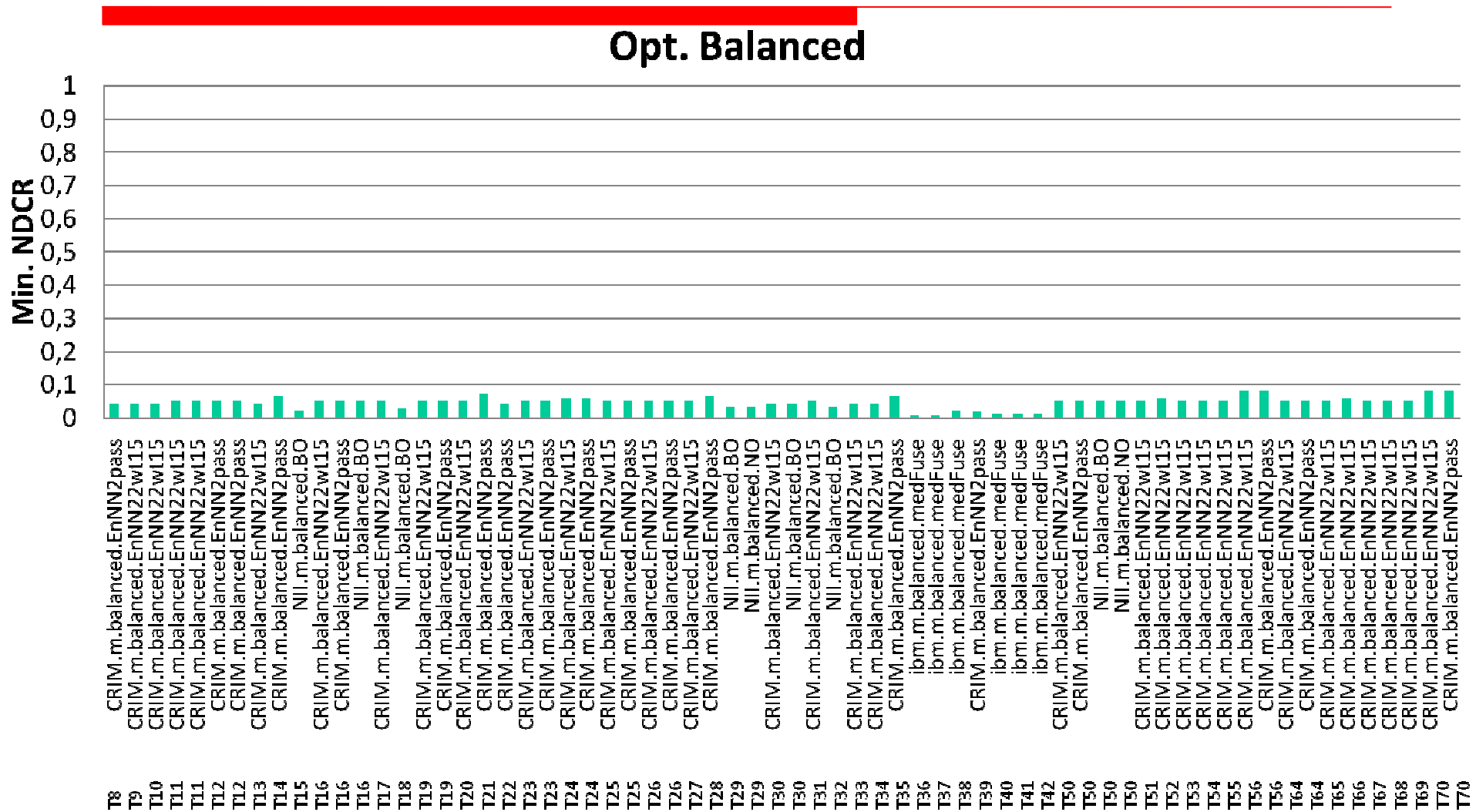
Top “video + audio” runs



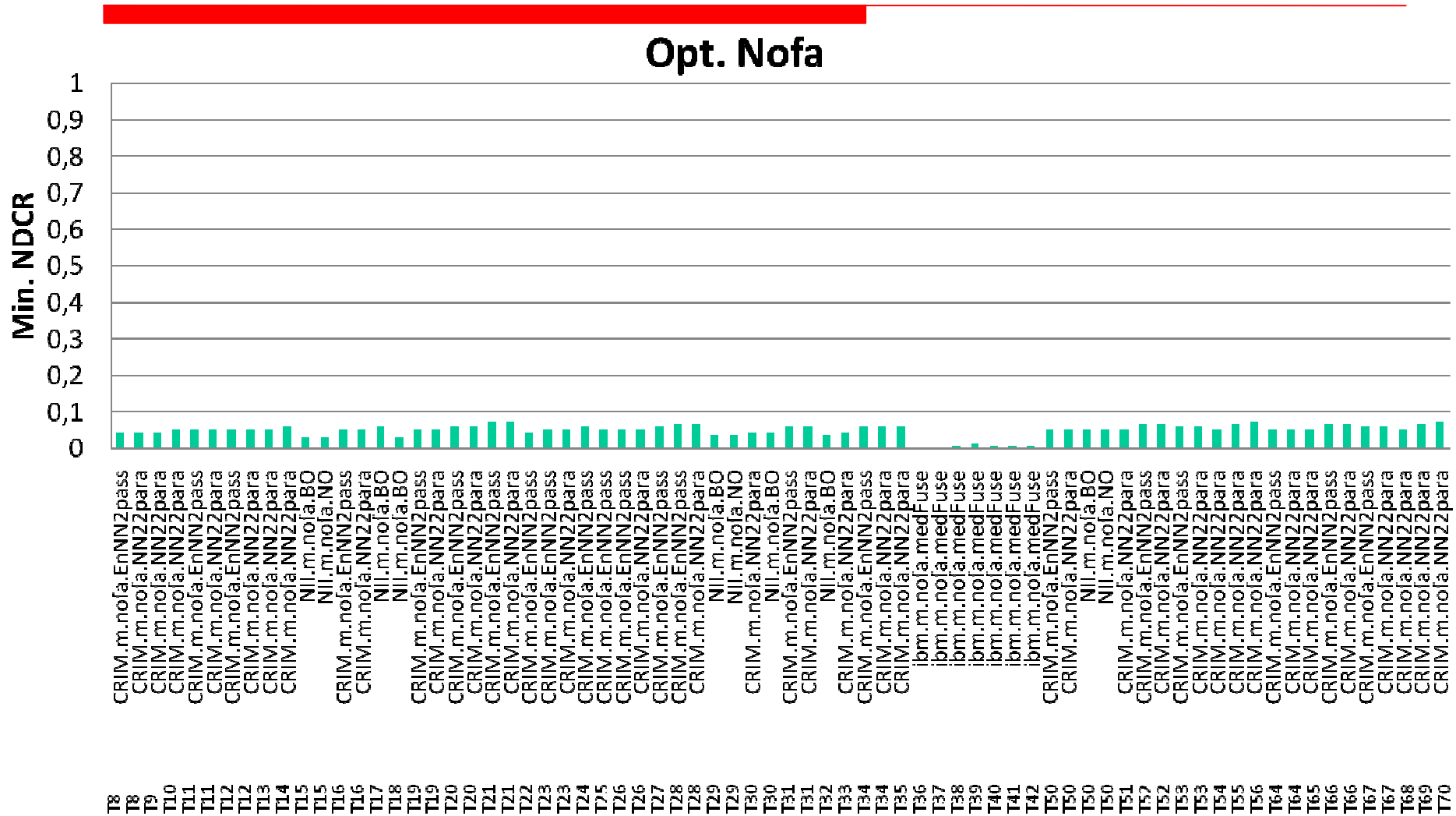
Top “video+audio” runs



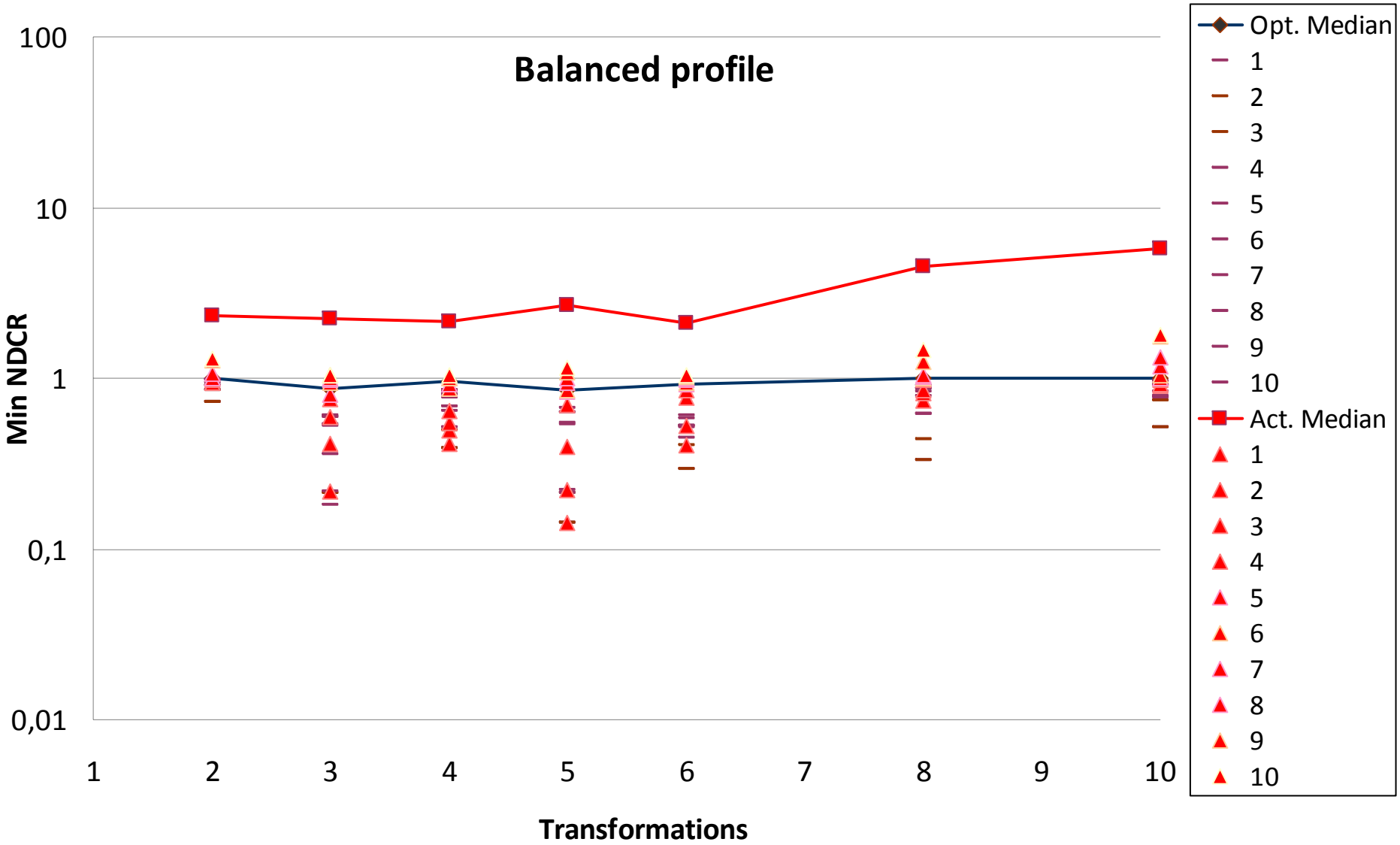
Top “video + audio” runs



Top “video+audio” runs



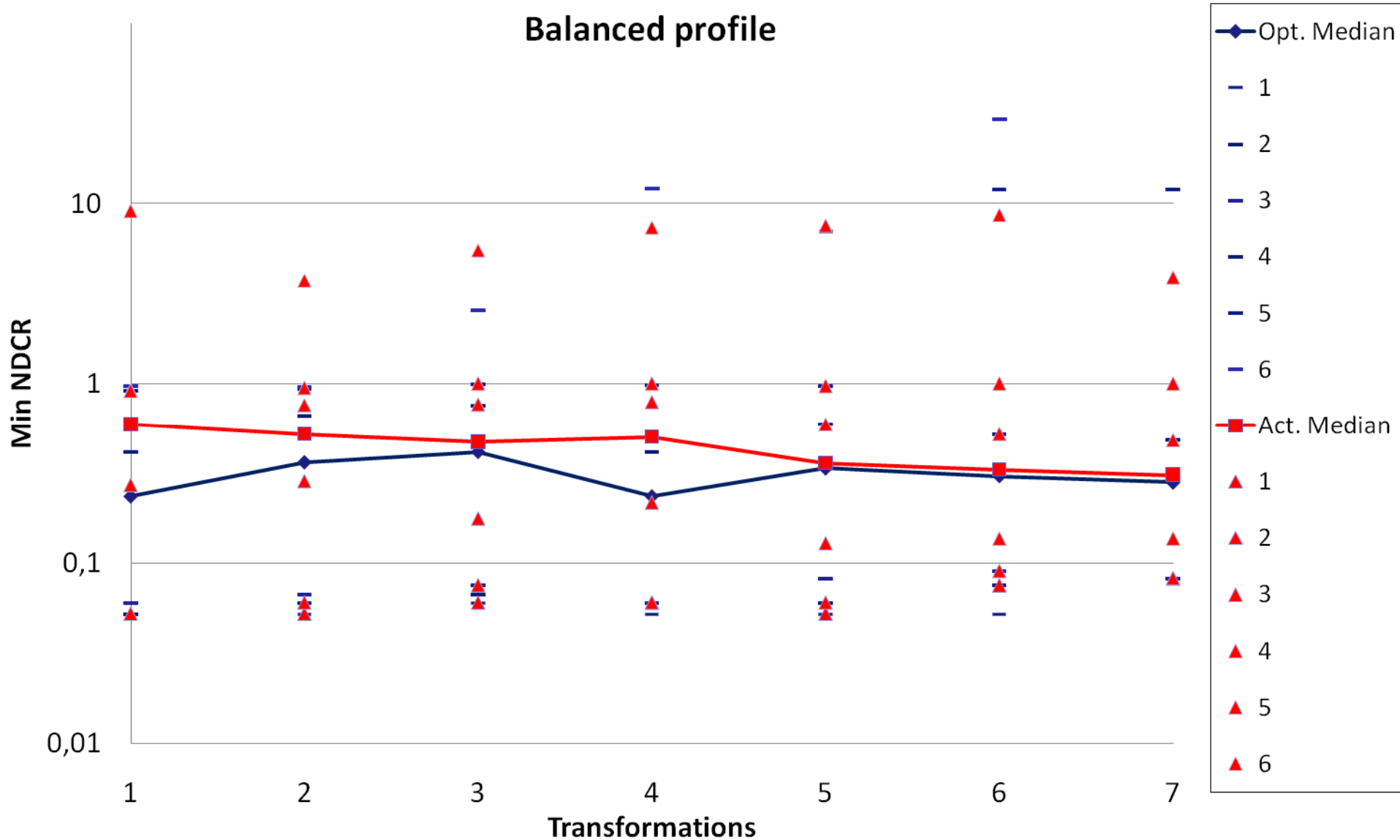
CBCD video-only detection (Top 10 performance)



T2: Pict. In Pict. T3: Insertion of patterns T4: Strong Re-encoding T5: Change of gamma

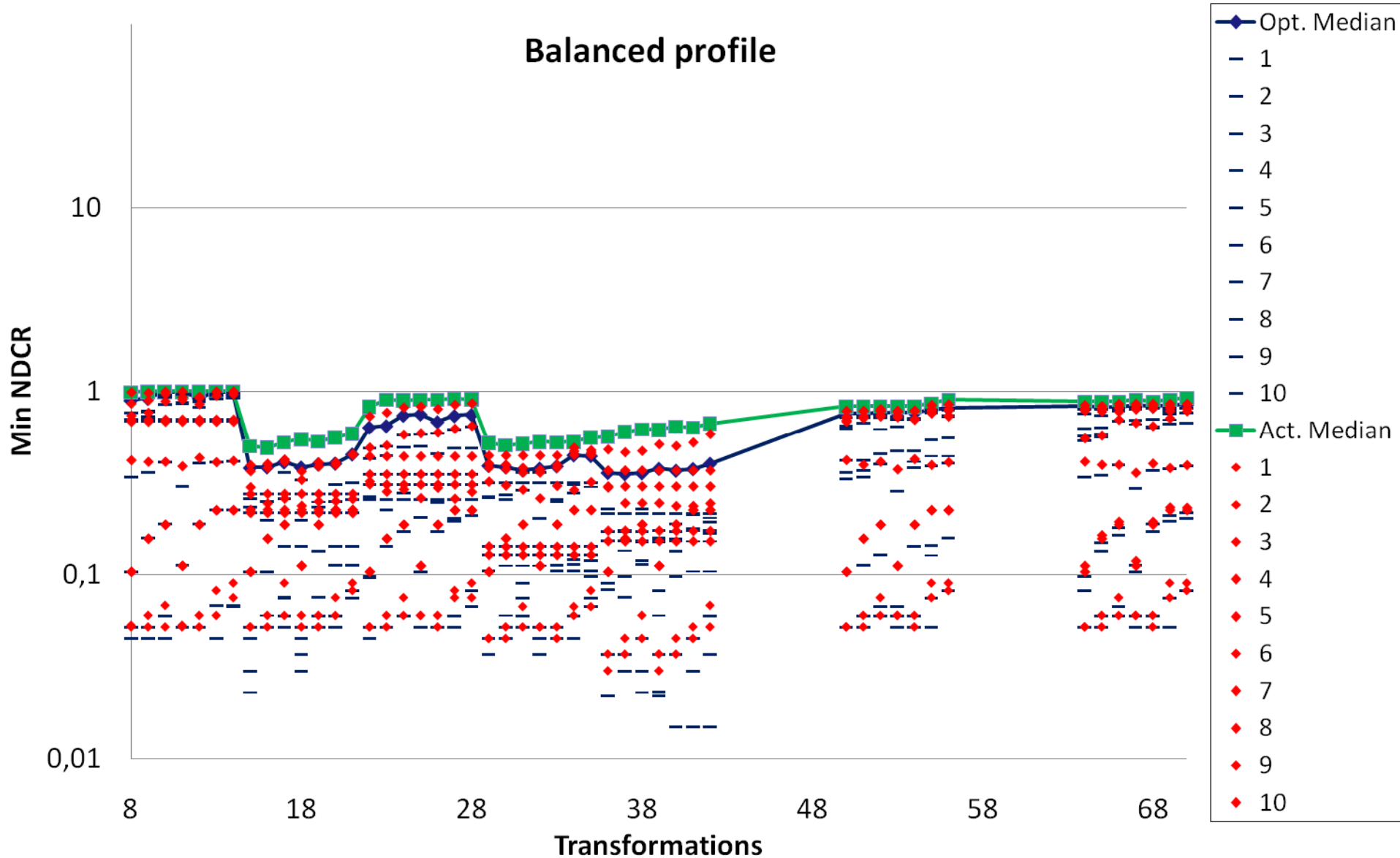
T6 : Frame dropping T8 : Post Production T10: Random combination of 3 transformations

CBCD audio-only detection (6 submitted runs)

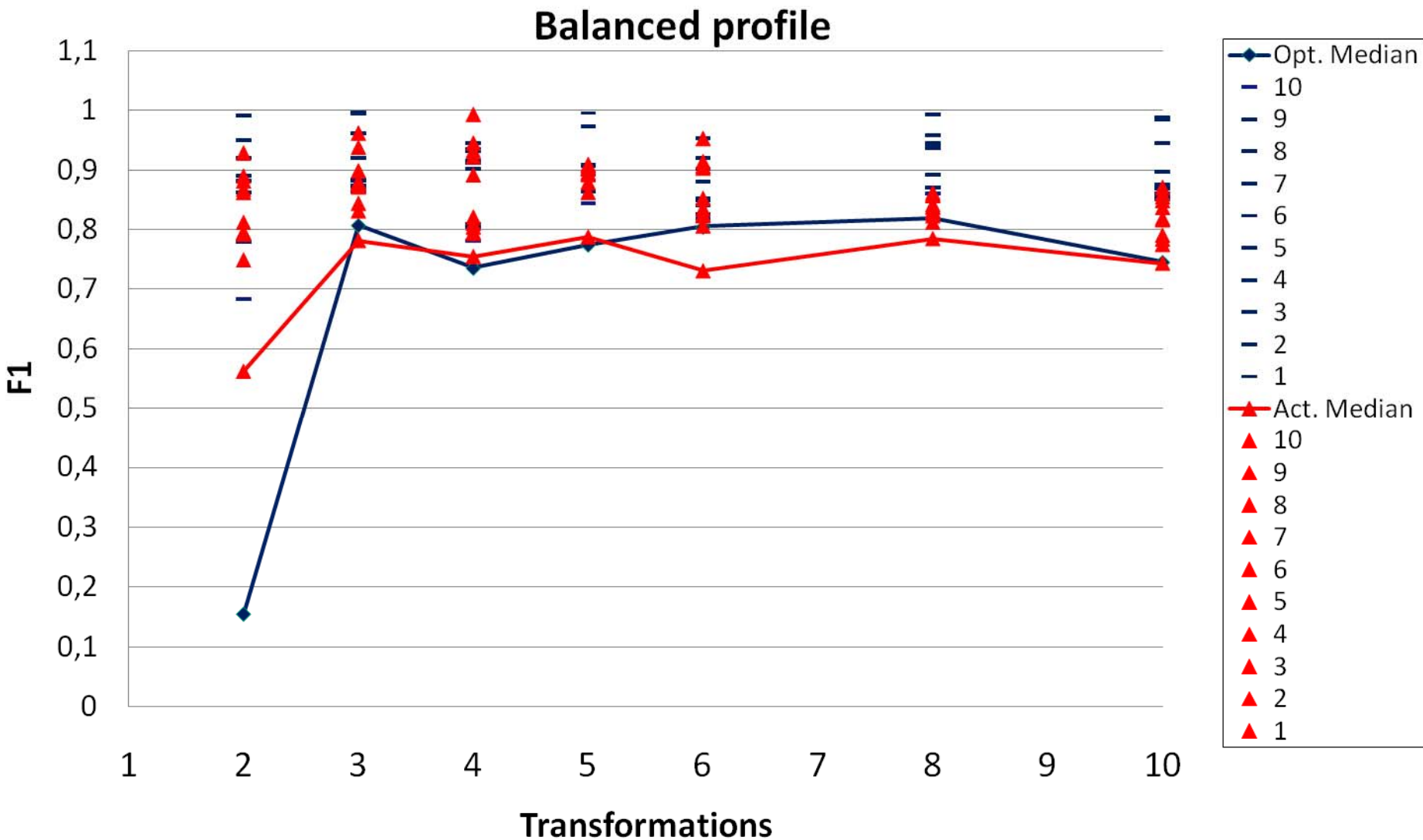


T1: nothing T2: mp3 compression T3: mp3 compression & multiband companding T4: bandwidth limit & single-band companding
T5 : mix with speech T6 : mix with speech, then multiband compress T7: bandpass filter, mix with speech, compress

CBCD video+audio detection(Top 10 performance)



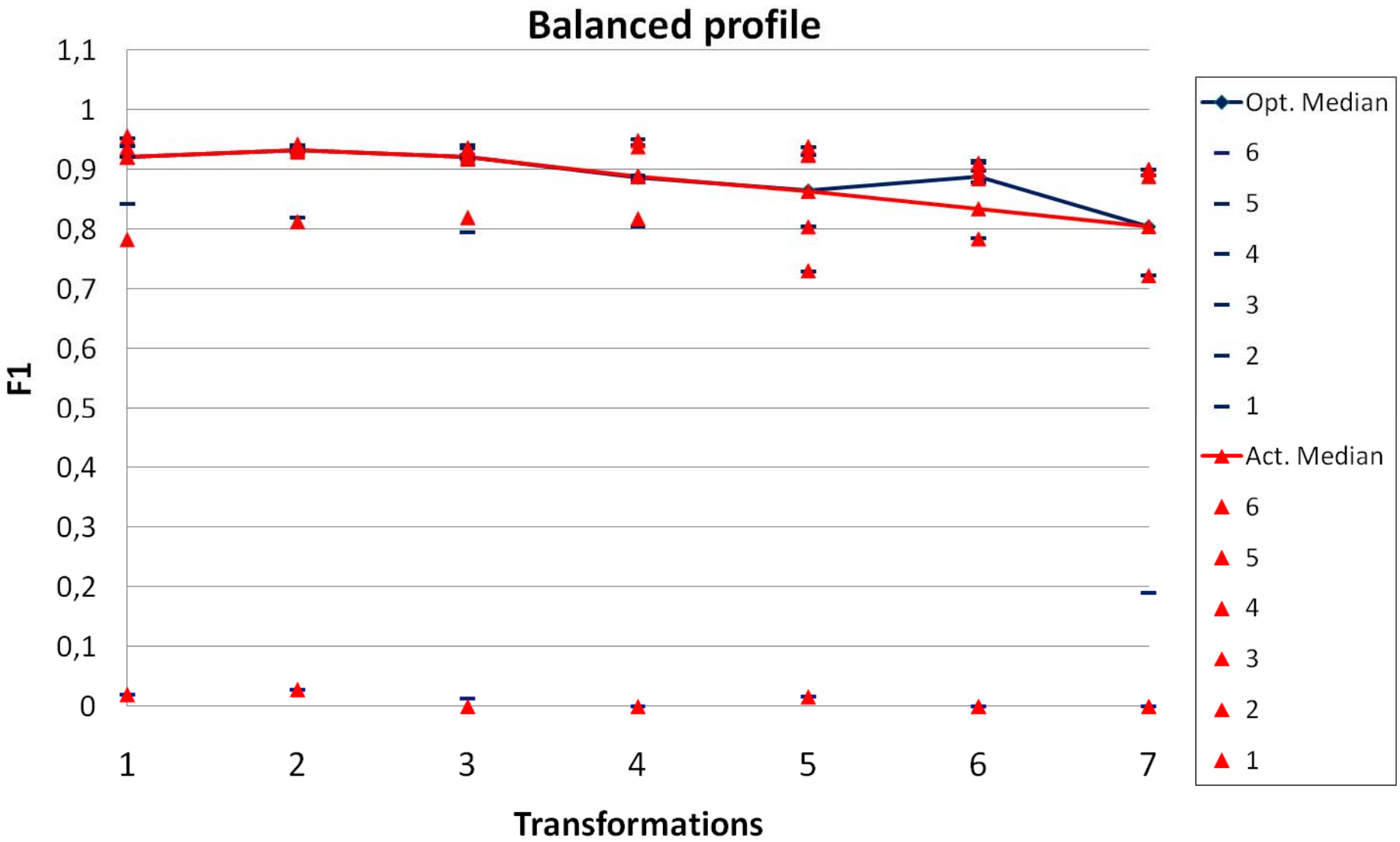
CBCD video-only localization (Top 10 performance)



T2: Pict. In Pict. T3: Insertion of patterns T4: Strong Re-encoding T5: Change of gamma

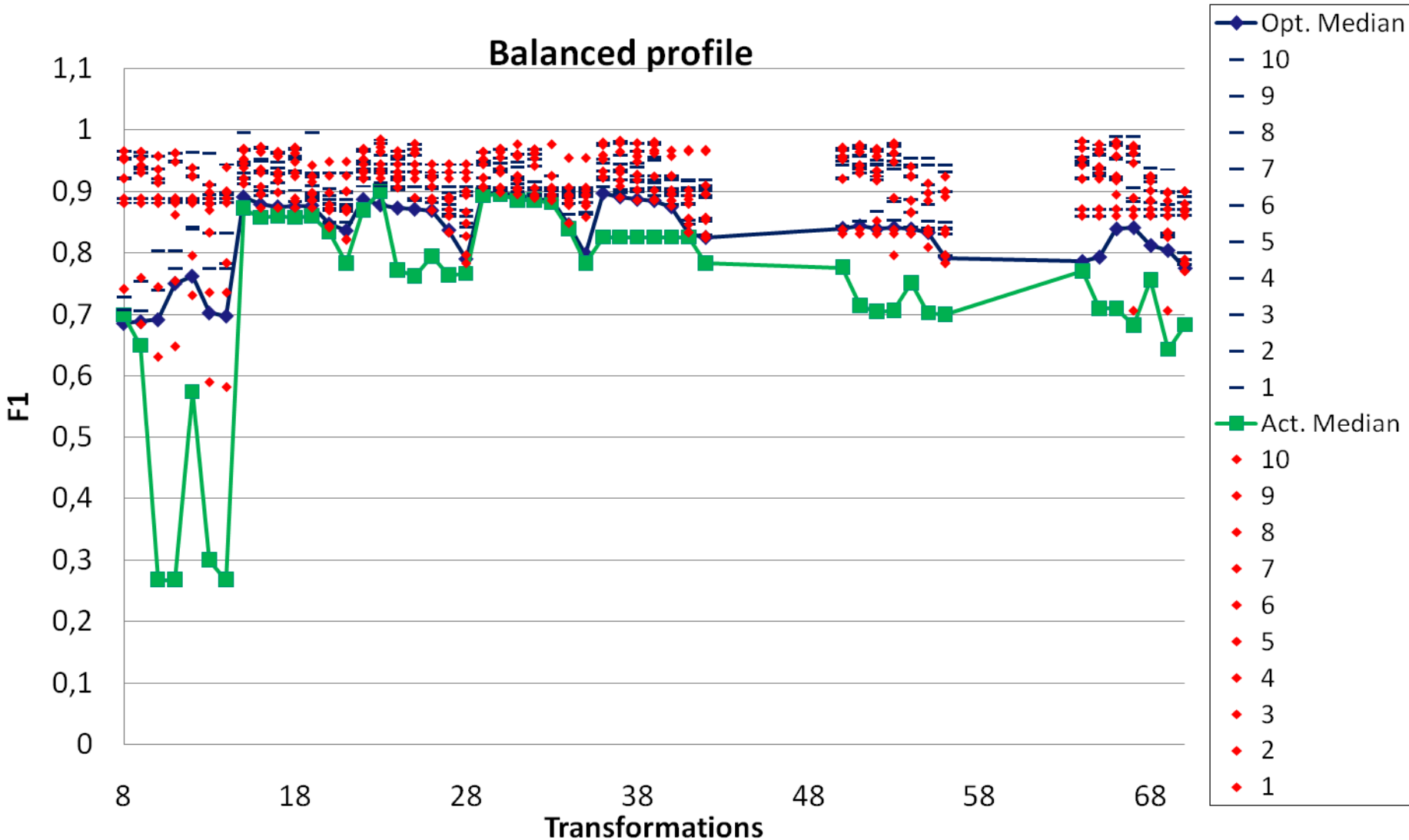
T6 : Frame dropping T8 : Post Production T10: Random combination of 3 transformations

CBCD audio-only localization (6 submitted runs)

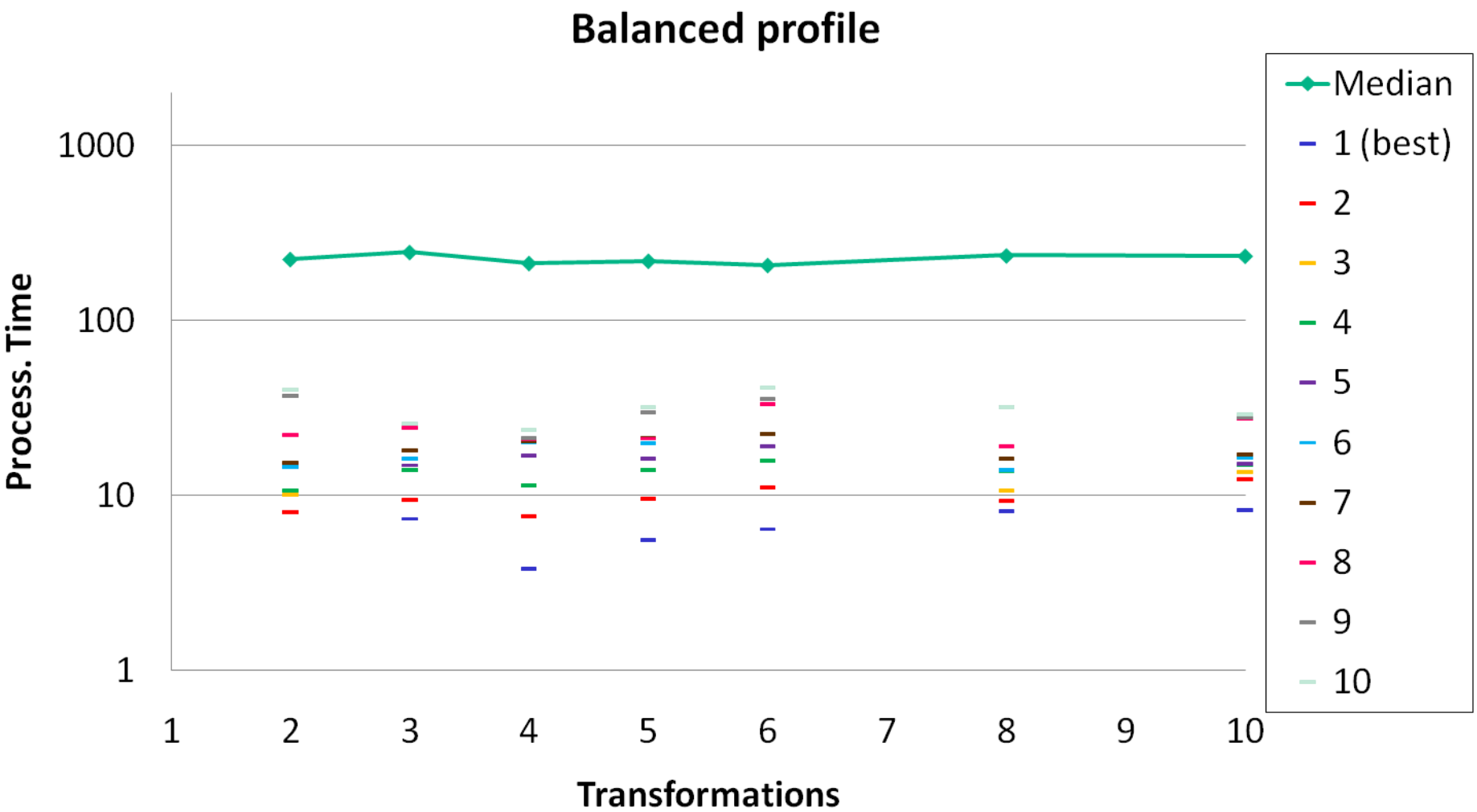


T1: nothing T2: mp3 compression T3: mp3 compression & multiband companding T4: bandwidth limit & single-band companding T5 : mix with speech T6 : mix with speech, then multiband compress T7: bandpass filter, mix with speech, compress

CBCD video+audio localization (Top 10 performance)



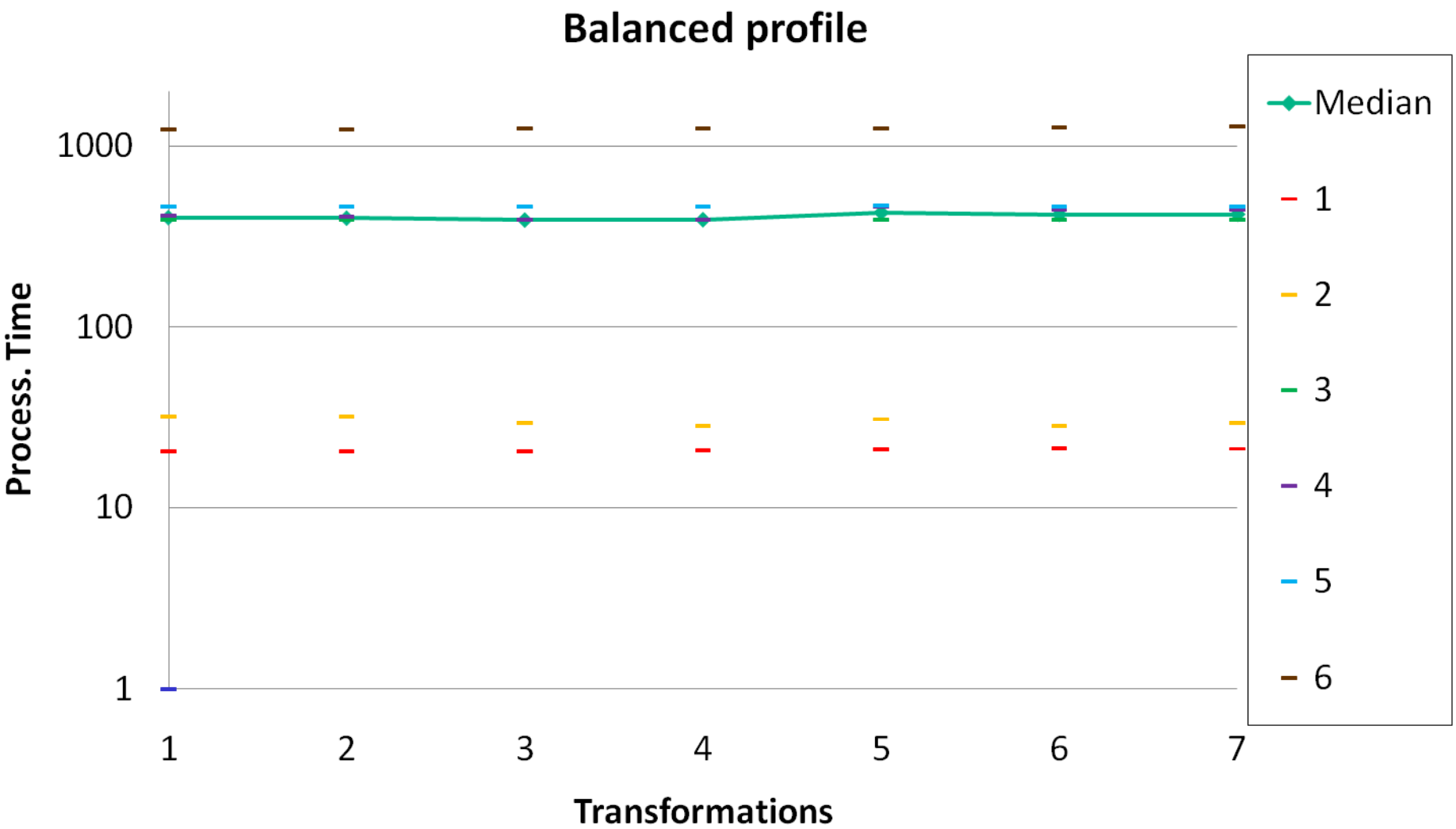
CBCD video-only efficiency (Top 10 performance)



T2: Pict. In Pict. T3: Insertion of patterns T4: Strong Re-encoding T5: Change of gamma

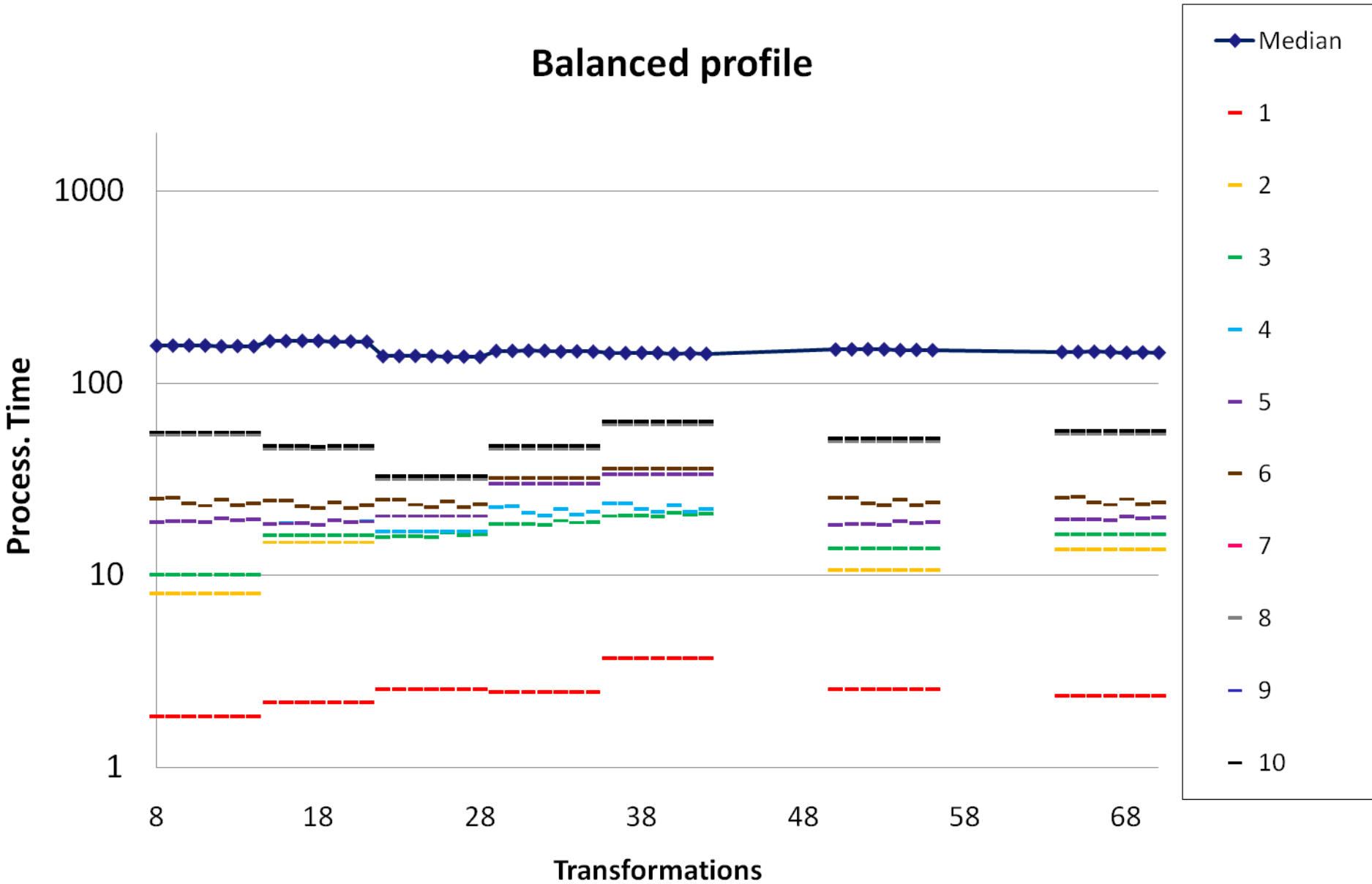
T6 : Frame dropping T8 : Post Production T10: Random combination of 3 transformations

CBCD audio-only efficiency (6 submitted runs)

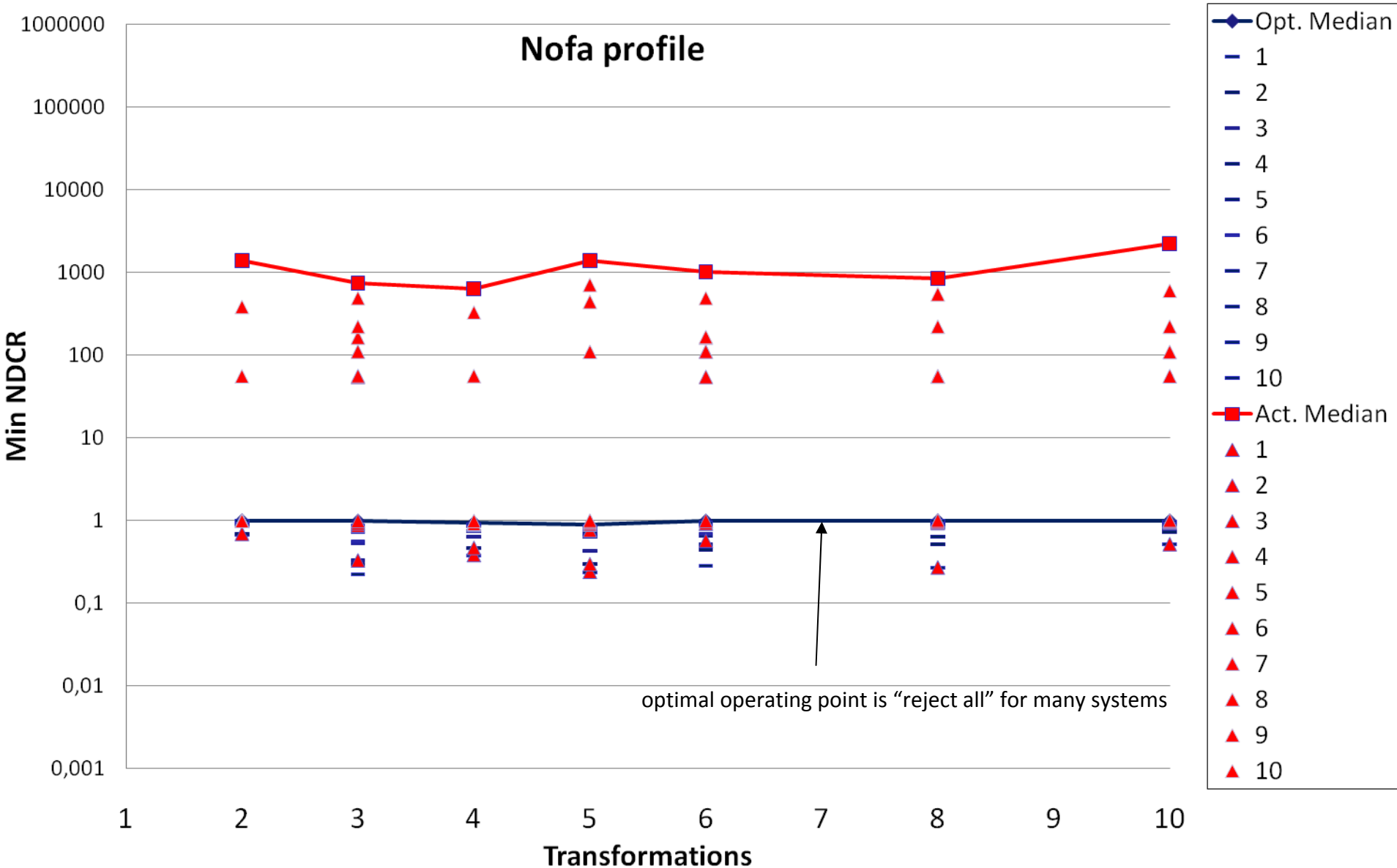


T1: nothing T2: mp3 compression T3: mp3 compression & multiband companding T4: bandwidth limit & single-band companding T5 : mix with speech T6 : mix with speech, then multiband compress T7: bandpass filter, mix with speech, compress

CBCD video+audio efficiency (Top 10 performance)



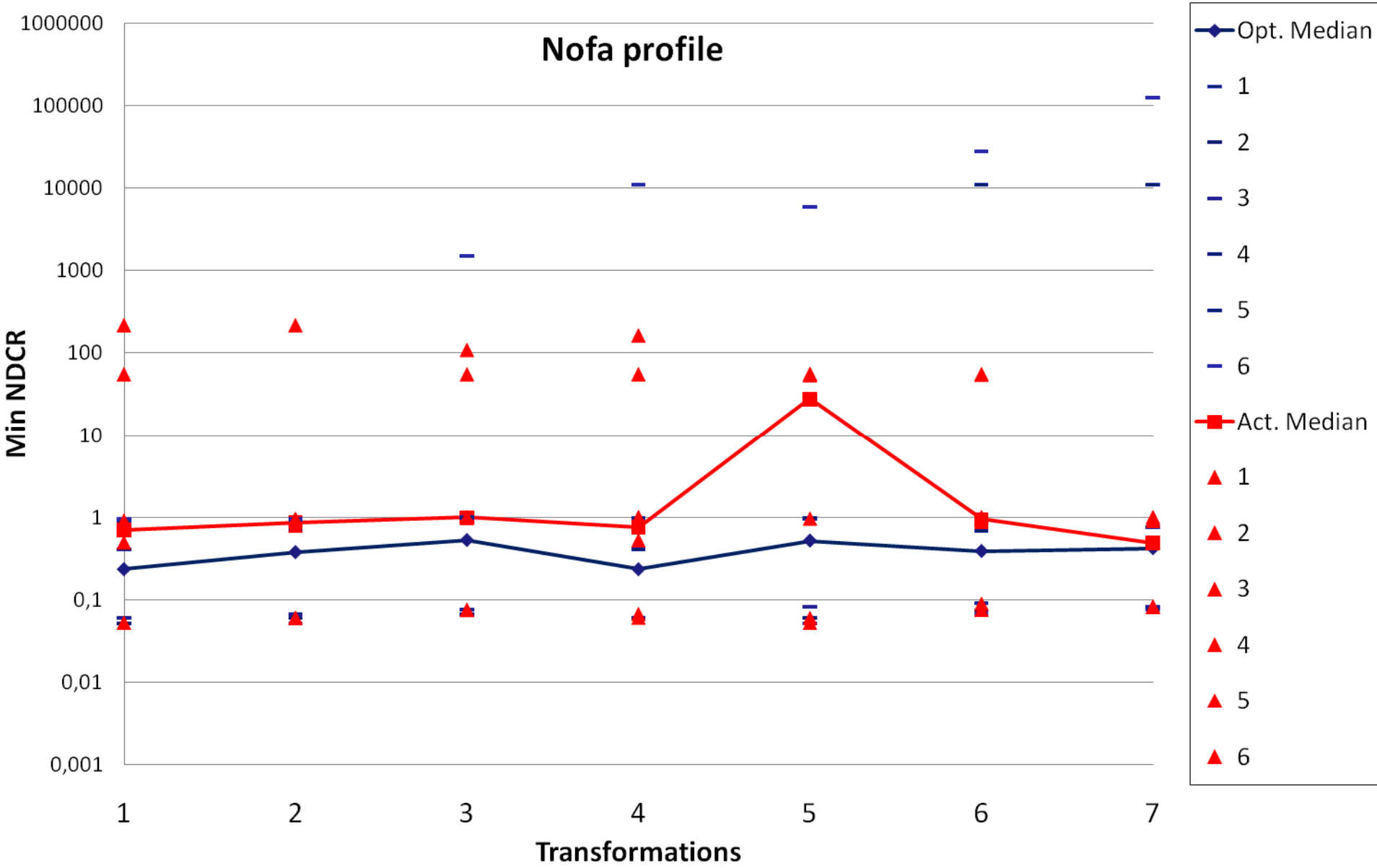
CBCD video-only detection (Top 10 performance per T)



T2: Pict. In Pict. T3: Insertion of patterns T4: Strong Re-encoding T5: Change of gamma

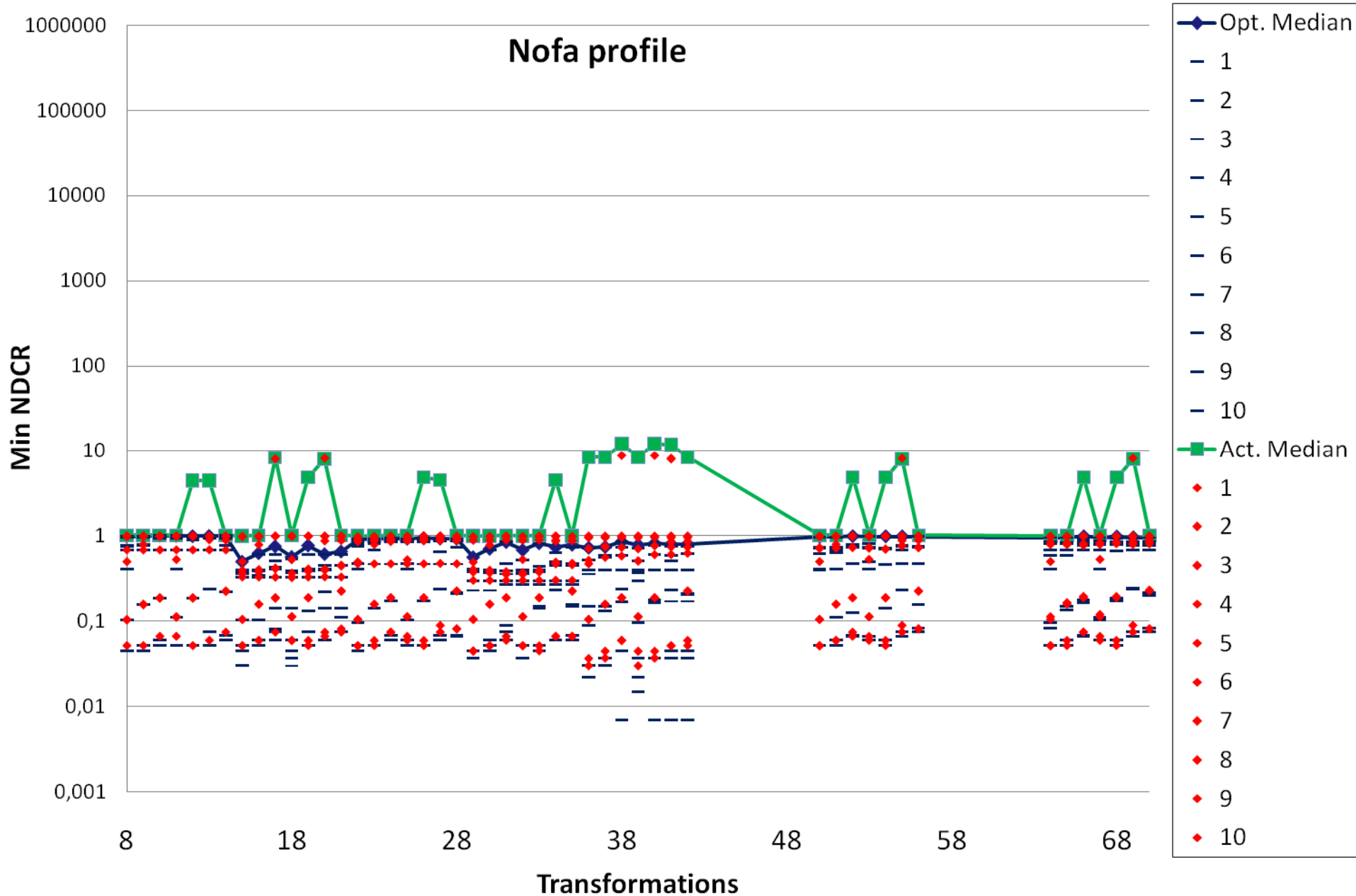
T6 : Frame dropping T8 : Post Production T10: Random combination of 3 transformations

CBCD audio-only detection (6 submitted runs)

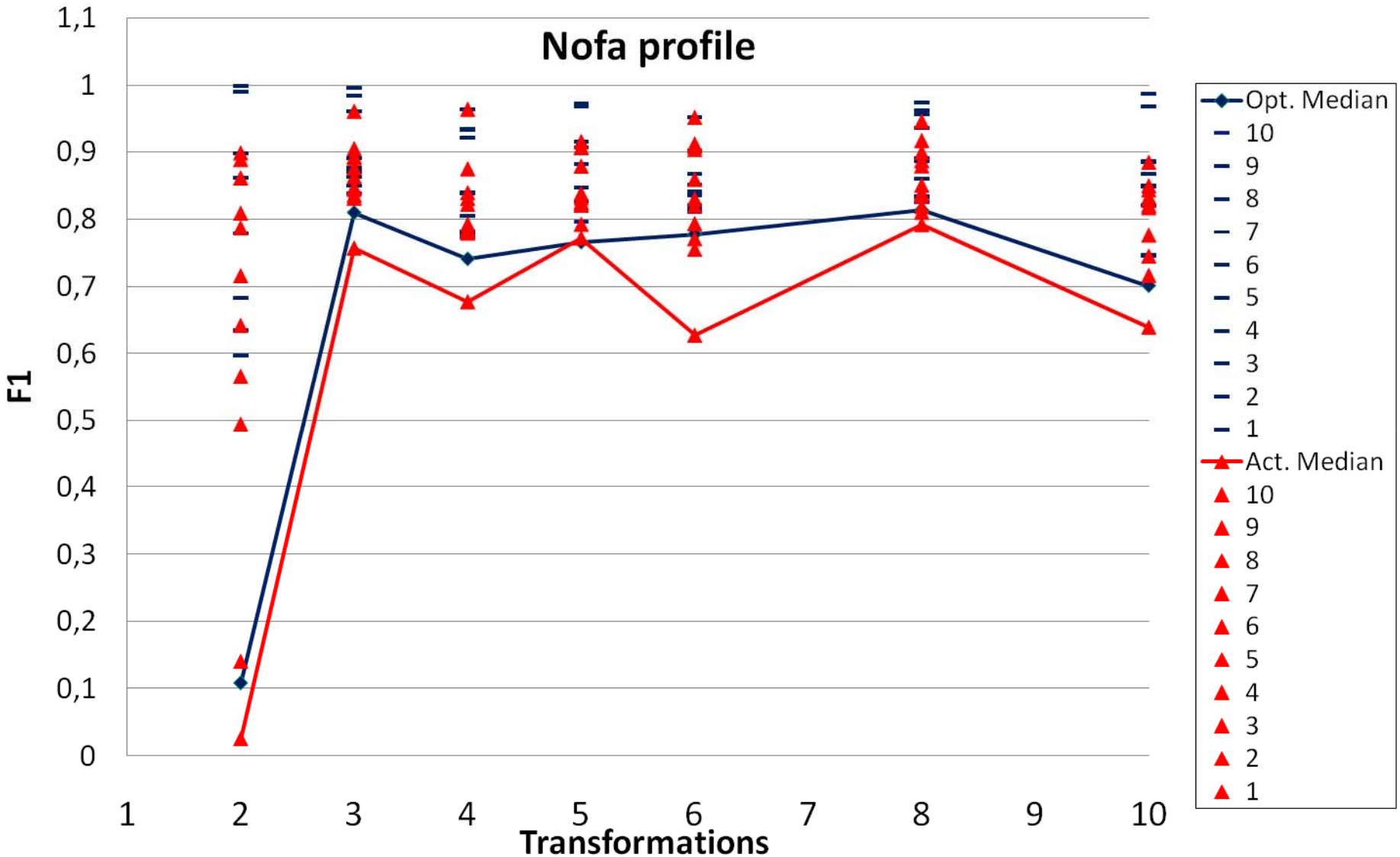


T1: nothing
T2: mp3 compression
T3: mp3 compression & multiband companding
T4: bandwidth limit & single-band companding
T5 : mix with speech
T6 : mix with speech, then multiband compress
T7: bandpass filter, mix with speech,

CBCD video+audio detection (Top 10 performance)



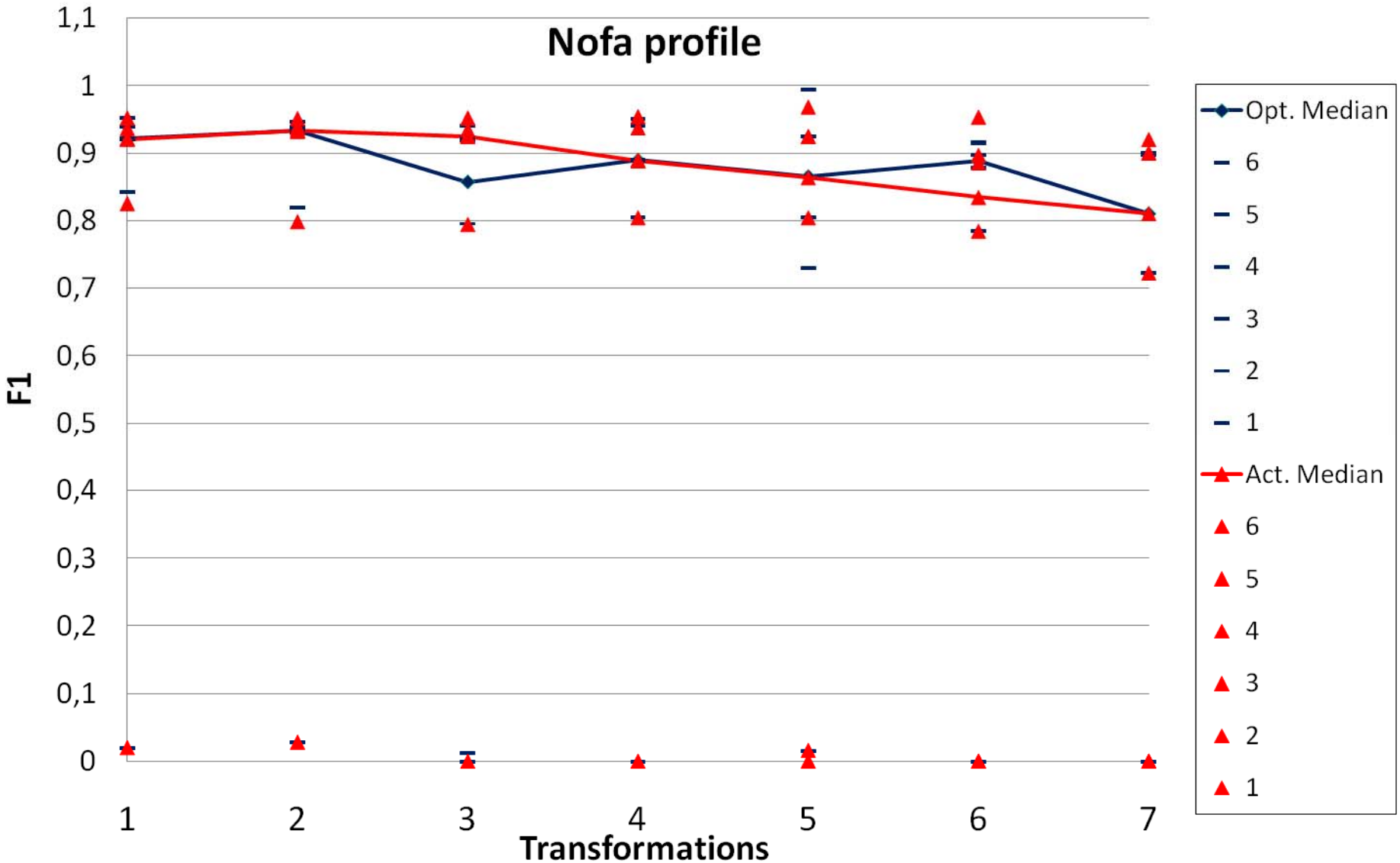
CBCD video-only localization (Top 10 performance)



T2: Pict. In Pict. T3: Insertion of patterns T4: Strong Re-encoding T5: Change of gamma

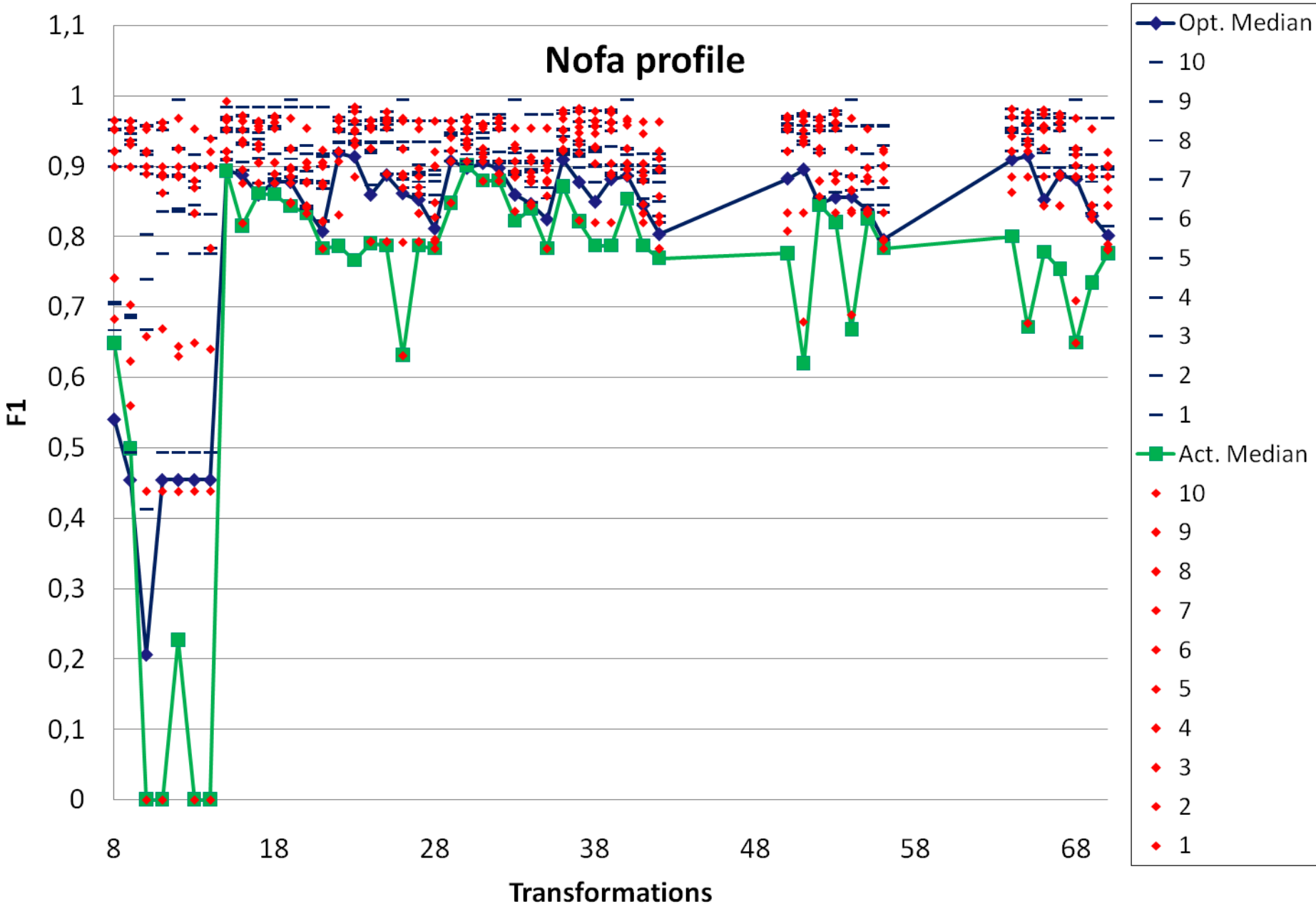
T6 : Frame dropping T8 : Post Production T10: Random combination of 3 transformations

CBCD audio-only localization (6 submitted runs)

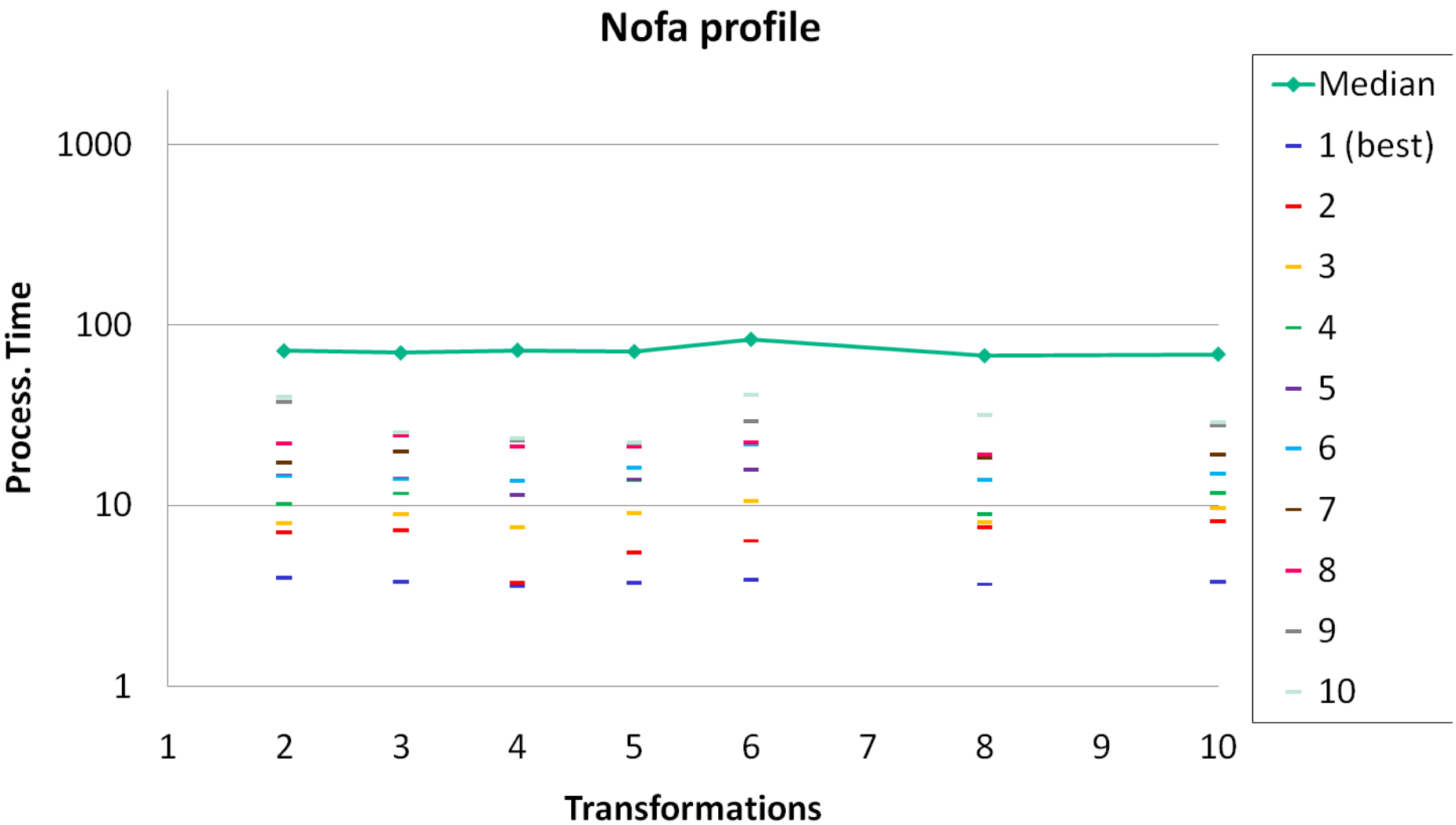


T1: nothing T2: mp3 compression T3: mp3 compression & multiband companding T4: bandwidth limit & single-band companding T5 : mix with speech T6 : mix with speech, then multiband compress T7: bandpass filter, mix with speech, compress

CBCD video+audio localization (Top 10 performance)



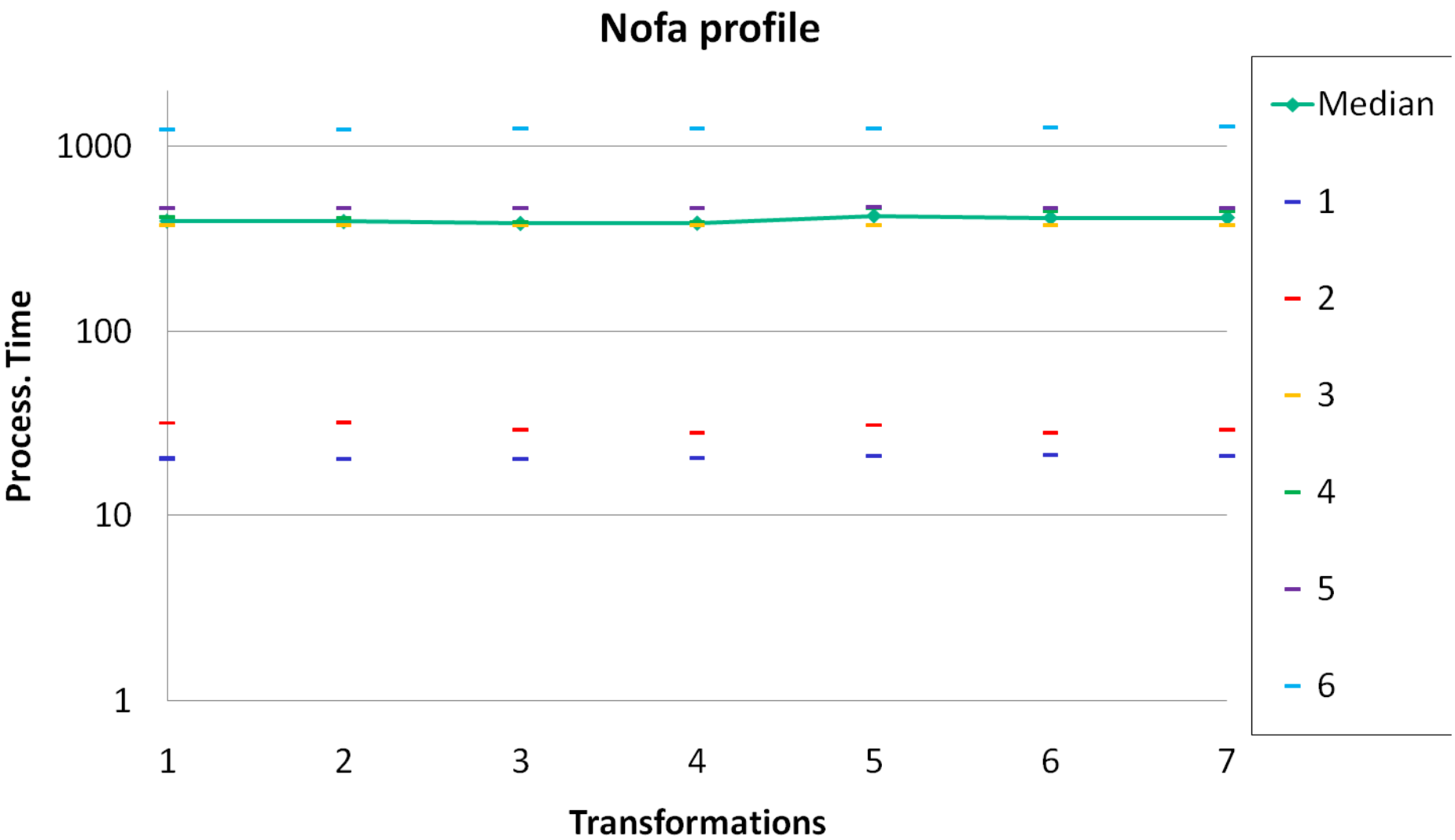
CBCD video-only efficiency (Top 10 performance)



T2: Pict. In Pict. T3: Insertion of patterns T4: Strong Re-encoding T5: Change of gamma

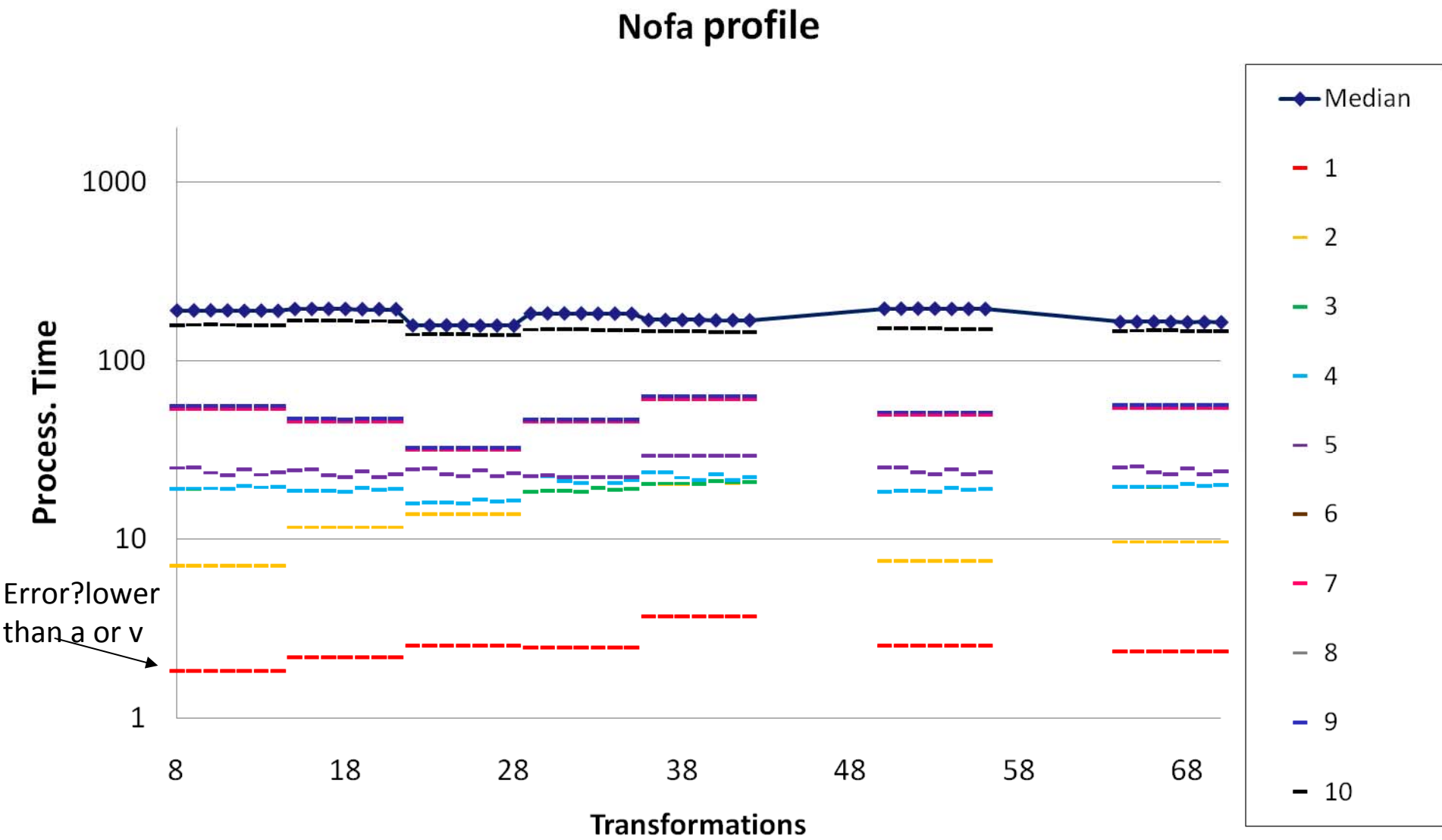
T6 : Frame dropping T8 : Post Production T10: Random combination of 3 transformations

CBCD audio-only efficiency (6 submitted runs)

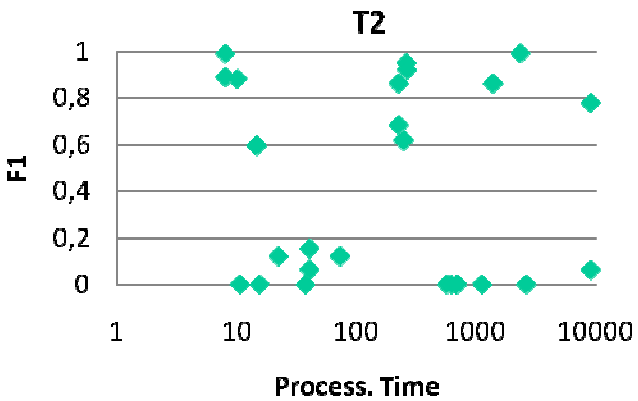


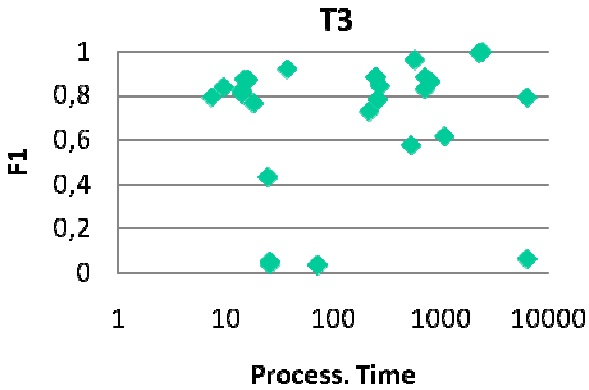
T1: nothing
T2: mp3 compression
T3: mp3 compression & multiband companding
T4: bandwidth limit & single-band companding
T5 : mix with speech
T6 : mix with speech, then multiband compress
T7: bandpass filter, mix with speech, compress

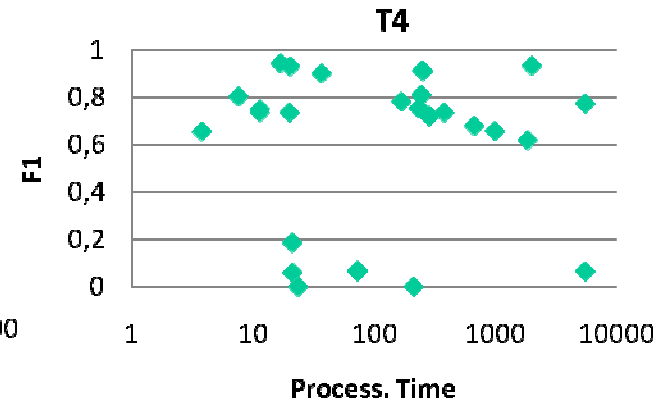
CBCD video+audio efficiency (Top 10 performance)

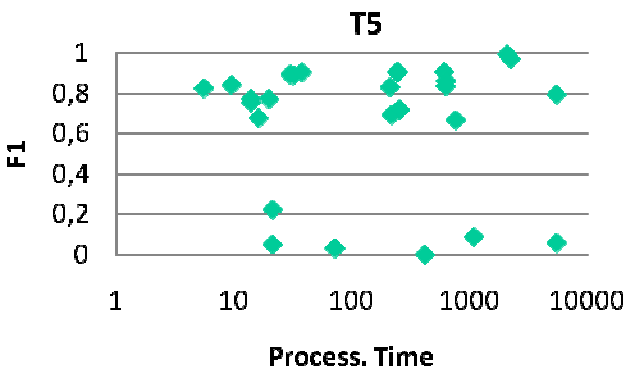


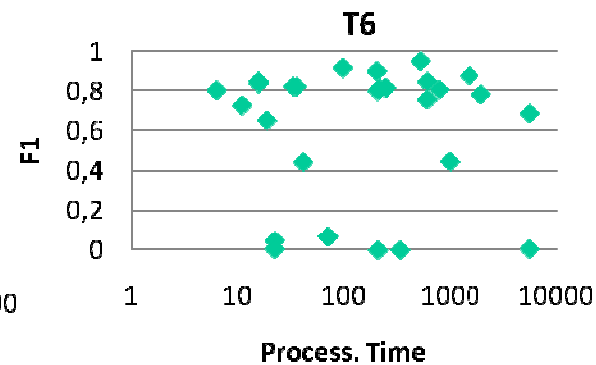
Video only – Balanced runs by transformations

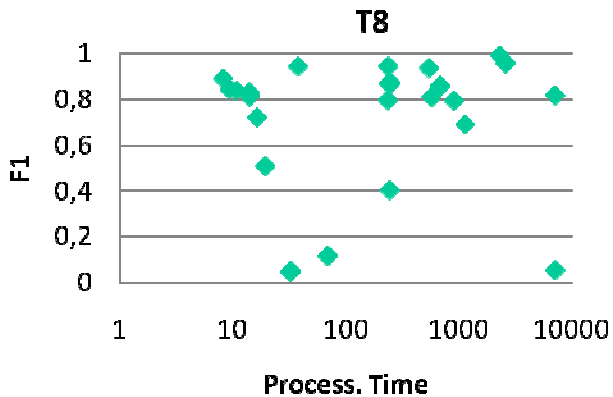


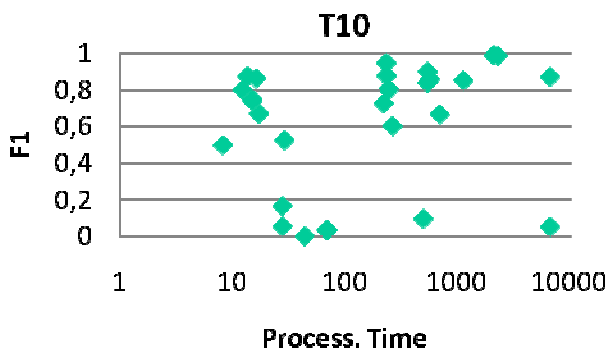






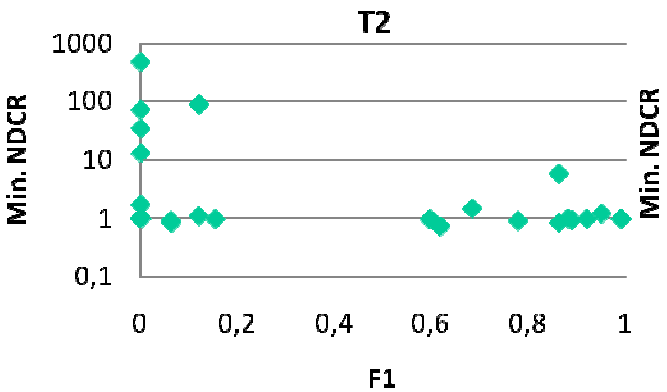


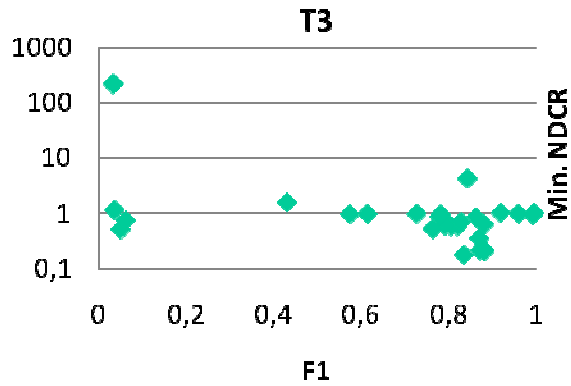


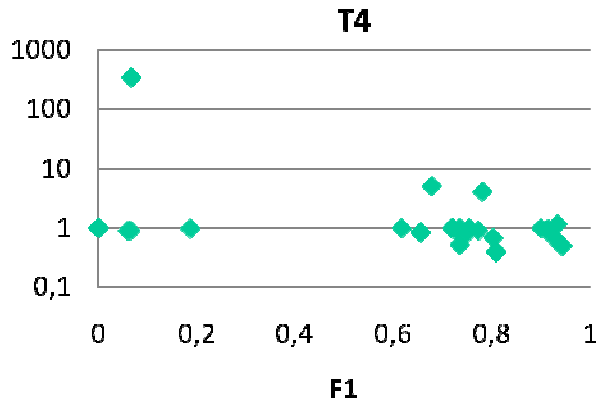


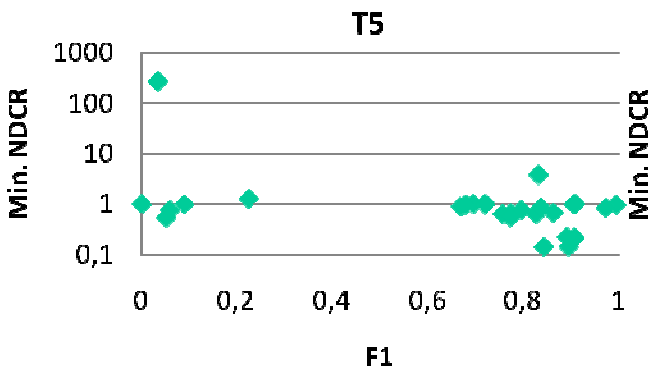
Increasing proc. time did not enhance localization. Few systems achieved high localization in small proc. time. Strong systems are efficient and precise.

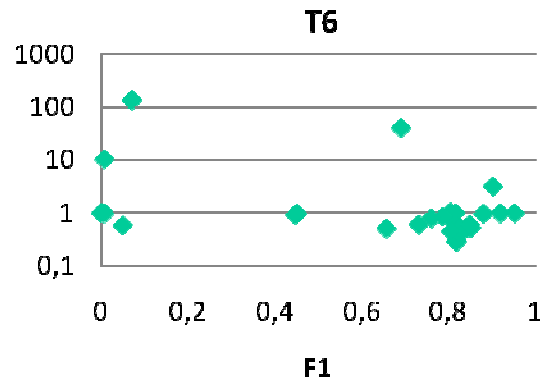
Video only – Balanced runs by transformations

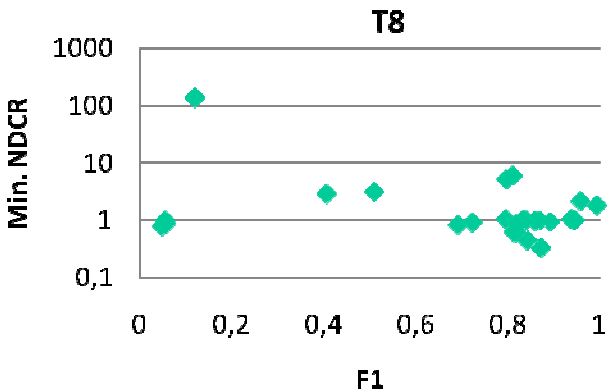


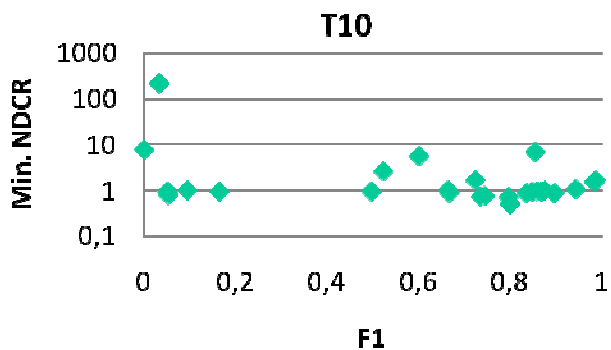






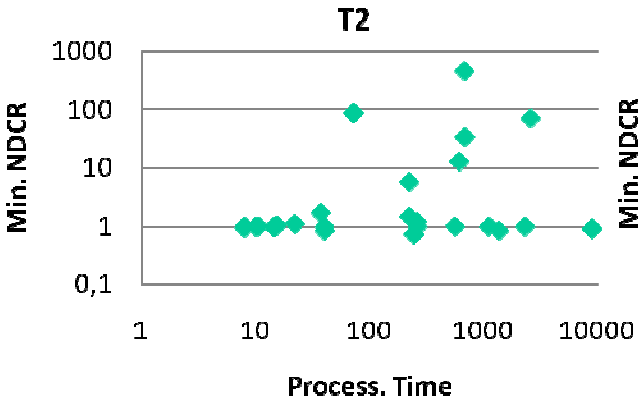


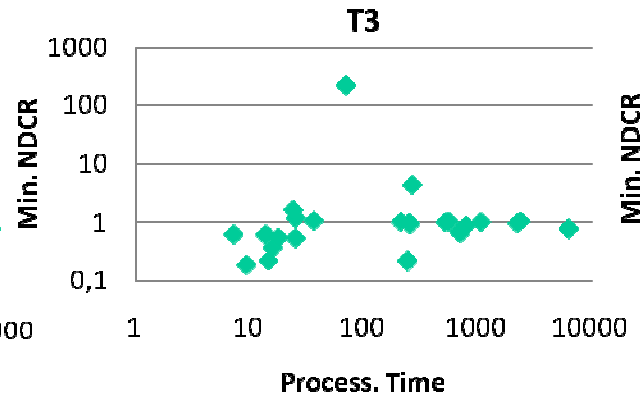


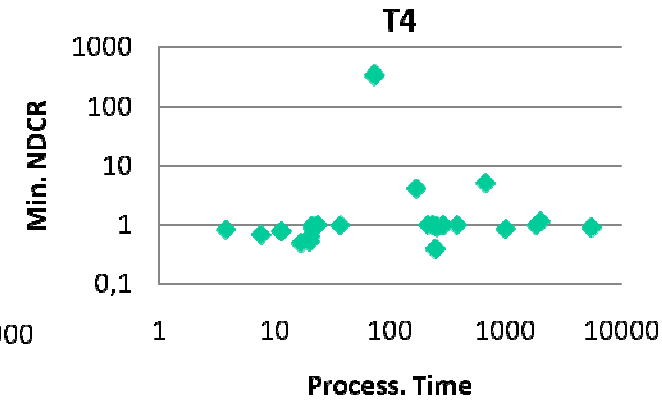


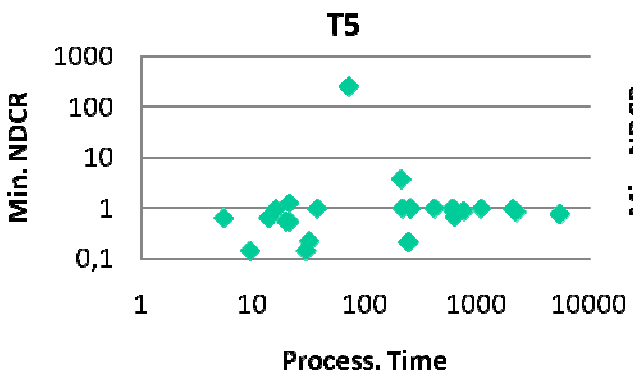
Most of the systems that are good in separating copies from non-copies (low NDCR) are also good in localization.

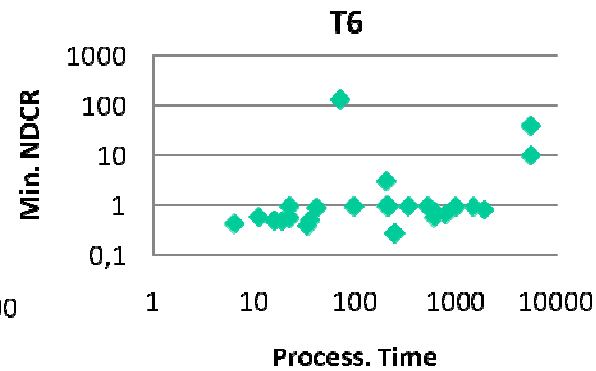
Video only – Balanced runs by transformations

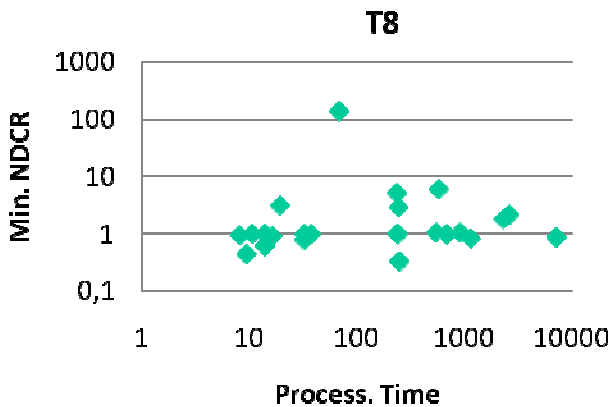


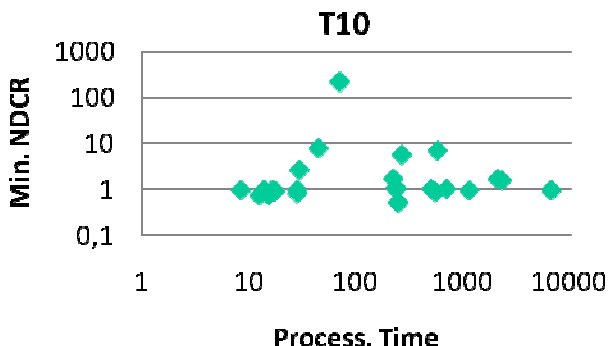






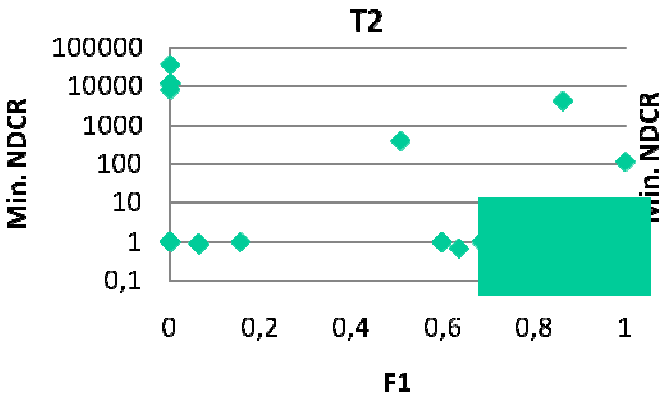


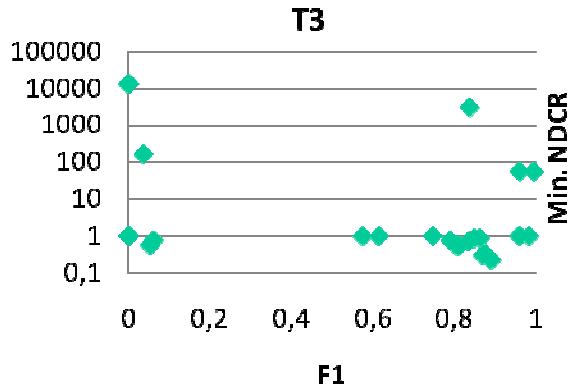


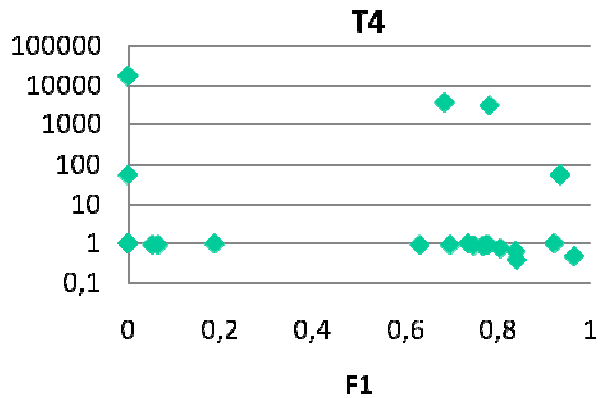


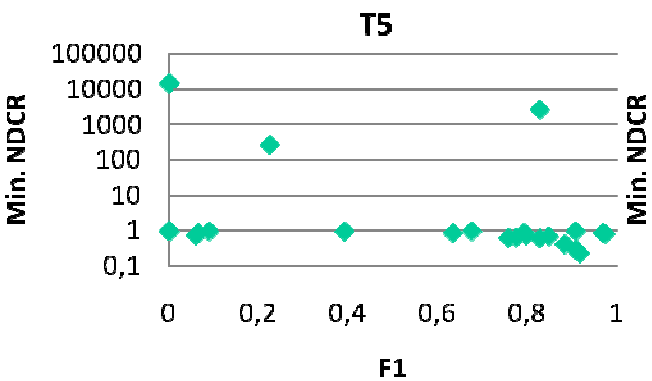
Increasing proc. time did not reduce the cost. Few good systems are fast with low cost.

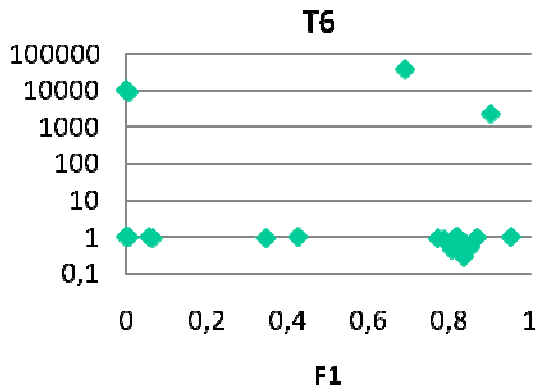
Video only – Nofa runs by transformations

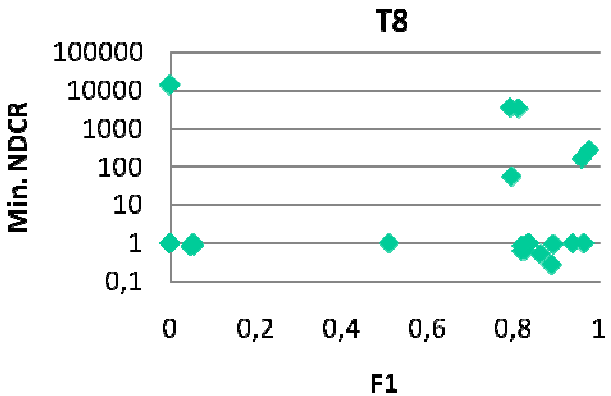


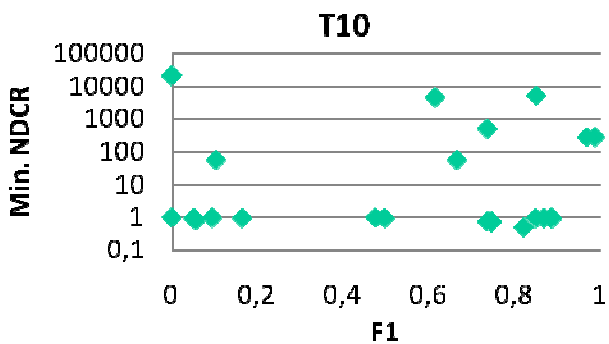






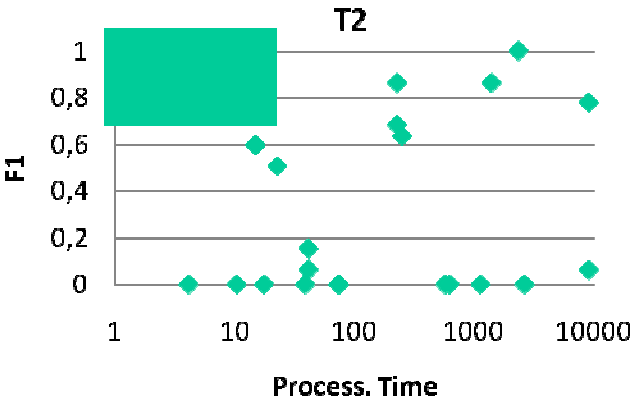


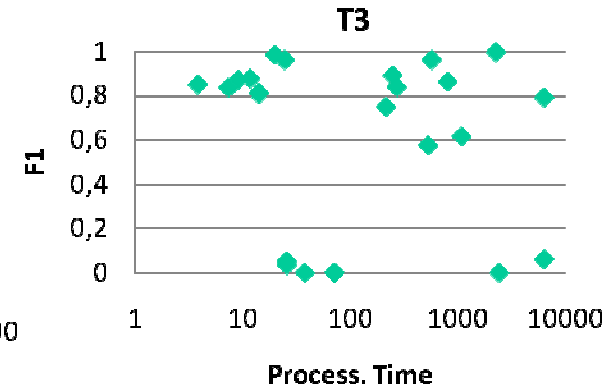


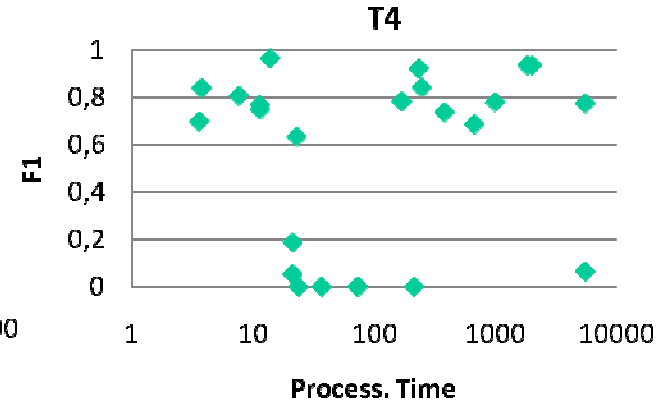


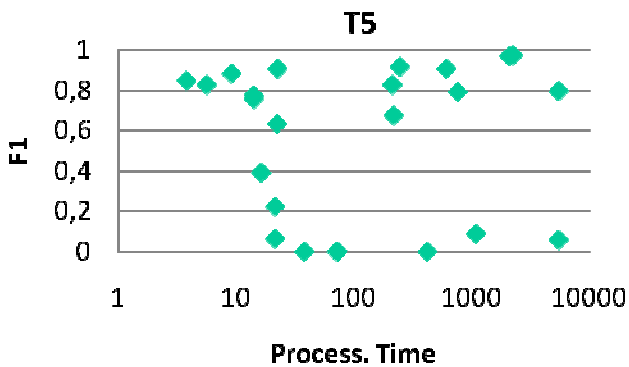
Most of the systems that are good in separating copies from non-copies (low NDCR) are also good in localization.

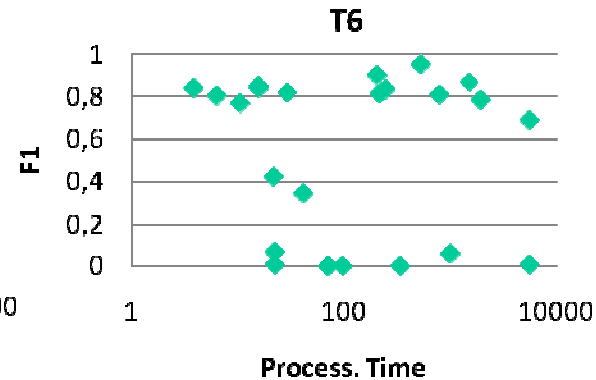
Video only – Nofa runs by transformations

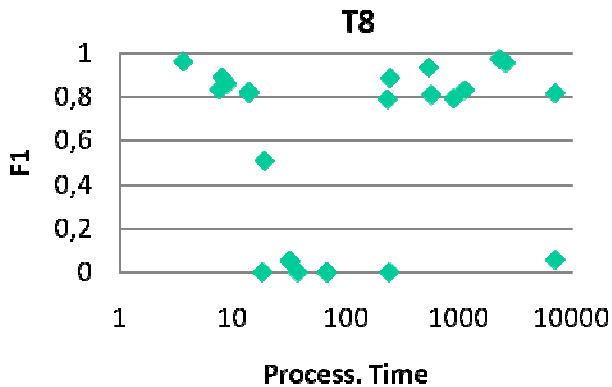


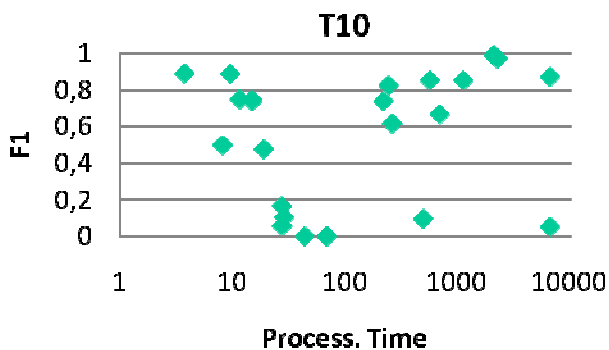






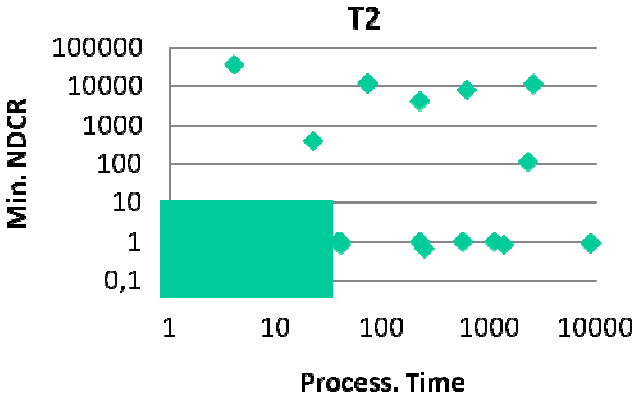


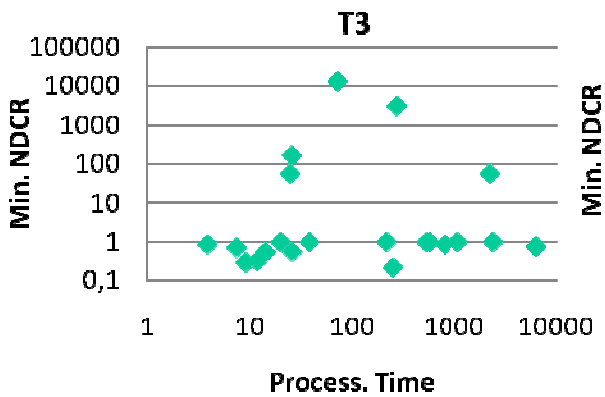


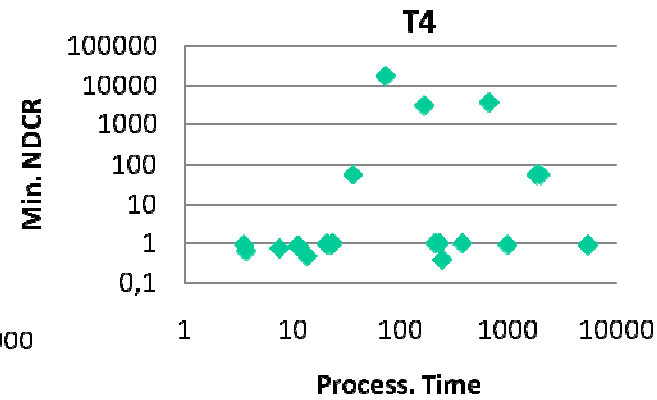


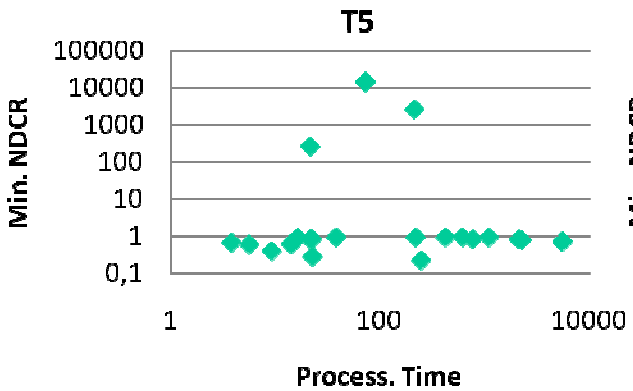
Increasing proc. time did not enhance localization. Few systems achieved high localization in small proc. time. Strong systems are efficient and precise.

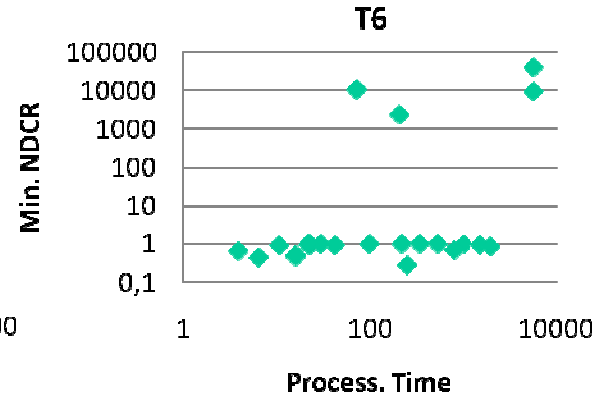
Video only – Nofa runs by transformations

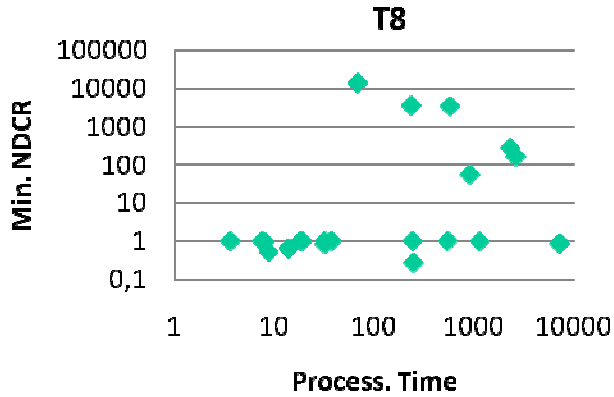


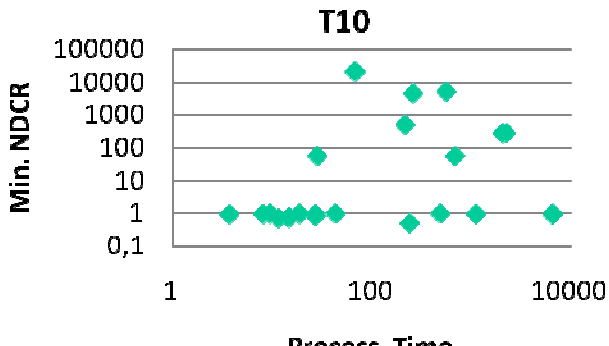






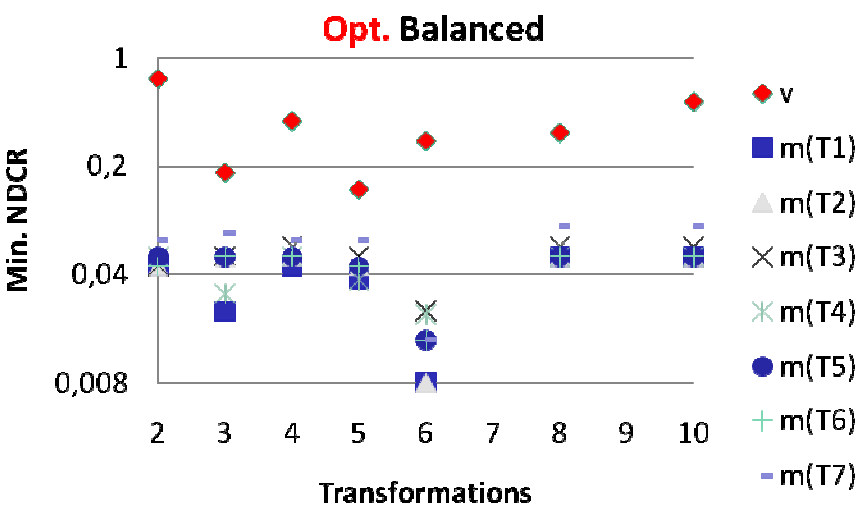


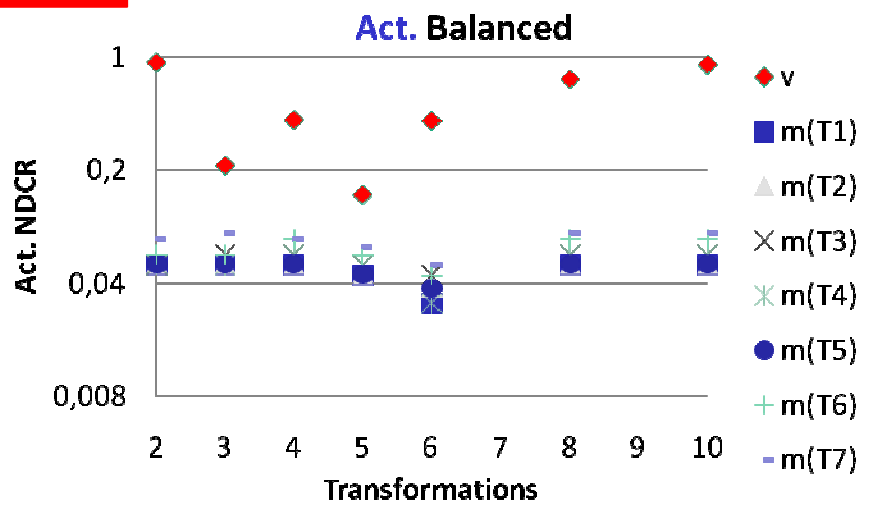


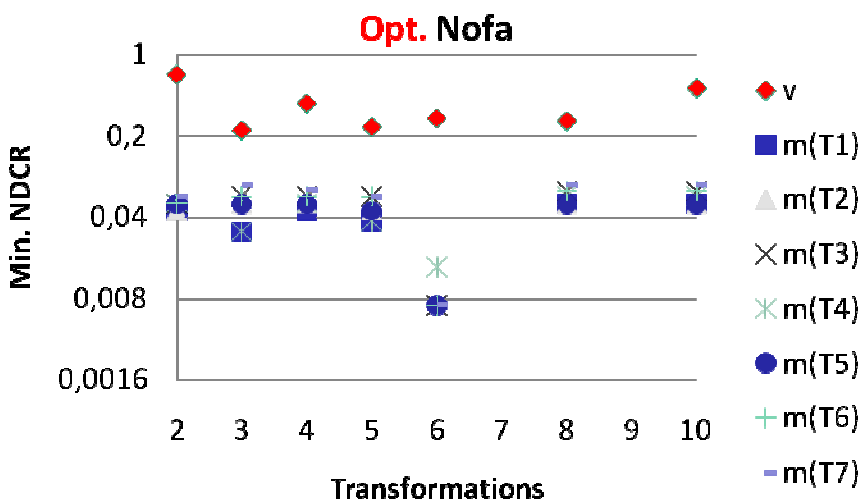


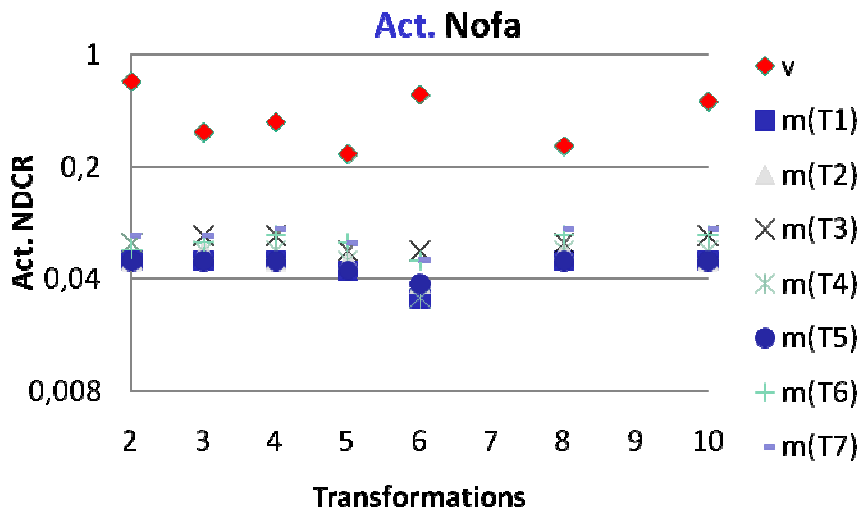
Increasing proc. time did not reduce the cost. Few good systems are fast with low cost.

Video+audio vs Video only (comparing best runs)



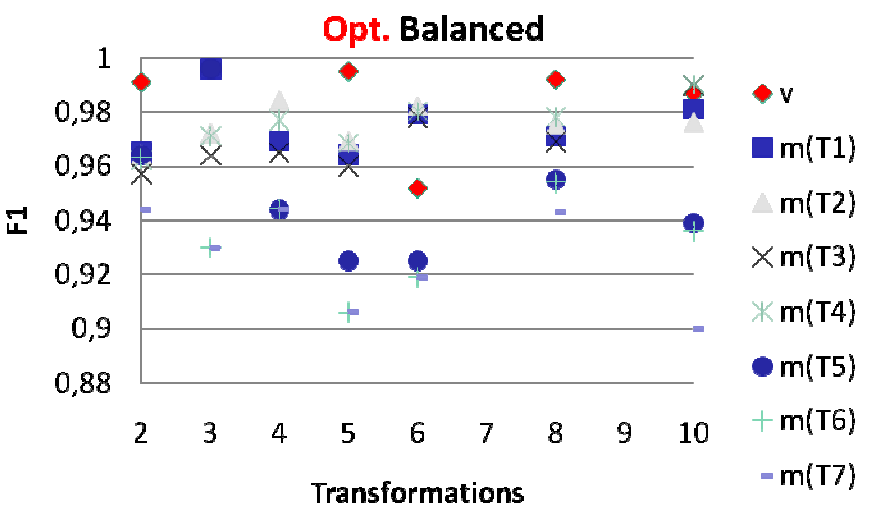


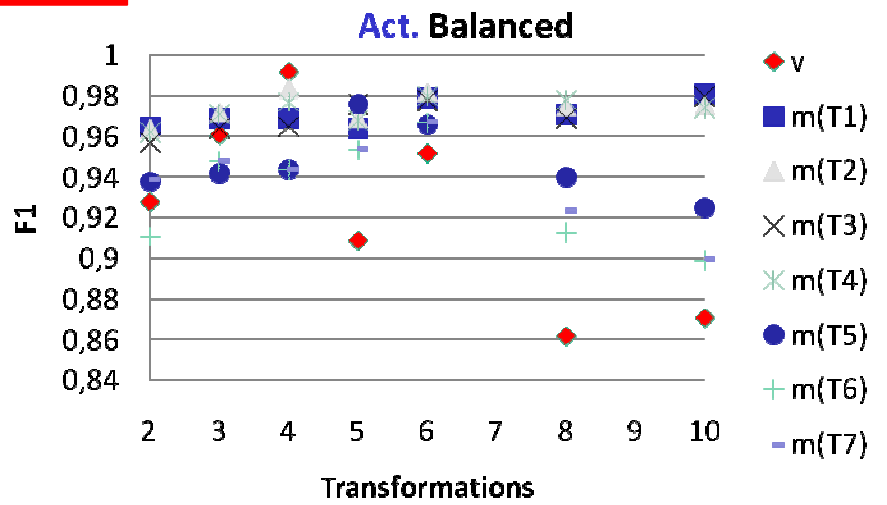


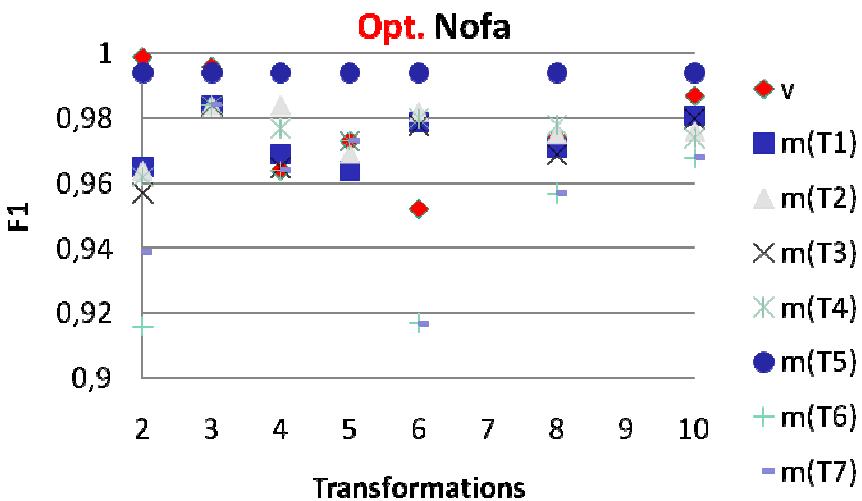


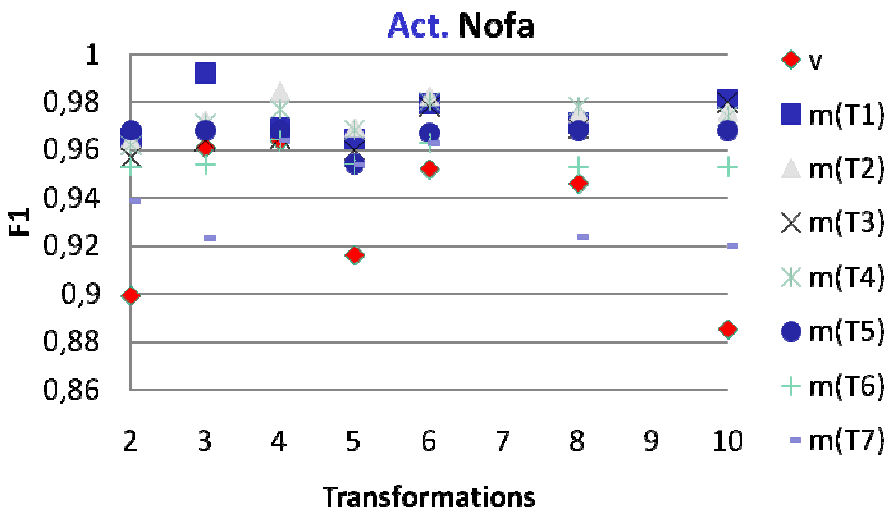
The m runs highly enhanced the detection accuracy across all transformations

Video+audio vs Video only (comparing best runs)



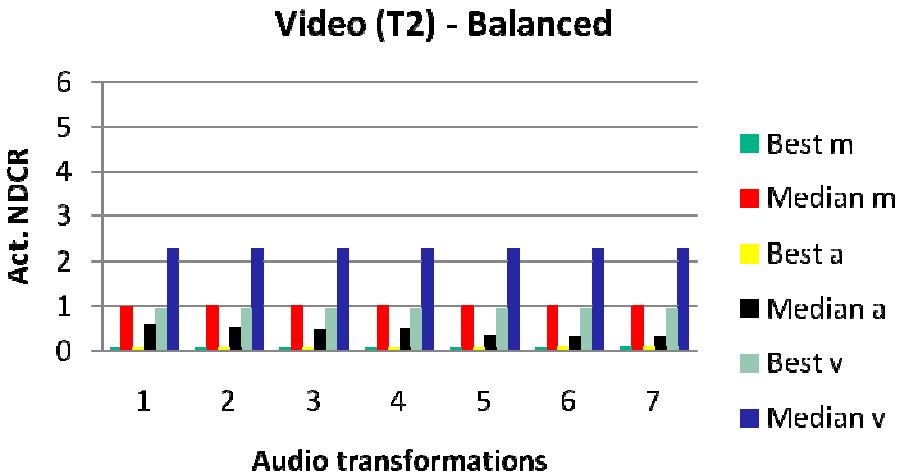


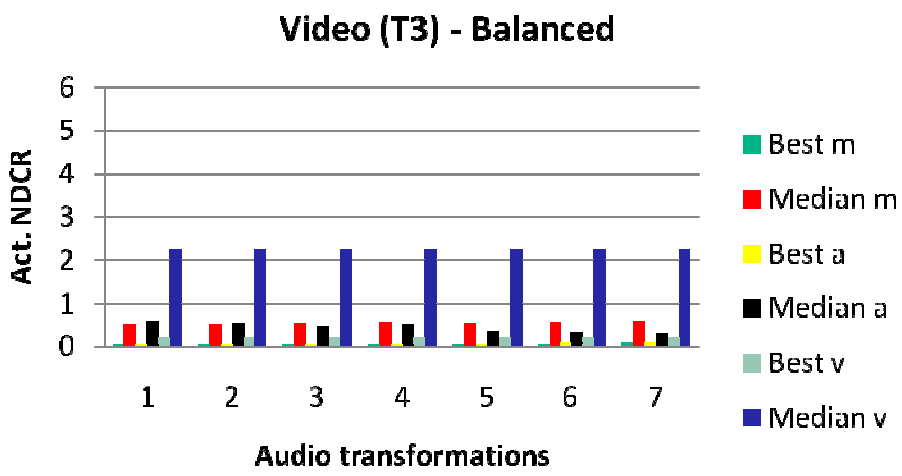


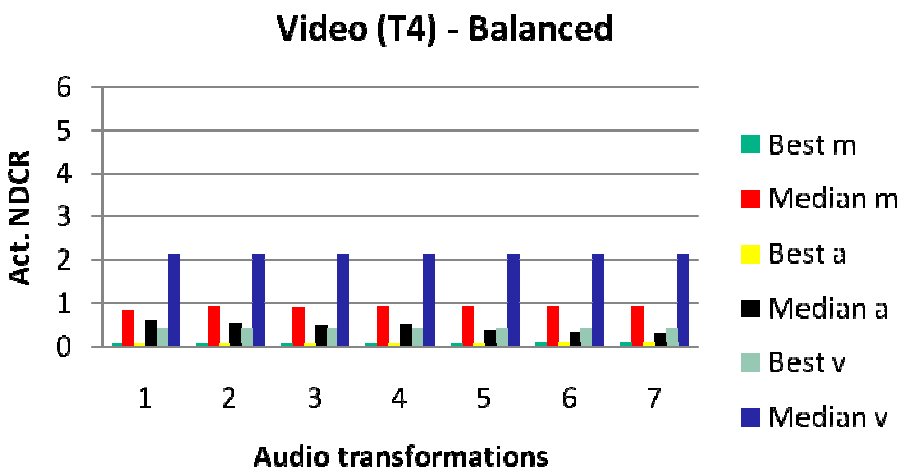


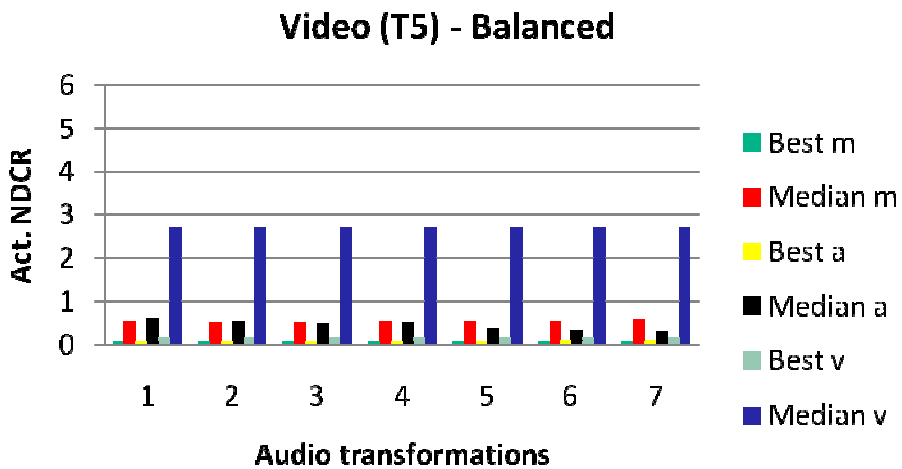
The m runs helped in the majority of transformations to enhance localization

Comparing a, v, and m best runs (Act. Balanced)

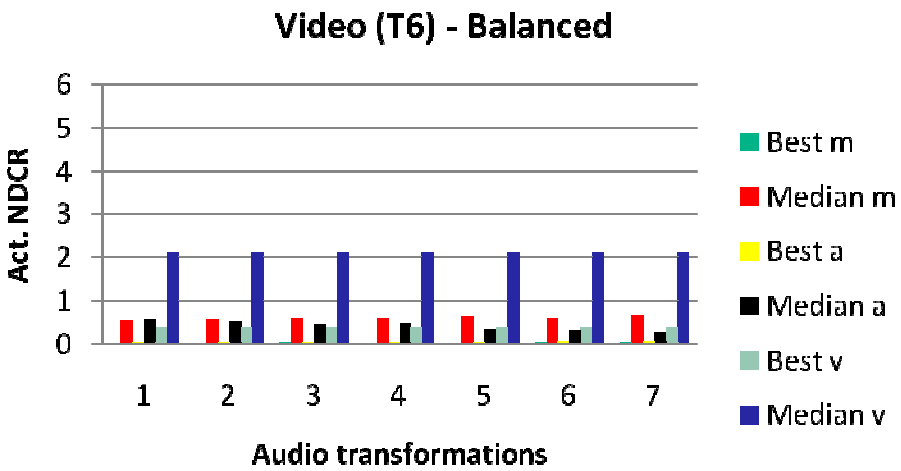


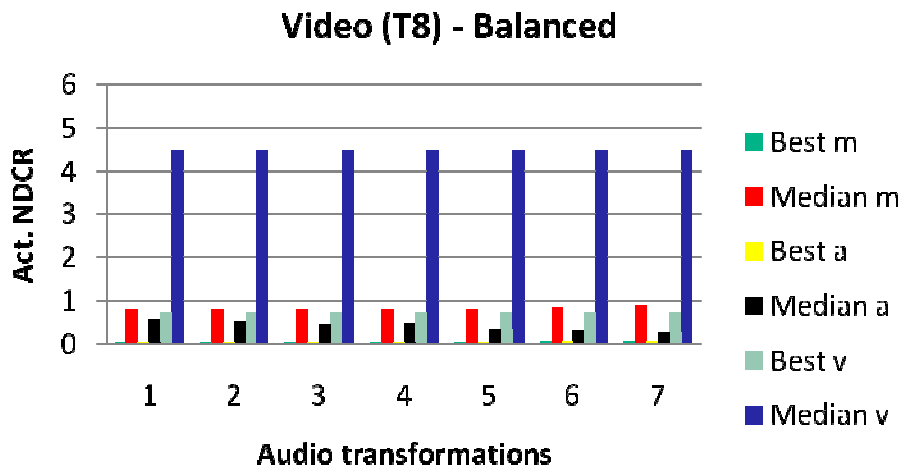


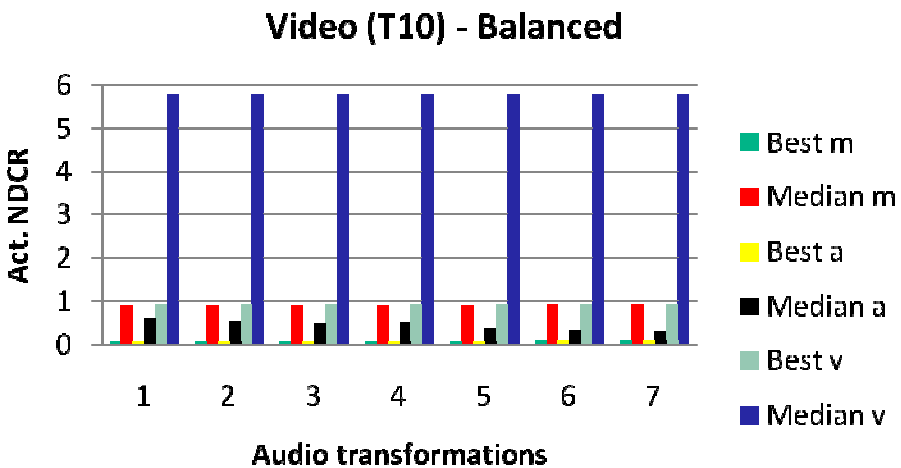




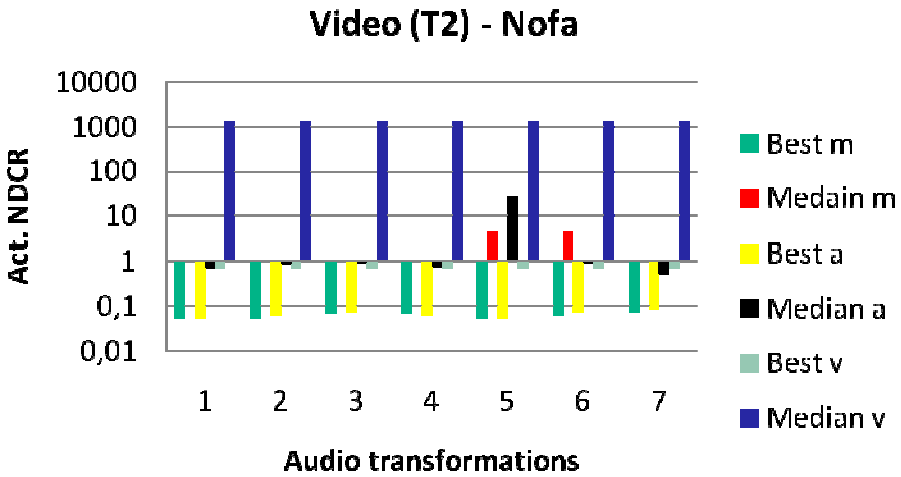
Comparing a, v, and m best runs (Act. Balanced)

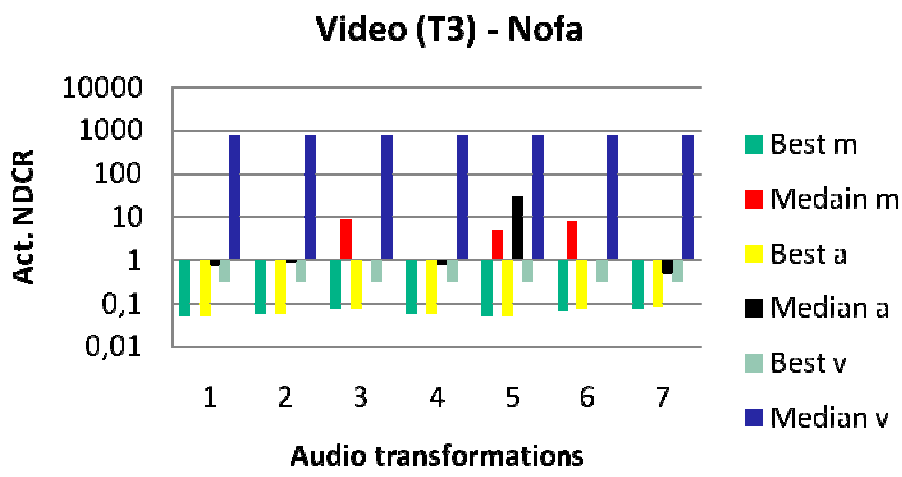


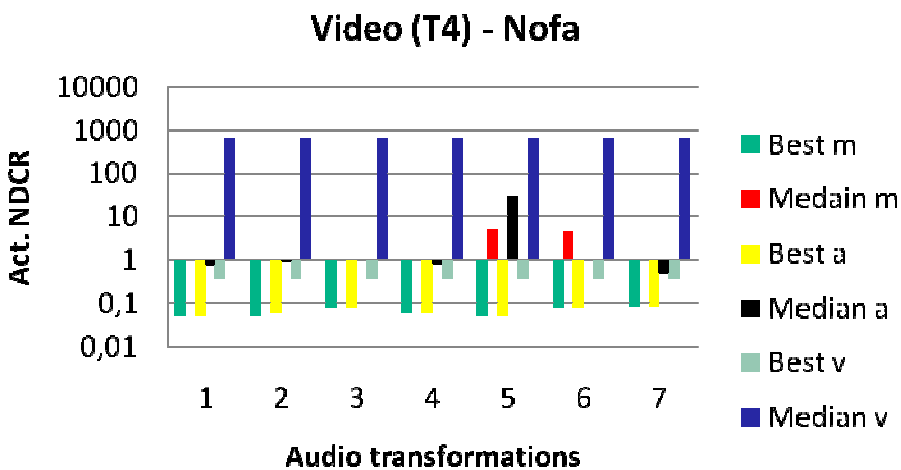


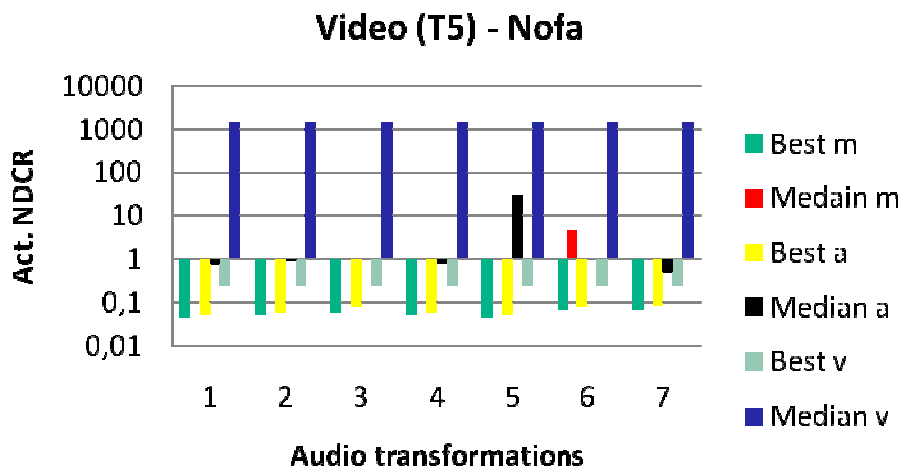


Comparing a, v, and m best runs (Act. Nofa)

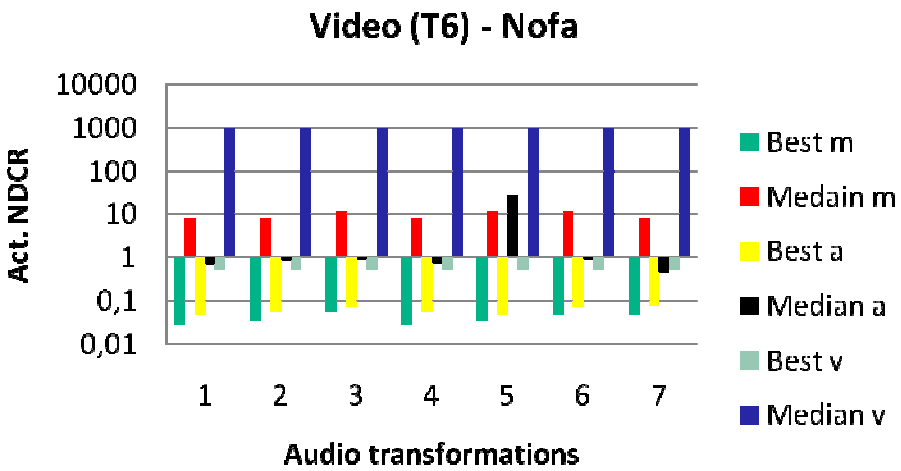


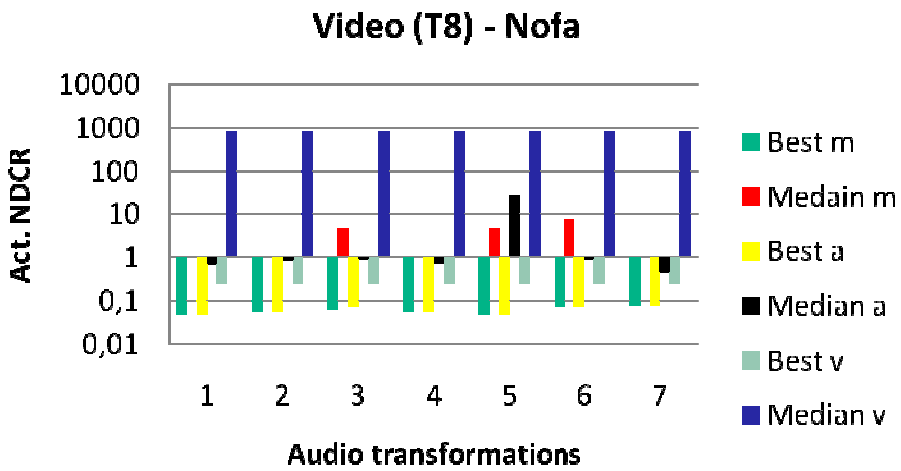


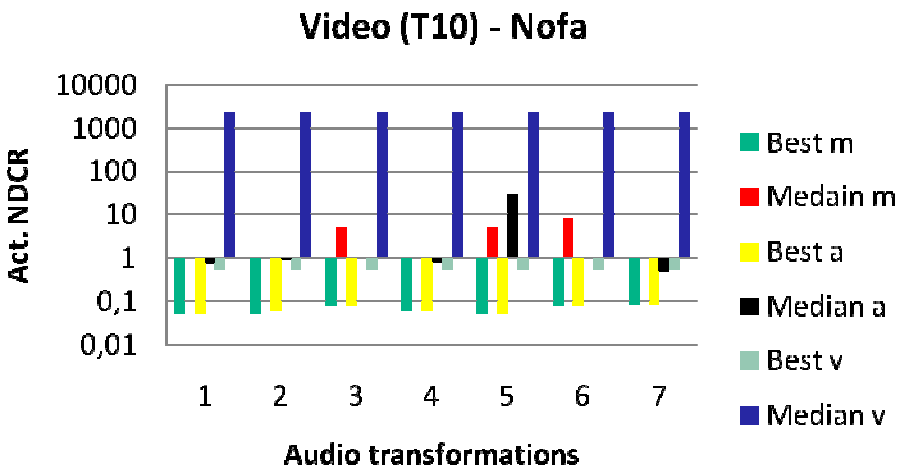




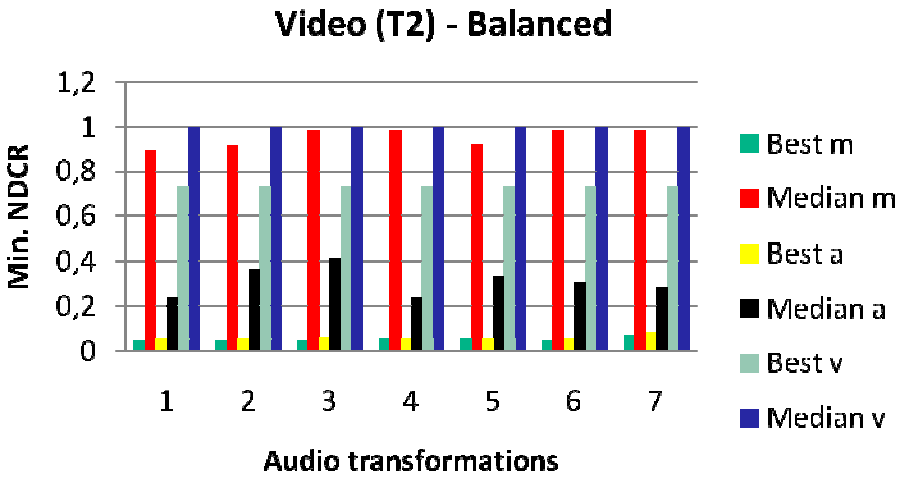
Comparing a, v, and m best runs (Act. Nofa)

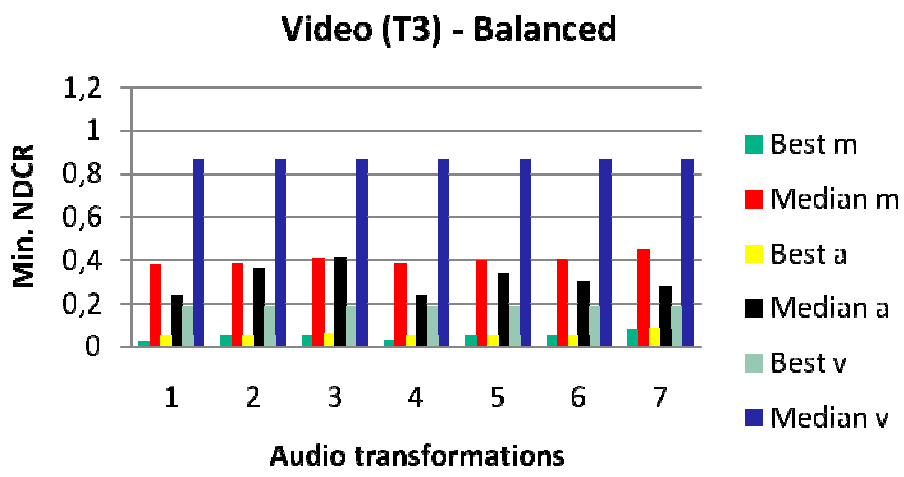


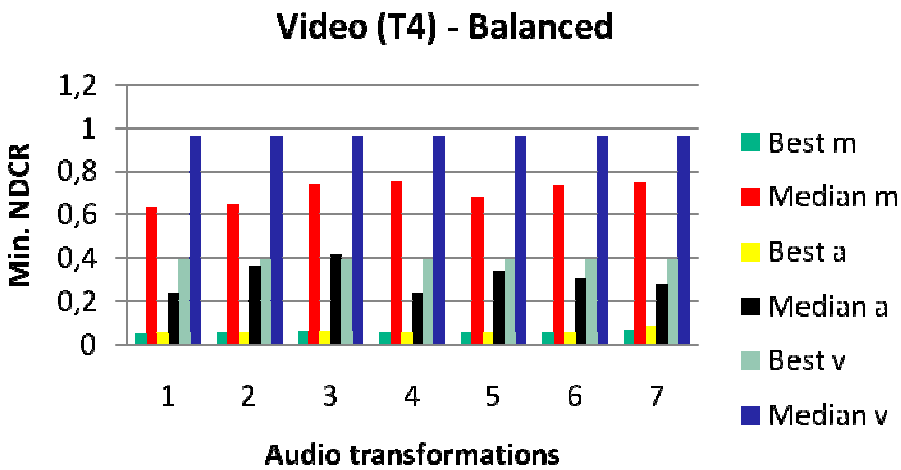


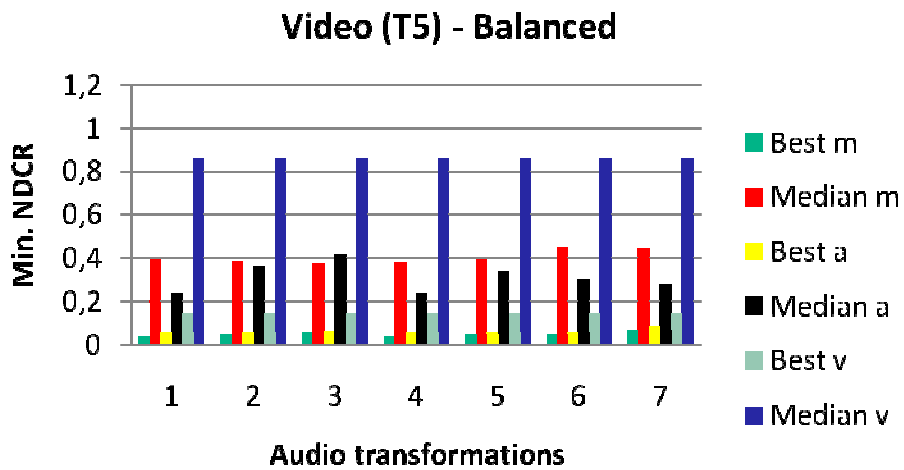


Comparing a, v, and m best runs (Opt. Balanced)

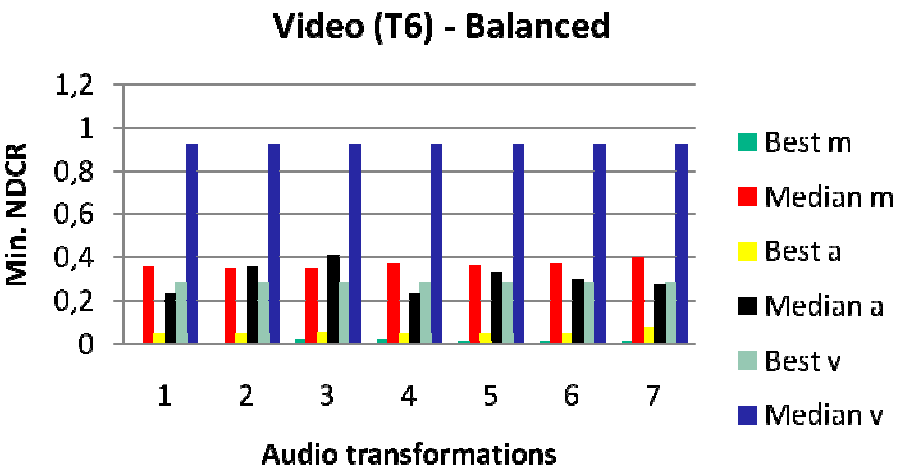


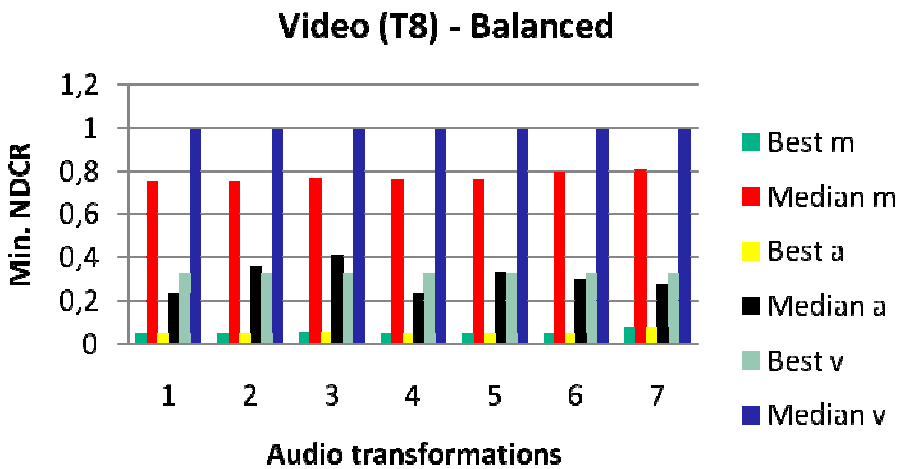


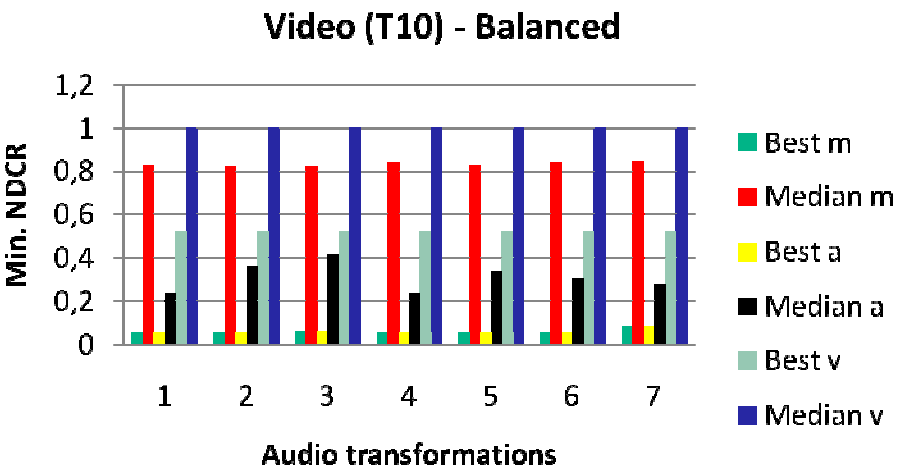




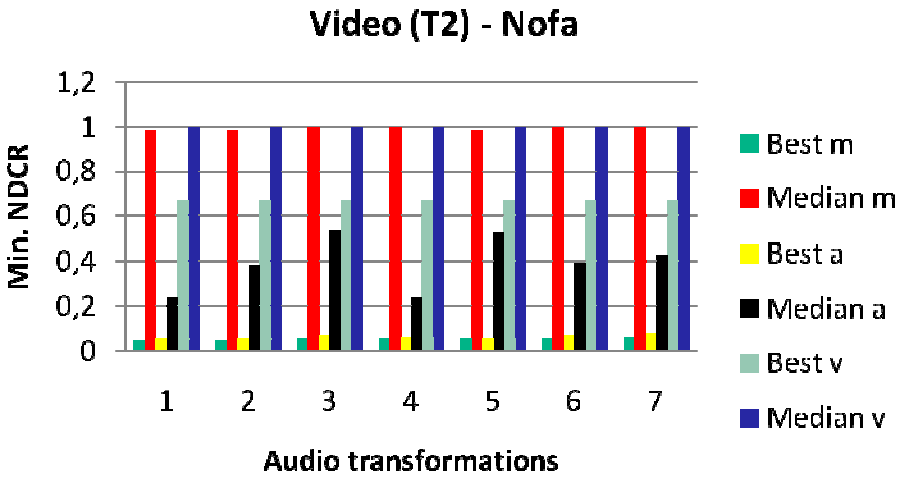
Comparing a, v, and m best runs (Opt. Balanced)

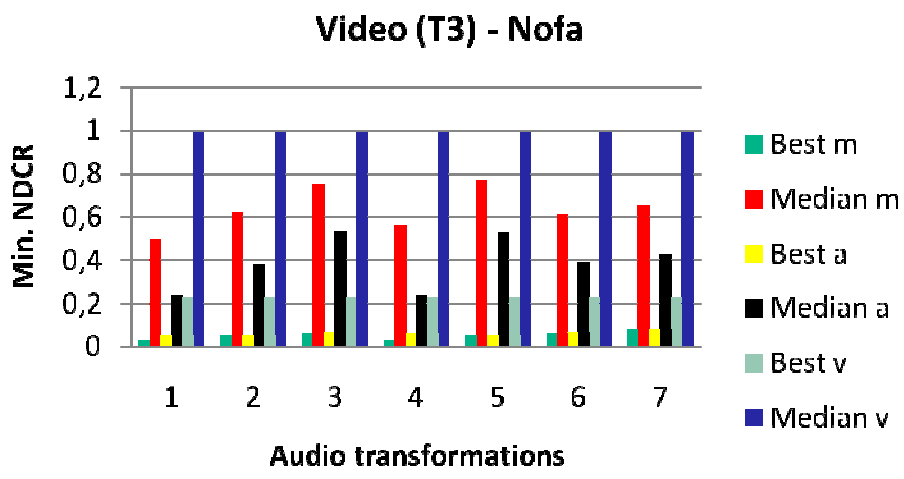


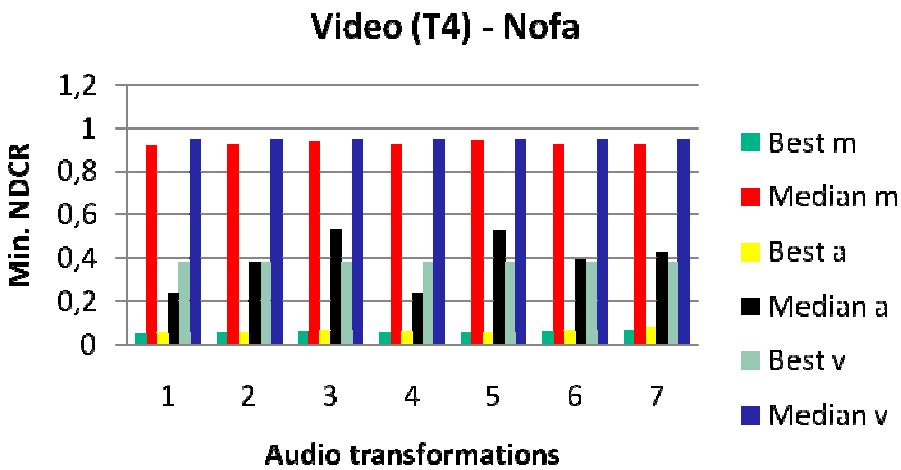


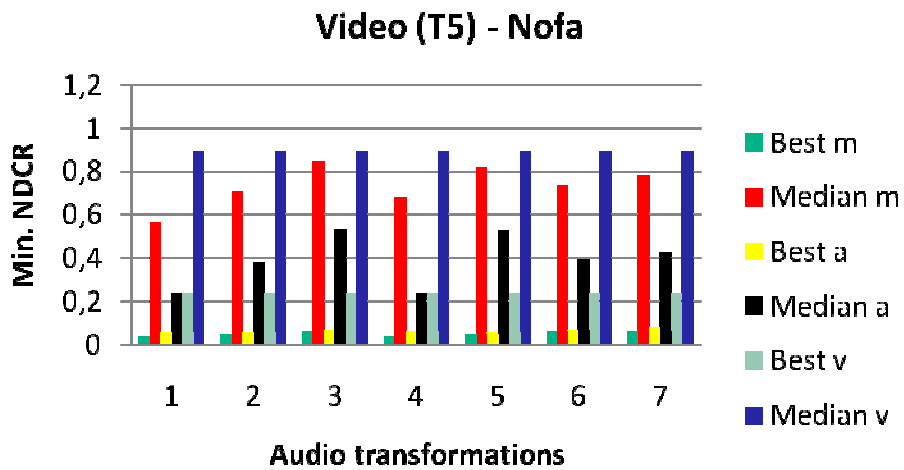


Comparing a, v, and m best runs (Opt. Nofa)

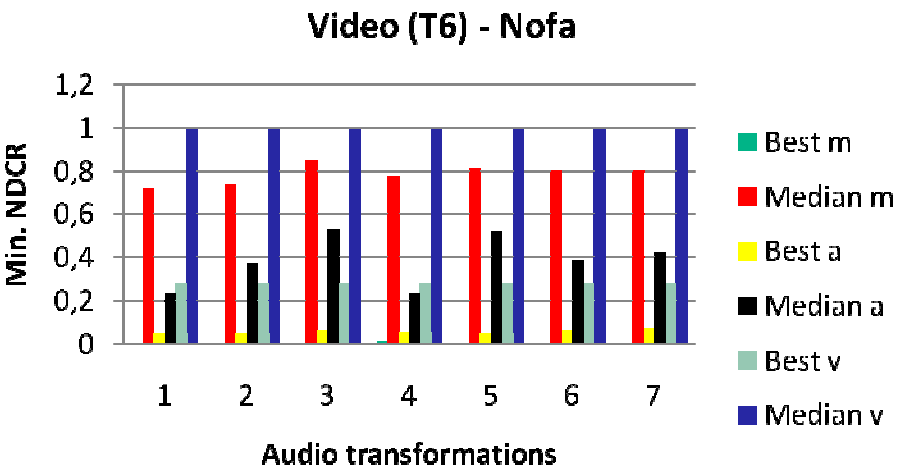


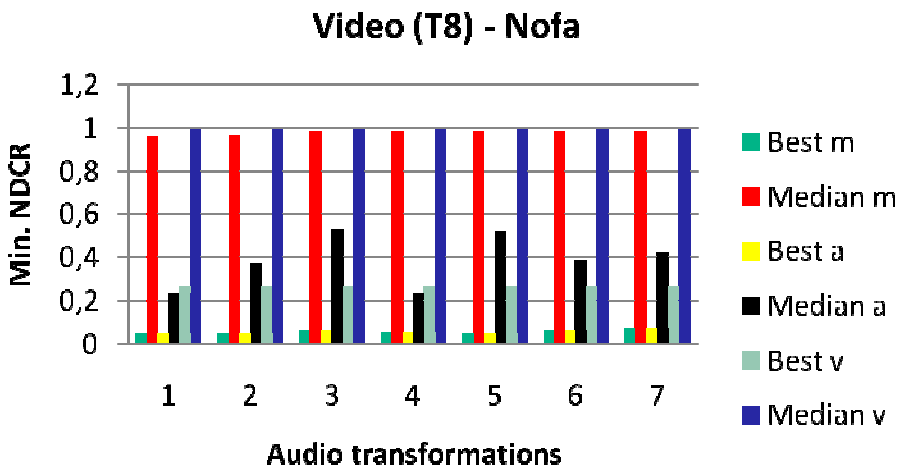


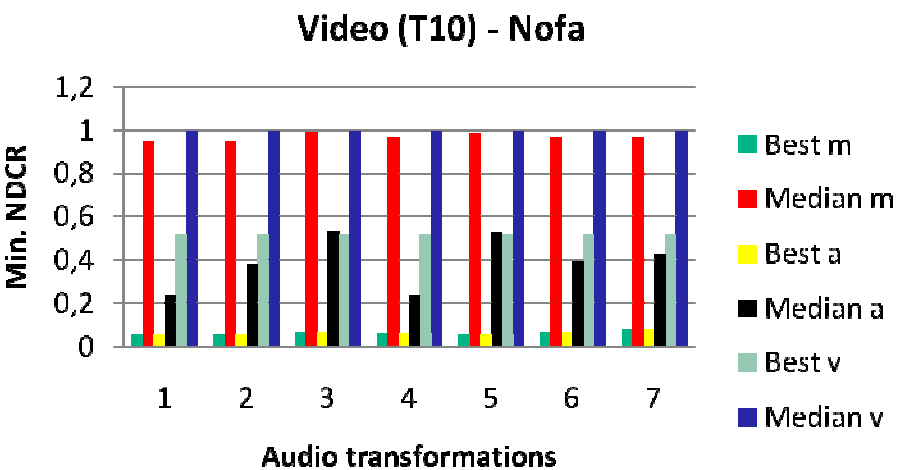




Comparing a, v, and m best runs (Opt. Nofa)







Lowest detection cost (NDCR) for individual transformations

	video only	audio only	mixed
noFA profile	MCG-ICT-CAS ATT NII IBM	CRIM TNO	CRIM
Balanced profile	MCG-ICT-CAS ATT TUBITAK	CRIM TNO	CRIM

Determining the optimal operating point

- New element for CBCD TV09, requires score normalization across queries
- For TV09 only some systems could do this
  - Large differences between actual and optimal results: big room for improvement
  - Huge impact on NDCR scores (esp. the video only runs)
  - Score normalization is critical

The influence of modalities

- Audio only detection results outperform video only
  - Easier? Techniques more mature?
- Combination of a+v improves upon a and v only
- Video only yields best localization results,
  - (still audio only systems have a higher median)
  - combination does not help
- Video only systems in general slightly faster

Comparison between noFA and balanced profiles

- tv08 discussion: teams are interested in a diversity of application profiles, noFA and balanced profiles were chosen for tv09
- Larger spread in NDCR for noFA profile (cost of a FA is high)
- noFA video only detection results slightly better than balanced

Trade-offs?

- the majority of low detection cost systems also have a good localization performance, but there is room for improvement here
- tv9 data suggests no trade-off between detection cost and speed, and between localization and speed
- Few systems perform well on all three measures

Three evaluation measures

- The cost based NDCR evaluation measure seems suitable to model a variety of application profiles
  - large potential for improvement
- The localization and performance evaluation measures can help systems to find a balance in the accuracy/size/speed trade-off
  - top systems achieve near perfect results (F1>0.95)
- Only a minority of systems performs faster than RT
  - room for improvement

Other Observations

- Complex transformations are indeed more difficult.
- Limited attraction for audio-only queries.
- Many new teams, several strong tv08 teams did not participate this year.
- Would not have been possible without major help from INRIA-IMEDIA, Laurent Joyeaux, Dan Ellis.

Some trends in tv09 within site experiments

- Fusion of distinct frame representations (fingerprints)
  - SIFT descriptors
  - Block based features
  - Global (edge histogram)
- Speed optimization
  - GPU based local feature extraction
- Transformation detection + transformation specific approaches
- Score normalization
  - Dice coefficient, sigmoid transformation
- Combination of audio and video:
  - AND or OR
  - linear combination

Impact on real-world scenarios?

- How well do these results carry over to real application scenarios?
- Is the query creation process realistic?
  - copying audio track
  - hard cuts (no gradual transitions)
  - query lengths
- Do we have accurate estimates of Rfa and Pmiss?
- How realistic are the transformations?
- Transformations a-priori known

Some suggestions for a potential tv10 task

- single application profile
- retain three measures
- rethink query creation process
  - Need data for different scenarios
- Near similar detection?

Questions

- Did any one found multiple copies for a given query?
- Can we repeat the task again in Tv2010 on IA dataset?
- Any new thoughts about application profiles? Did the balanced/nofa achieved their goals?
- Any thoughts about “near similar” detection tasks?