A Data-driven Cognitive Salience Model for Objective Perceptual Audio Quality Assessment

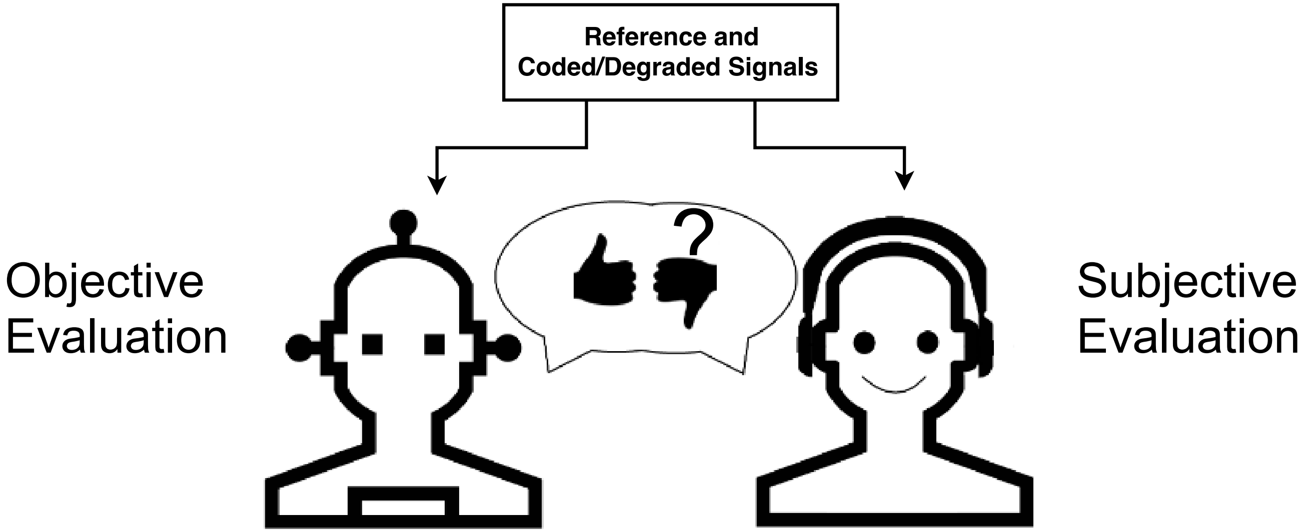
Pablo Delgado and Jürgen Herre

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Fraunhofer IIS

Motivation

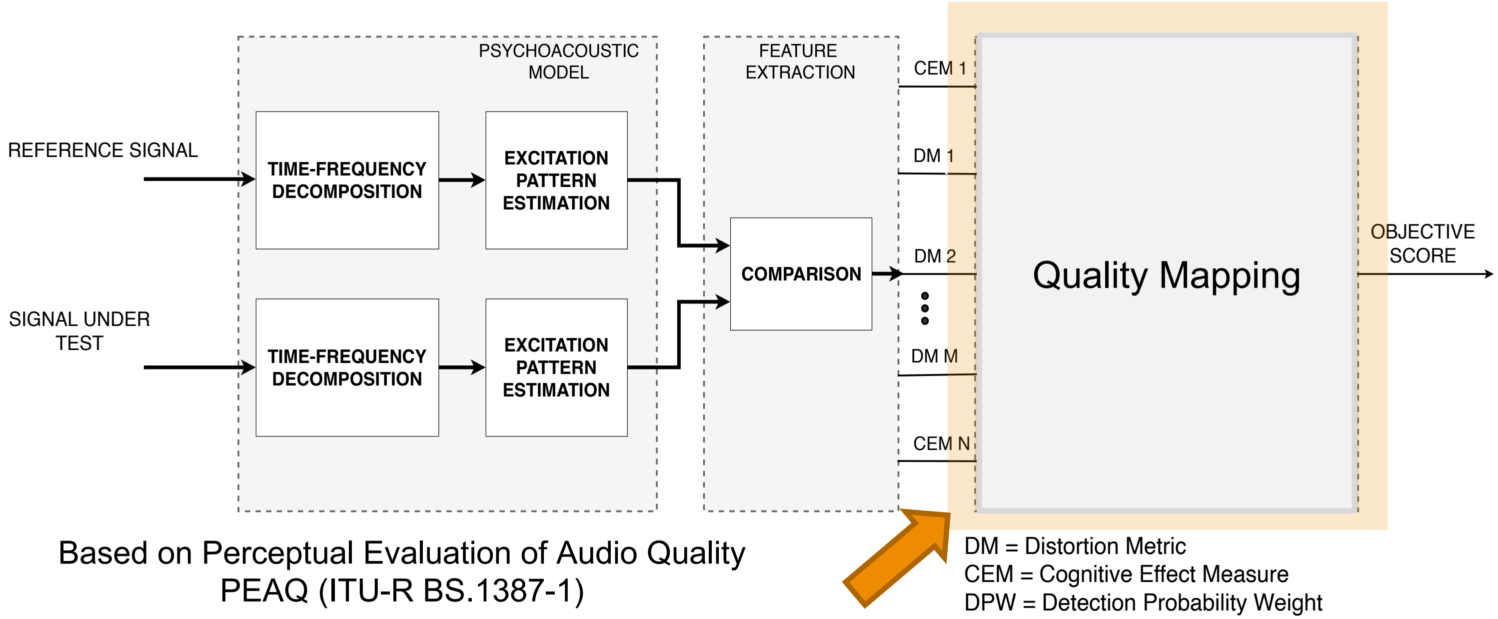
- Objective Quality Assessment Systems (OQAS) analyze signals to predict perceived quality degradation as reported by subjects (i.e., the subjective quality) on a listening test:
  - Can be used for audio codec selection, real-time monitoring, etc..
  - They save time and resources (as alternatives to listening tests)
  - Mostly based on a model of human perception/psychometric findings.



© AudioLabs, 2022

Motivation

- OQAS Architecture



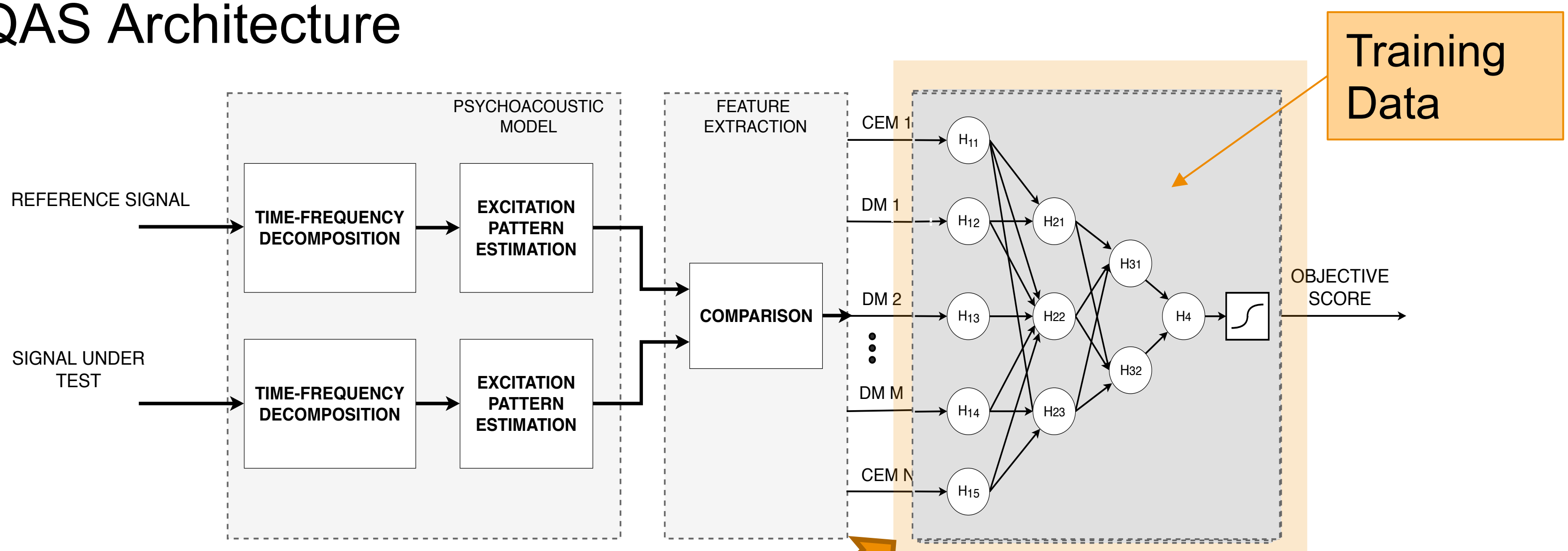
Based on Perceptual Evaluation of Audio Quality
PEAQ (ITU-R BS.1387-1)

DM = Distortion Metric
CEM = Cognitive Effect Measure
DPW = Detection Probability Weight

- Metric-to-quality mapping stage
  - Considered a model of **auditory cognition** (beyond-peripheral processes)
  - **Weighted** combination of different metrics into a single quality score
  - Weights reflect the **importance** of each metric **in describing quality degradation**.

Motivation

- OQAS Architecture



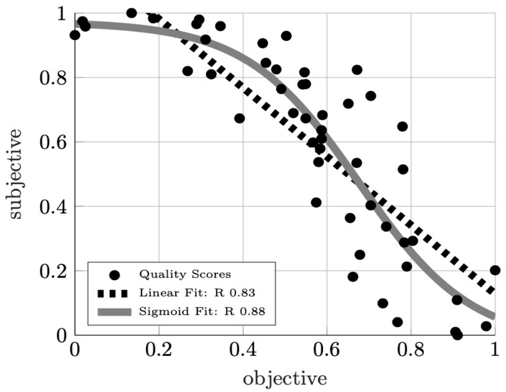
Based on PEAQ (ITU-R BS.1387-1)

DM = Distortion Metric
CEM = Cognitive Effect Measure
DPW = Detection Probability Weight

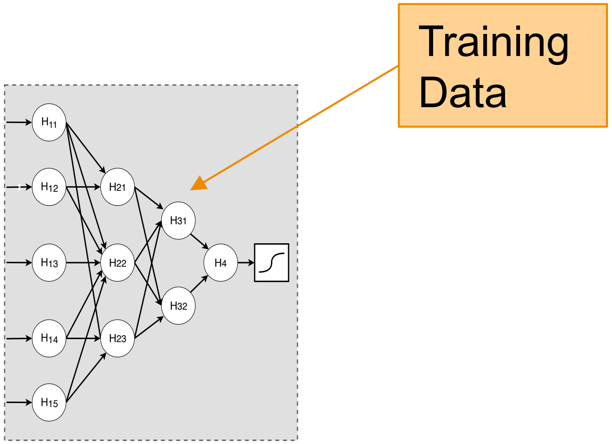
- Metric-to-quality mapping stage
  - Usually implemented as a **multivariate statistical learning model** (Linear Regression, Splines, SVM, ANN and others…)
  - The mapping function that links metric values to overall quality score, using **subjective listening test data as target**.

Motivation

- The learning algorithm tasks:
  1. To estimate DM-to-quality (nonlinear) mapping functions (due **peripheral effects**: threshold and compression effects, loudness perception, artefact detection, etc..)



  2. To model interactions between features (**cognitive effects**: a distortion's **perceived severity** depends on the **strength** of other competing distortions.)
     - Mapping function **gradients change** according to the values of the input vector on a multidimensional space.



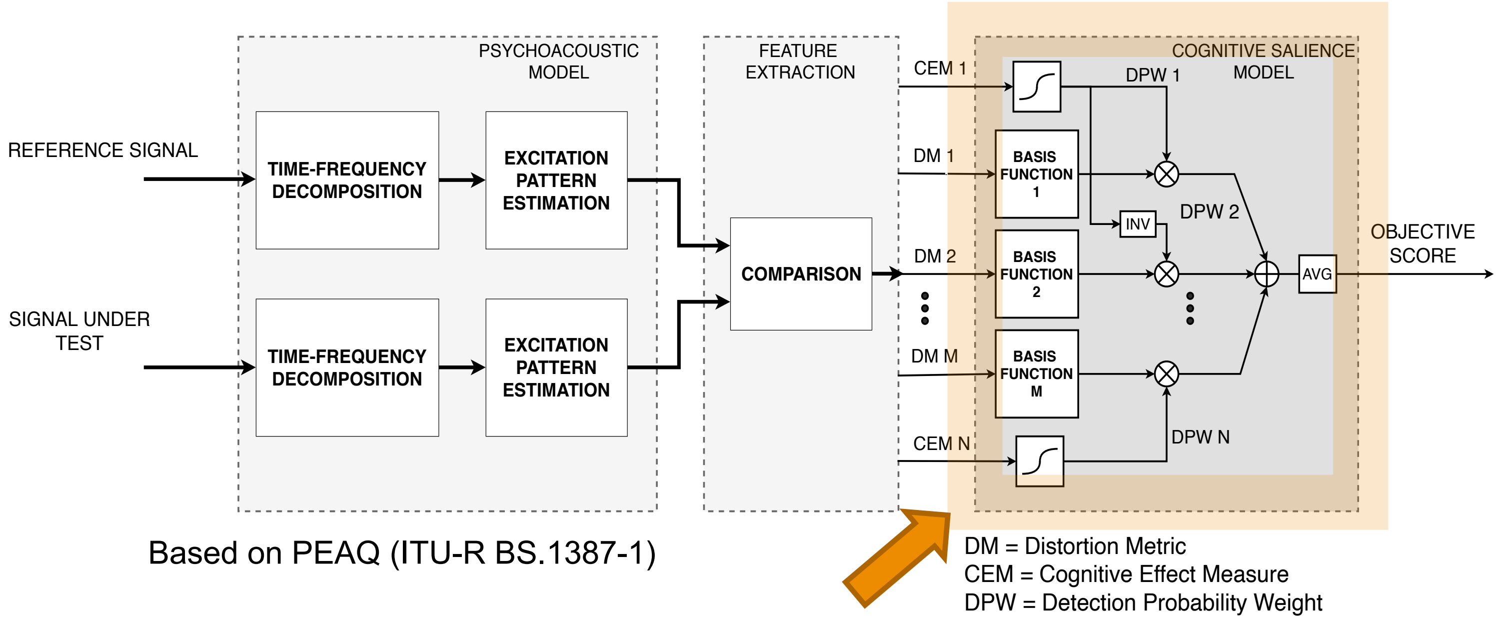
Reliable listening test data is usually expensive and at most, a couple hundred data points are available at a time.

These tasks take need many free model parameters to estimate with **scarce** data.

© AudioLabs, 2022

Method: Data-Driven Cognitive Salience Model

- Cognitive Salience Model (CSM) as quality mapping stage:



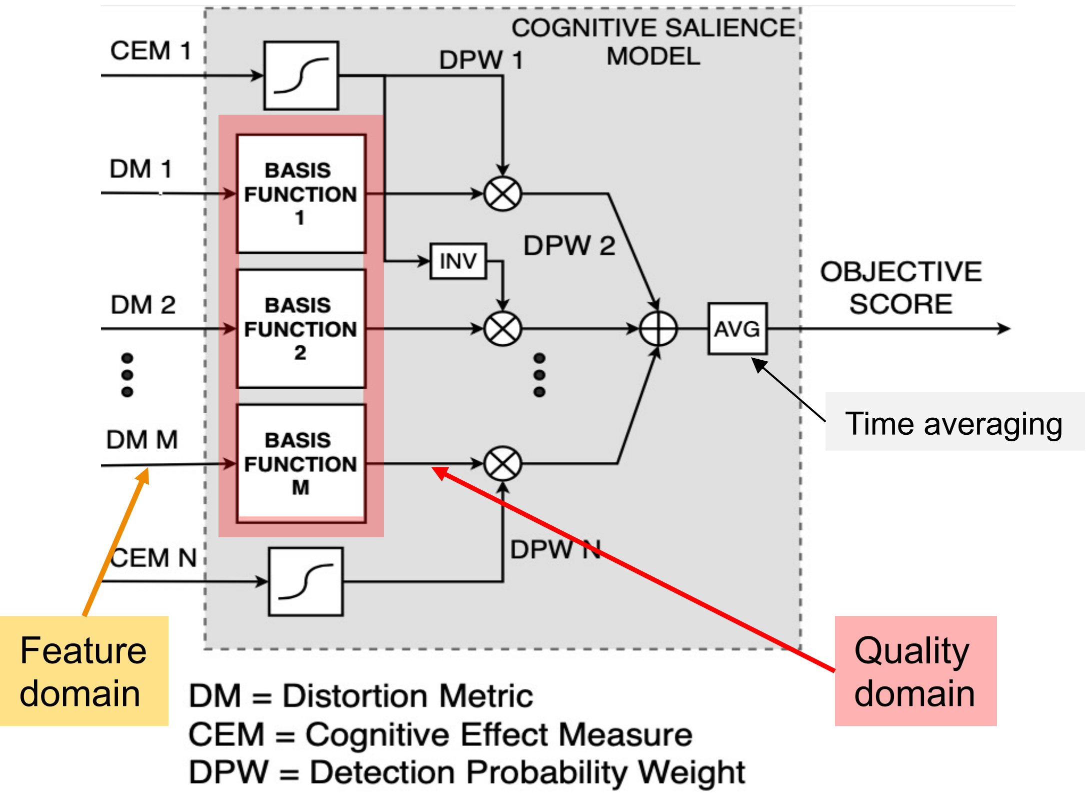
Based on PEAQ (ITU-R BS.1387-1)

DM = Distortion Metric
CEM = Cognitive Effect Measure
DPW = Detection Probability Weight

- Perceptually-motivated architecture:
  1. Pre-mapped DM-to-quality mapping Basis Functions
  2. Limits in the number of **feature interactions** using the concept of **distortion salience**

Method: Data-Driven Cognitive Salience Model

- DM-to-quality mapping

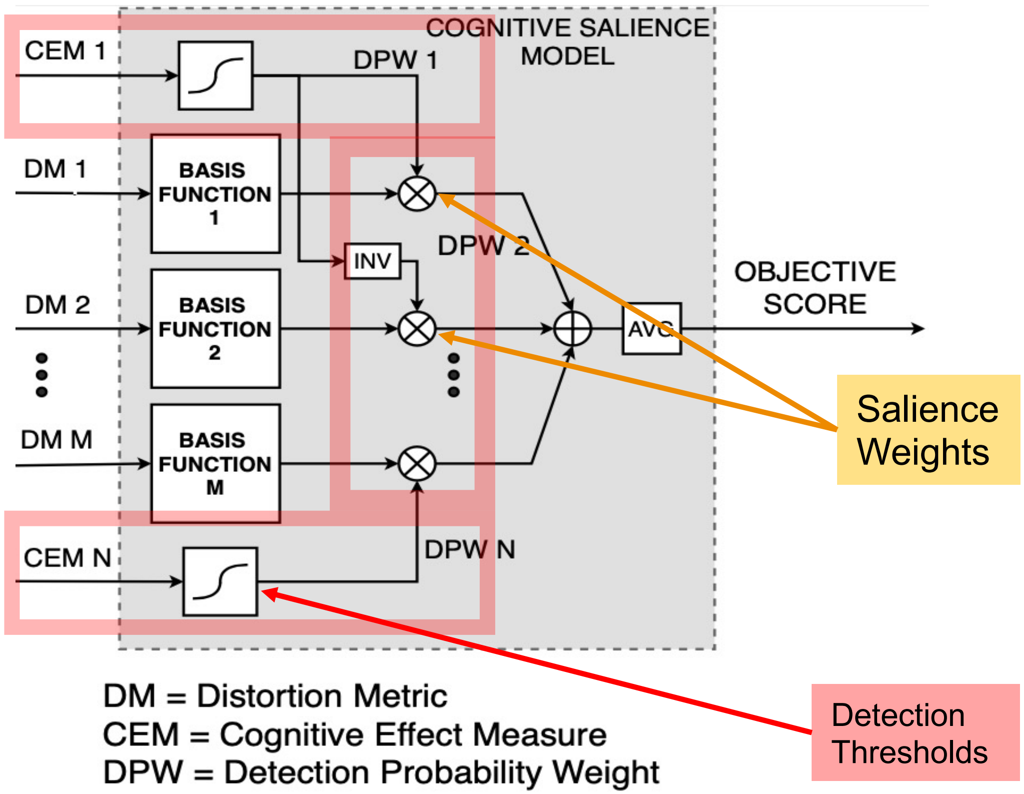


- Basis Functions (BF)
  - Estimated **separately** for **each of the DM** using a **LT database** on signals with isolated audio coding artefacts [*].
  - MUSHRA-like Method (ITU-R BS.1534)
  - The **isolated audio coding artefacts** minimize distortion metric **interactions** → favor BF independence
  - Estimation Method: Multivariate Adaptive Regressive Splines (MARS)

* Dick et al. "*Generation and Evaluation of Isolated Audio Coding Artifacts*" Audio Engineering Society Convention 143, Oct. 2017

Method: Data-Driven Cognitive Salience Model

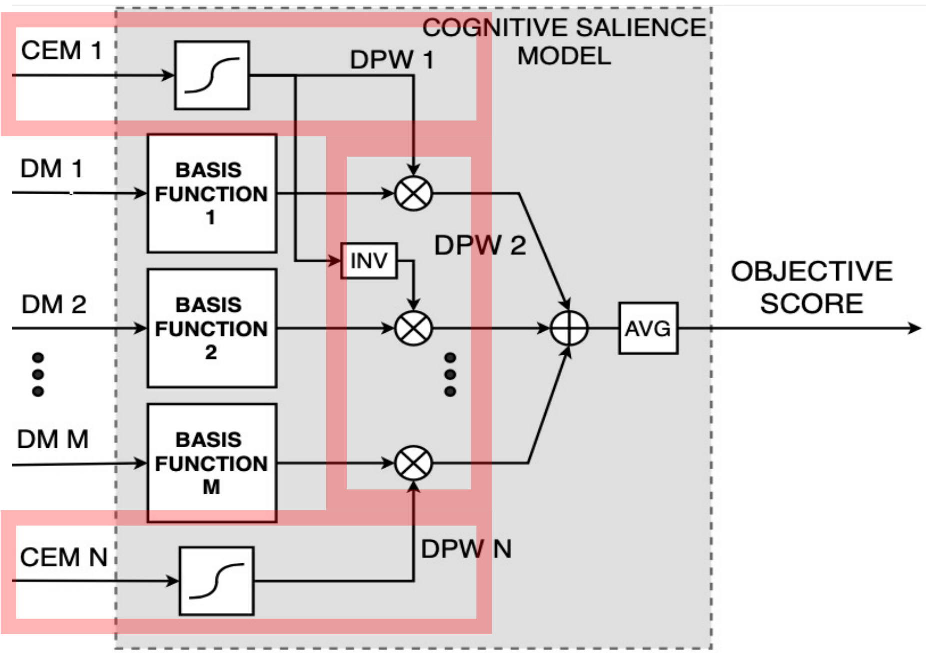
- Interactions Model



- Interactions
  - Cognitive effects can **predict distortion salience** → dynamically **weight** DM importance in overall quality according to CEM values
  - Model cognitive effect detection and saturation thresholds using [0,1] sigmoids (DPW)

Method: Data-Driven Cognitive Salience Model

- Data-driven model selection procedure



DM = Distortion Metric
CEM = Cognitive Effect Measure
DPW = Detection Probability Weight

- **Meaningful** CEM-DM interactions are selected based on the values of an **interaction metric** on an **LT database**

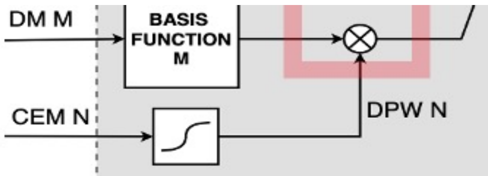
- Two sigmoid parameters per meaningful interaction (crossover point and slope) **are fitted** to **optimize** the values of the **interaction metric**

© AudioLabs, 2022

Method: Data-Driven Cognitive Salience Model

- Data-driven model selection
  - **Interaction** metric for M (**correlation** between **CEM** and **DM salience**):

$$\mathcal{C}_{m,n} = \left| \frac{\sum_{j=1}^{J}(\mathcal{S}_m(j) - \overline{\mathcal{S}}_m)(DPW_{m,n}(j) - \overline{DPW}_{m,n})}{\sqrt{\sum_{j=1}^{J}(\mathcal{S}_m(j) - \overline{\mathcal{S}}_m)^2}\sqrt{\sum_{j=1}^{J}(DPW_{m,n}(j) - \overline{DPW}_{m,n})^2} \right|$$



  - Where S_m is a **salience** metric defined as:

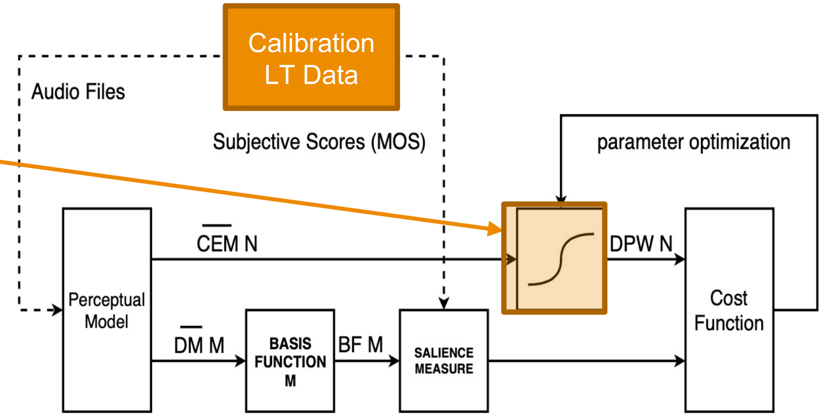
$$\mathcal{S}_m(j) = \frac{\sum_{i=1}^{I}(y_{ij} - \overline{y}_j)(BF_{mij} - \overline{BF}_{mj})}{\sqrt{\sum_{i=1}^{I}(y_{ij} - \overline{y}_j)^2}\sqrt{\sum_{i=1}^{I}(BF_{mij} - \overline{BF}_{mj})^2}}$$

Per each input signal j

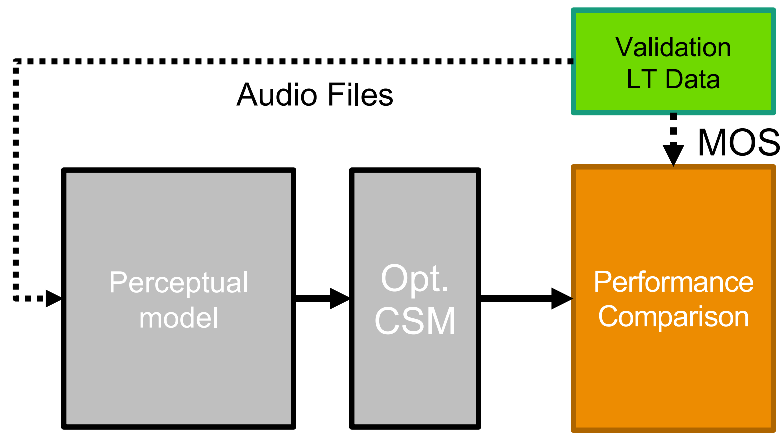
That measures **correlation** between y_{ij}, (the MOS of signal j over all treatments i) and the respective BF_m output (in the quality domain) of the DM basis function.

Experiment: Model Selection, Optimization, Validation

- Two disjoint sets will be used
  - The **Calibration/Optimization Dataset**
    - CEM/DM selection and DPW optimization
    - **7** condition/treatments
    - **168** MOS data points.
  - The **Validation Dataset** *independent* LT data on which the proposed system will be evaluated
    - **9** conditions/treatments (not in the optimization dataset)
    - **216** MOS data points.



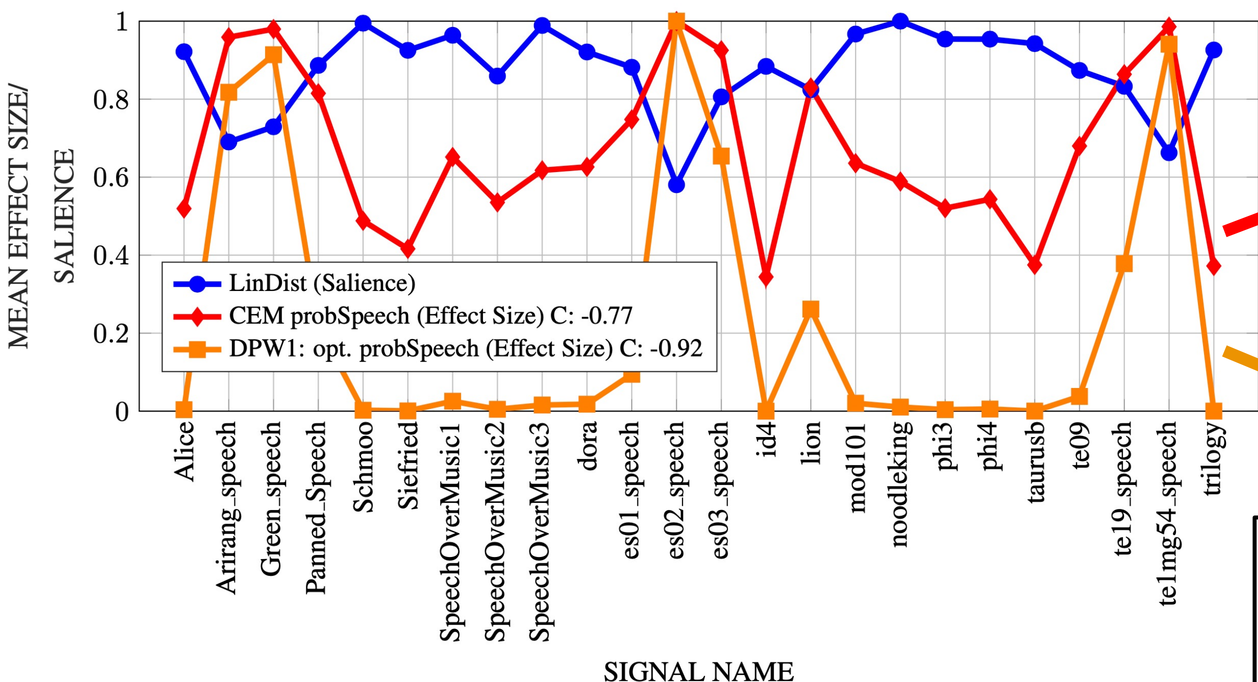
- Subjective LT Database[*].
  - **24 signals** (music, speech and mixed content), 3 codecs, bitrates: **16 to 96 kbps**, > **25000** individual subj. scores pooled into MOS (MUSHRA).

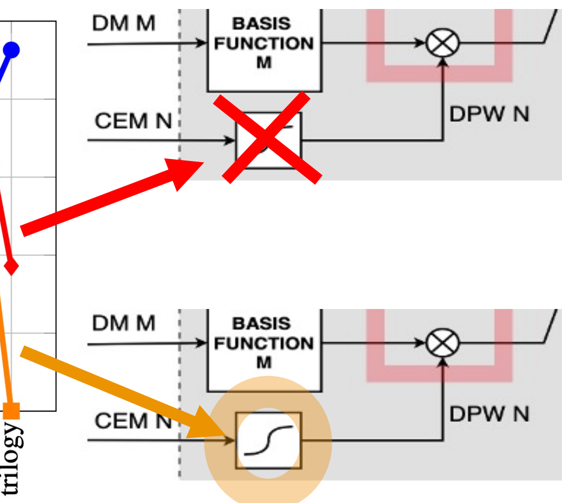


* Universal Speech and Audio Coding (USAC) Verification Tests "USAC Verification Test Report N12232" ISO ISO/IEC JTC1/SC29/WG11 2011

Results: Model Selection

- Interaction optimization example
  - Linear distortion (DM) **salience** (blue line) is **lower** when the probability of the signal being speech-like (CEM, red line) is **higher**.
  - The DPWs **threshold the CEM** values (orange line) through a **sigmoid** function with two **optimized** parameters.





Optimized sigmoid parameters increase the interaction metric value

Results: Model Selection

- Selected Interactions and C_m values (before and after optimization)

Weight	CEM	Target DM	C (CEM/DPW)	Equation
DPW1	probSpeech	LinDist	-0.77 / -0.92	DPW1 = 1- probSpeech_th_lin
DPW2	probSpeech	NoiseLoudness	0.67 / 0.80	DPW2 = probSpeech_th_nl
DPW3	probSpeech	MissingComponents	-0.20 / -0.37	DPW3 = 1-DPW2
DPW4	EPN	LinDist	-0.40 / -0.70	DPW4 = 1-EPN_th_lin
DPW5	EPN	SegmentalNMR	0.1 / 0.25	DPW5 = (EPN_th_sgm)(1-PDEV_th_sgm)
DPW5	PDEV	SegmentalNMR	-0.18 / -0.21	-

- The weights of the distortion metrics (DPW) depend on the values of cognitive effect size metrics:

Target DMs: PEAQ Advanced MOVs (ITU-R BS.1387-1)
EPN: amount of disturbance perceptual streaming (PS) [*]
PDEV: informational masking of disturbances (IM) [*]
probSpeech: probability of signal being speech-like [**]

- Negative interaction metric values denote **decreasing salience** with **increasing effect size**
- DPW3 selected despite lower C values, because it is complementary to DPW3
- DPW5 was combined based on PS/IM complementary relationship reported in [*]

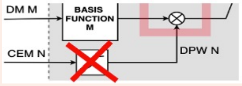
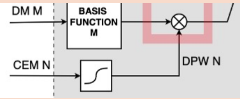
* J. Beerends "The Role of Informational Masking and Perceptual Streaming in the Measurement of Music Codec Quality" Audio Engineering Society Convention 100, May 1996
** G. Fuchs "A robust speech/music discriminator for switched audio coding" EUSIPCO, Sept. 2015

Results: Validation

- System Validation Performance Metrics (on unseen data)

System	R	RMSE*
ViSQOL NSIM	0.82	5.6
PEAQ DI	0.69	8.1
DM + CEM	0.84	5.1
PROPOSED	0.86	4.6
PROPOSED (Opt.)	**0.90**	**3.7**

- R: objective versus subjective score correlation
- RMSE*: error of predictions outside of confidence interval (ITU-R P.1401)

System Name	Inputs	Quality Mapping	Training Data
ViSQOL NSIM (Hines et al.)	NSIM	3rd order polynomial (ITU-T P.1401)	-
PEAQ DI (ITU-R BS.1387)	PEAQ Advanced MOV	ANN in ITU-R BS.1387	ITU DBs listed in ITU-R BS.1387
DM+CEM	Proposed MOVs and CEMs	ANN with similar settings in ITU-R BS.1387. Approach inspired in [*]	Isolated Artefacts + Optimization DB
PROPOSED	Proposed MOVs and CEMs	CSM (No DPW) 	Isolated Artefacts (DM-to-quality) + Optimization DB (Interaction Selection)
PROPOSED (Opt)	Proposed MOVs and CEMs	CSM (Optimized DPW) 	Isolated Artefacts (DM-to-quality) + Optimization DB (Interaction Selection and DPW optimization)

* Barbedo et al. "A New Cognitive Model for Objective Assessment of Audio Quality" J. Audio Eng. Soc. (vol. 53 p. 22-31), 2005

© AudioLabs, 2022

Summary and Conclusions

- We proposed a Cognitive Salience Model (CSM) as a feature-to-quality mapping stage that explicitly models interactions of cognitive effects and distortion metric saliences in quality perception

- On a diverse set of unseen validation data, two systems using the CSM outperformed a system with a general-purpose ANN mapping stage, with the same input features and training data.

- The CSM systems also outperformed two state-of-the-art quality measurement systems.

Future Work

- Improve model selection criteria:
  - This study: based on strong values of the interaction metric. However, combined interactions improved performance despite relatively weak interaction metric values.
- Further validation on data
  - More listening test data to validate the model
  - More diverse signal degradations stemming from other applications
- Consider other beyond-peripheral effects as predictors of distortion salience in the CSM (e.g., release of masking through co-modulation)
- Stereo/spatial audio: consider interactions between cognitive effects and perceived spatial image distortion metrics

© AudioLabs, 2022

Thank you for your time!

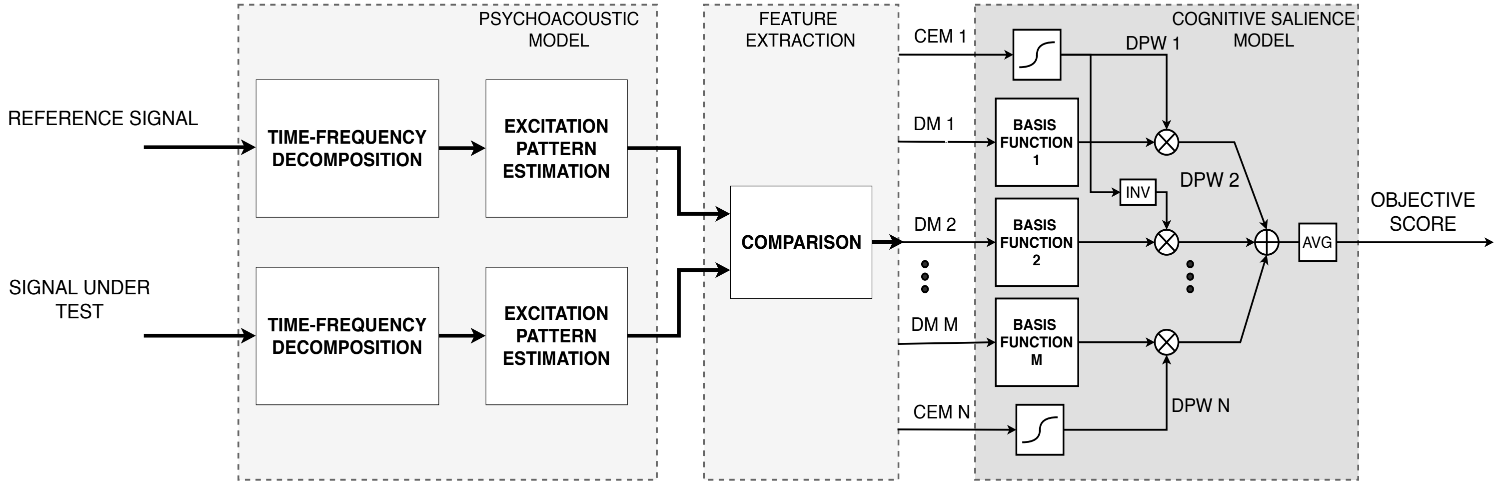
Roadmap

- Motivation
- Method
- Experiment
- Results
- Discussion
- Summary and Conclusions

© AudioLabs, 2022

Motivation

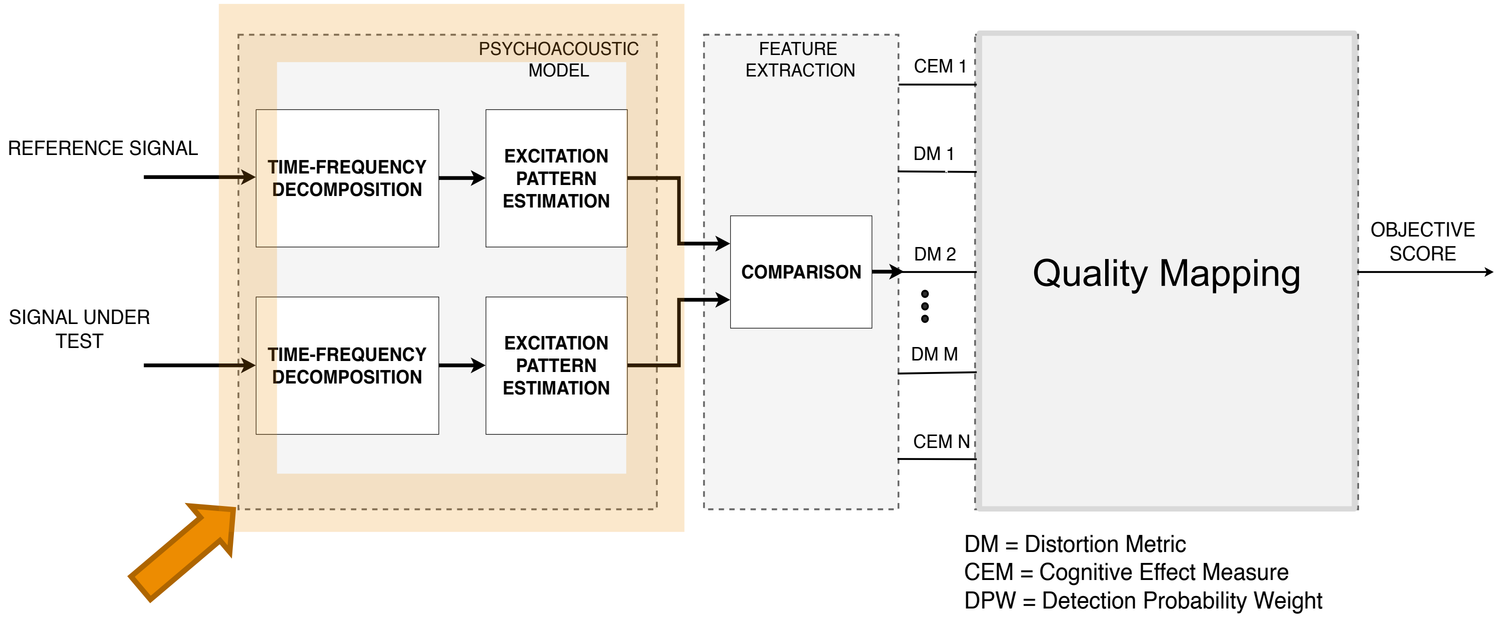
- OQMS Architecture



DM = Distortion Metric
CEM = Cognitive Effect Measure
DPW = Detection Probability Weight

Motivation

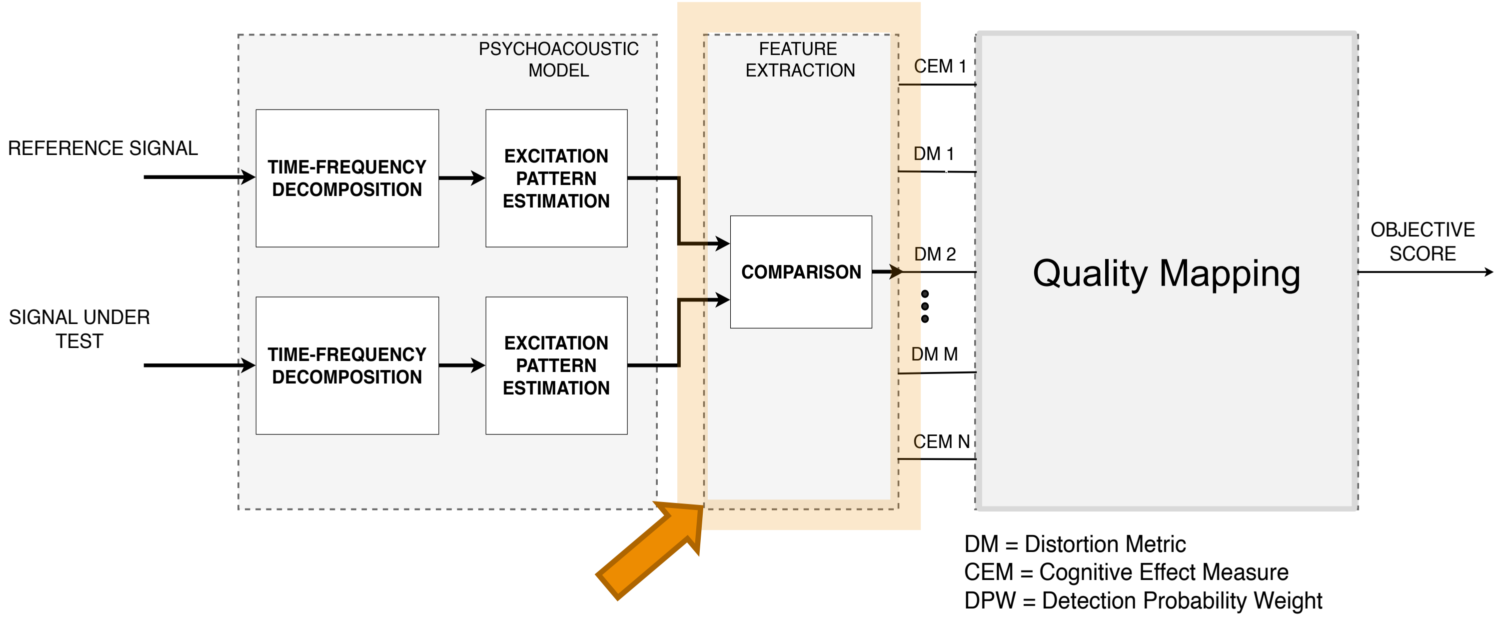
- OQMS Architecture



- Psychophysical representation of input signals (ITU-R BS.1387-1)
  - Pre-conditioning: time alignment, DC offset removal, silence removal, etc…
  - Cochlear model: time/frequency decomposition, NL filter bank, ….
  - Excitation patterns: simultaneous and non-simultaneous masking models (peripheral hearing phenomena), pattern adaptation and others…

Motivation

- OQMS Architecture



DM = Distortion Metric
CEM = Cognitive Effect Measure
DPW = Detection Probability Weight

- Feature extraction and comparison, two types of features:
  - **Distortion Metrics** (derived from PEAQ's Advanced Version): linear distortions, modulation disturbance, noise loudness, harmonic structure of errors
  - Extension: **Cognitive Effect Metrics** for informational masking (IM), perceptual streaming (PS), and probability of signal being speech-like (probSpeech)