Objective and subjective quality evaluations for speech and audio

Lectures for the course:
Digital and Audio Coding

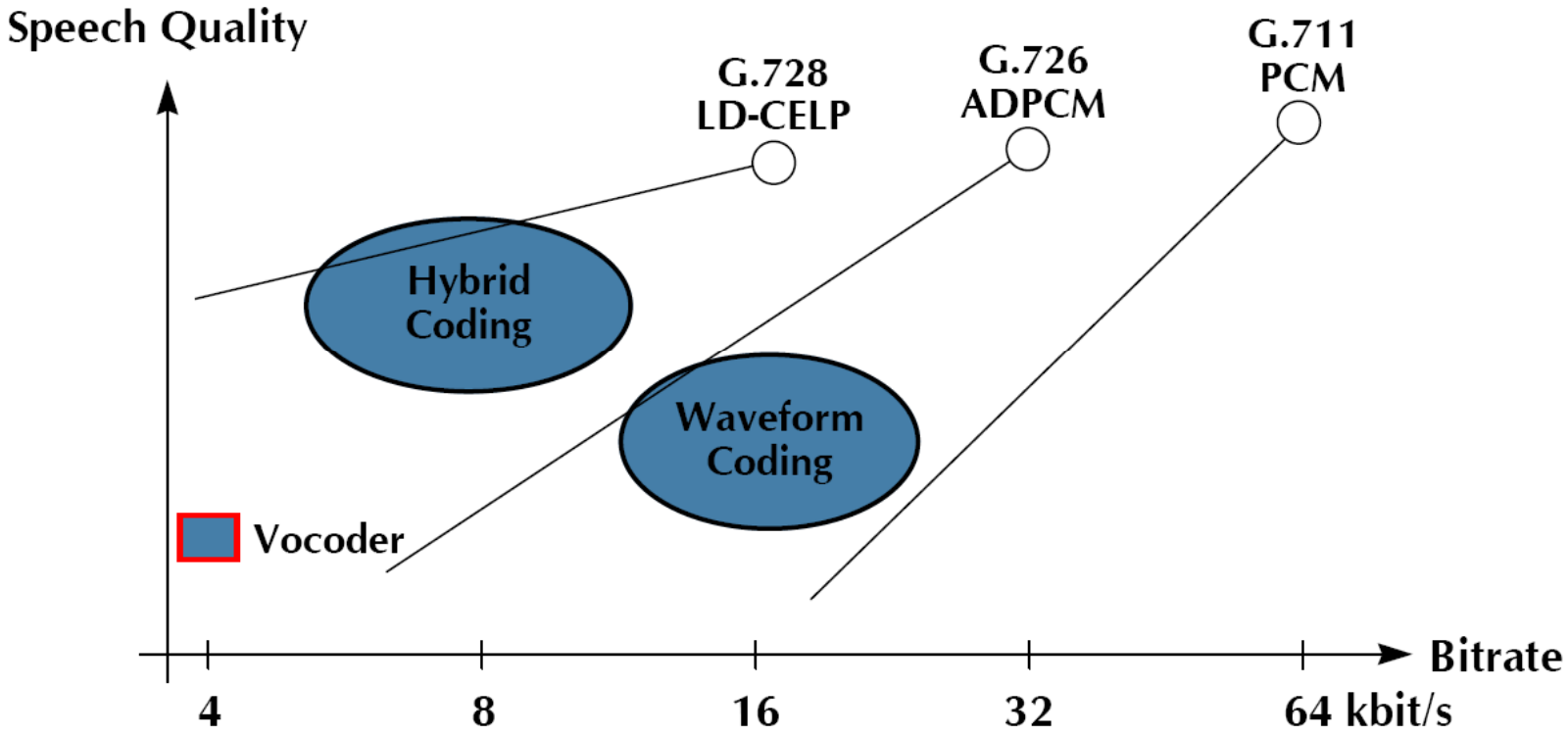
IDIAP Research Institute
Petr Motlicek and Mathew M. Doss

Outline

- Introduction – focus on speech quality
- Different physiological characteristics
- Speech intelligibility
- Voice quality
  - Subjective tests
    - Conversational quality
    - Listening quality
  - Objective tests
    - Non-Intrusive
    - Conversational Quality
    - Intrusive
      - SNR, SSNR, …
      - Perceptual domain measures
        - » Speech quality (PSQM, PESQ)
        - » Audio quality (PEAQ)
        - » Desired scheme, advantages, drawbacks of current algorithms
        - » PESQ, PEAQ analysis
          - Examples for PESQ, PEAQ

Introduction

- QoS (quality of service) testing is one of the key issues in modern telecommunications. Whether it is during the development of VoIP equipment, setting up networks or while operating a mobile network, one will always be faced with the problem to determine and optimize the speech quality.
- The speech quality of digital signals ==> function of the available bit rate
- Modern speech coding techniques:
  - allow for bit rates of 8kbit/s and less, to transmit a speech conversation.
  - This coding gain compared to wide band audio codecs can be achieved by focusing on the modeling of the human speech tract.
  - As a consequence, the codec is highly adapted to transmit speech signals, and music signals or natural sounds will be significantly distorted.



Speech Quality in the context of coding techniques

VoIP

- In the case of VoIP typically the following codecs are being used: G.711 (64 kbit/s), G.723 (5.4 and 6.3 kbit/s), G.728 (16 kbit/s) and G.729 (8 kbit/s), as well as GSM Full-Rate.
- Internet protocol (IP) networks: neither sufficient bandwidth for the voice traffic, nor a constant, acceptable delay. Dropped packets and varying delays introduce distortions not found in traditional telephony.
- In addition, if a low bit-rate codec is used in VoIP to achieve a high compression ratio, the original waveform can be significantly distorted.
- All these factors can affect psychological parameters like intelligibility, naturalness, and loudness that determine the overall speech quality.

Coding Scheme	MOS	Subjective Interpretation
64kbit/s PCM A-law	**4.3**	good, almost excellent
32kbit/s ADPCM DECT	**3.8**	good
13kbit/s GSM Full-Rate	**3.4...3.7**	fair, almost good
...	...	

Different physiological characteristics of speech quality and their dominant dependencies on physical network characteristics.

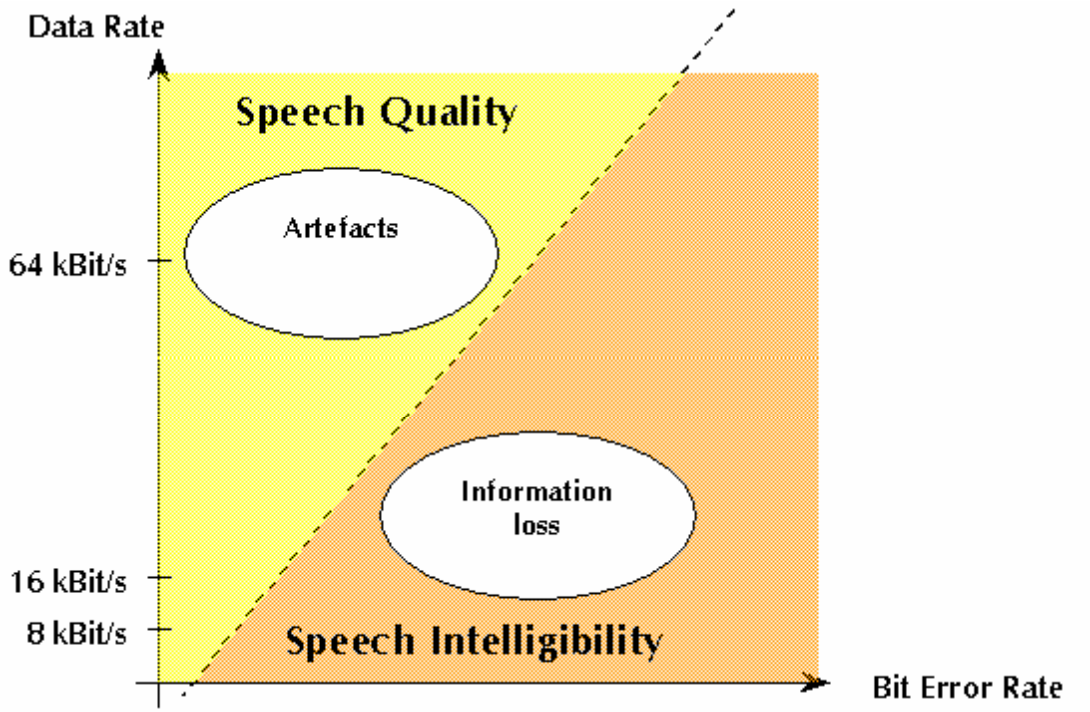
Intelligibility: measures the quality of the perception of the meaning or information content of what the talker has said. Sometimes called CLARITY: how much info can be extracted from conversation.

Naturalness: the degree of fidelity to the talker's voice.

Loudness: the absolute loudness level at the listener's side.

	Psychological Parameters			
Physical Parameters	Intelligibility	Naturalness	Loudness	*Quality*
Signal Level	+	+	+	+
Noise	+			+
Freq. Response	+	+	+	+
Distortion	+	+		+
Delay	+			+
Echo	+			+
Packet Loss	+			+

Speech intelligibility and quality versus data-rate



- the higher the bit rate, the more likely a good speech quality (not only intelligibility) will be obtained. However, the effect of bit errors increases with a lower data rate due to the increased lack of redundancy.

➔ As a summary, speech quality is first interfered by artifacts but with lower bit rates and thus more sensitivity to errors, speech intelligibility is interfered by information loss.

Speech Intelligibility

- ➔"clarity“ == how much information can be extracted from a conversation.
- depends on a large variety of factors, and only few are well understood: For example, certain frequency bands are more important.
  - for intelligibility than others: 250-800 Hz is less important for speech intelligibility than 1000-1200 Hz.
- also depends on the speech material:
  - Complete sentences are much better understood than a sequence of unrelated words due to the logical word flow in a sentence.
  - Subjective test procedures are defined based on spoken syllables, however these procedures cause too much effort to be applied in the daily operation.

Intelligibility tests

- An example: Diagnostic Rhyme Test (DRT):
  - uses a set of isolated words to test for consonant intelligibility in the initial position. The test consists of 96 word pairs that differ by a single acoustic feature in the initial consonant.
- The Modified Rhyme Test (MRT): an extension to the DRT. It tests for both initial and final consonants.
  - A set of six words is played one at a time and the listener marks which word he/she thinks he/she hears.

TESTING VOICE QUALITY

Broad classes of speech quality metrics

- Subjective measures: humans listening to a live or recorded conversation and assigning a rating to it.
- Objective measures: computer algorithms designed to estimate quality degradation in the signal.

- Speech quality: a complex psycho-acoustic phenomenon within the process of human perception.
  - necessarily subjective, even different people interpret speech quality differently.
- Objective measures are widely used due to several advantages:

Comparison of Subjective and Objective Methods for Quality Estimation. The symbol "+" is used to denote that the method is advantageous over the other method, denoted by "-".

	Subjective Measures	Objective Measures
Cost	-	+
Reproducibility	-	+
Automation	-	+
Unforeseen Impairments	+	-

Subjective tests

- human participants assess the performance of a system in accordance with opinion scale.
- Two general categories:
  - conversational quality measures
  - listening quality measures

Cont.

- Due to the lack of international standards for measuring the perceived voice quality, until a few years ago, the only widely accepted assessment procedures for voice quality were listening tests.
- first standardized within the ITU-T (International Telecommunication Union, Geneva, (former CCITT), http://www.itu.org

Conversational quality

- how listeners rate their ability to converse during the call
- It includes listening quality, ass well.
- Interactive communication scenarios, subjects are asked to complete a task over phone.
- Evaluation: efficacy of the performance of the task – quality measure for effects like delay, echo, loudness, …

Listening quality

- Listeners rate what they hear during the call.
- ignores effects as echoes at the talker side, transmission delays.

A) Absolute Category Ratings (ACR) test (Recommendation P.800)

Impairment	Grade
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

- pool of listeners rate a series of audio files using the impairment scale. After obtaining individual scores, the mean opinion for each audio file is obtained.
- Useful method for testing telephone band speech signal.
- The recommended test method for listening-only tests is the "Absolute Category Rating" (ACR) method.
- Used for assessment of speech codecs since 1993.
- 5 grade impairment scale applied.
- Testing is done without a comparison to an undistorted reference.
- large pool of listeners: 20 – 50 test subjects with identical series of speech fragments.
- Done under controlled conditions.
- MOS – mean opinion score (the most widely used method to evaluate overall speech quality).

B) Degradation category rating (DCR) test

- Listeners hear the reference and the test signals sequentially, and are asked to compare them.
- Degradation MOS – measure how different distortion in speech are perceived.

Very annoying	1
Annoying	2
Slightly annoying	3
Audible, but not annoying	4
Inaudible	5

Comparison Category Rating (CCR)

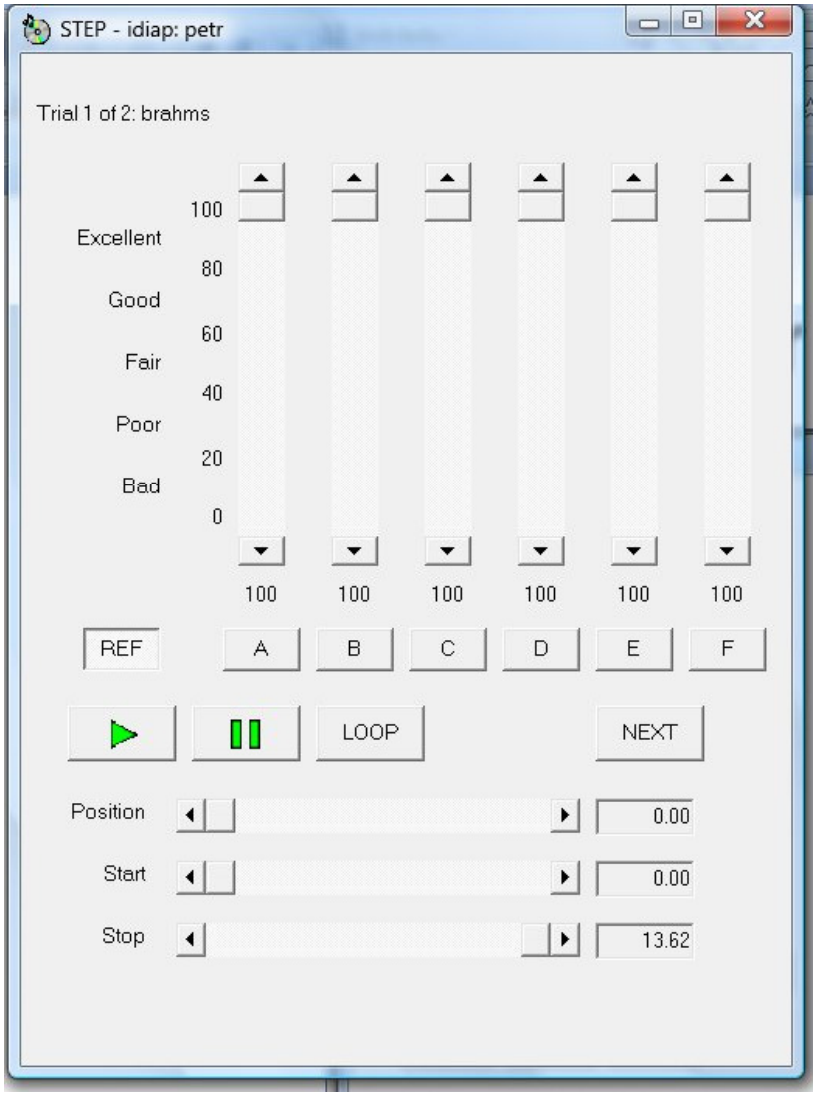
- Variation of DCR – listeners identify the quality of the second stimulus relative to the first one on the scale.
- DCR is more common in audio quality assessment, while speech coding systems are typically assessed by an ACR test.

Much better	3
Better	2
Slightly better	1
About the same	0
Slightly worse	-1
Worse	-2
Much worse	-3

MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA): ITU-R BS.1534

- Example of a DCR test
- a method for the subjective assessment of intermediate quality level of coding systems.
- MUSHRA: a double-blind multi-stimulus test method with a hidden reference and hidden anchors.
- In this test, the subjects are required to score the stimuli according to the continuous quality scale from 0 to 100.
- The listener records his/her assessment of the quality with the use of sliders on an electronic display.

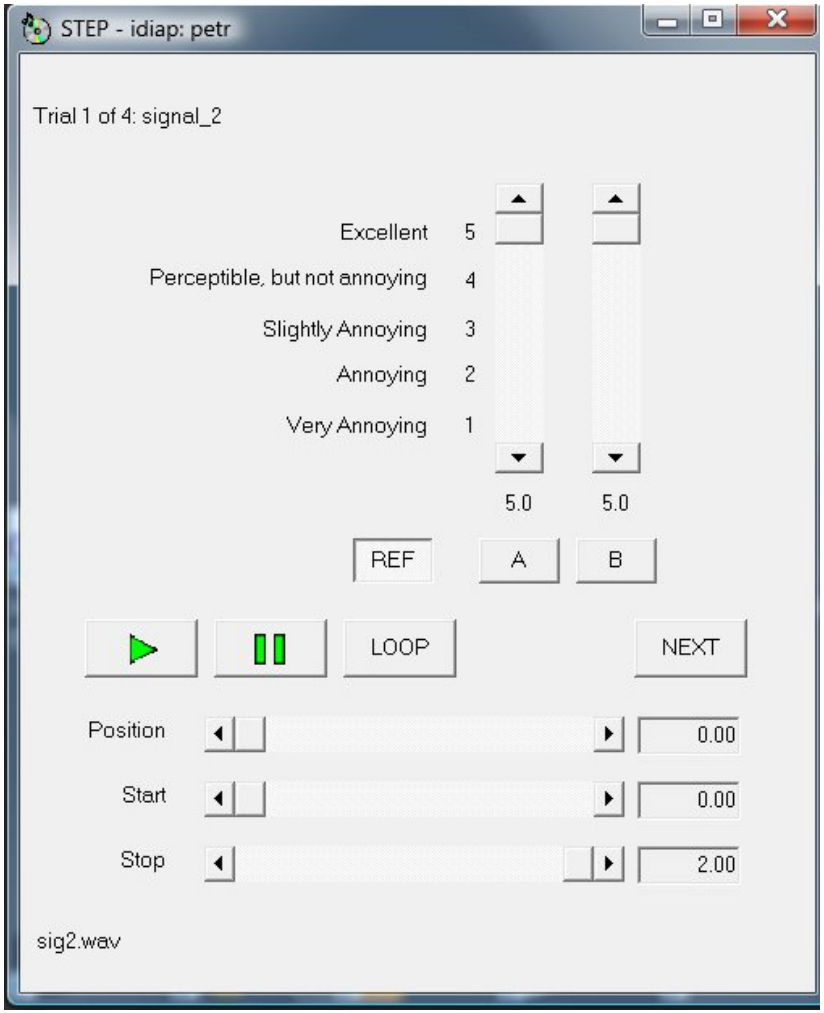
Graphical user interface for the MUSHRA test: the test subject can compare the files under test (buttons A-F) with the original signal (button REF).
It is defined by ITU-R recommendation BS.1534.



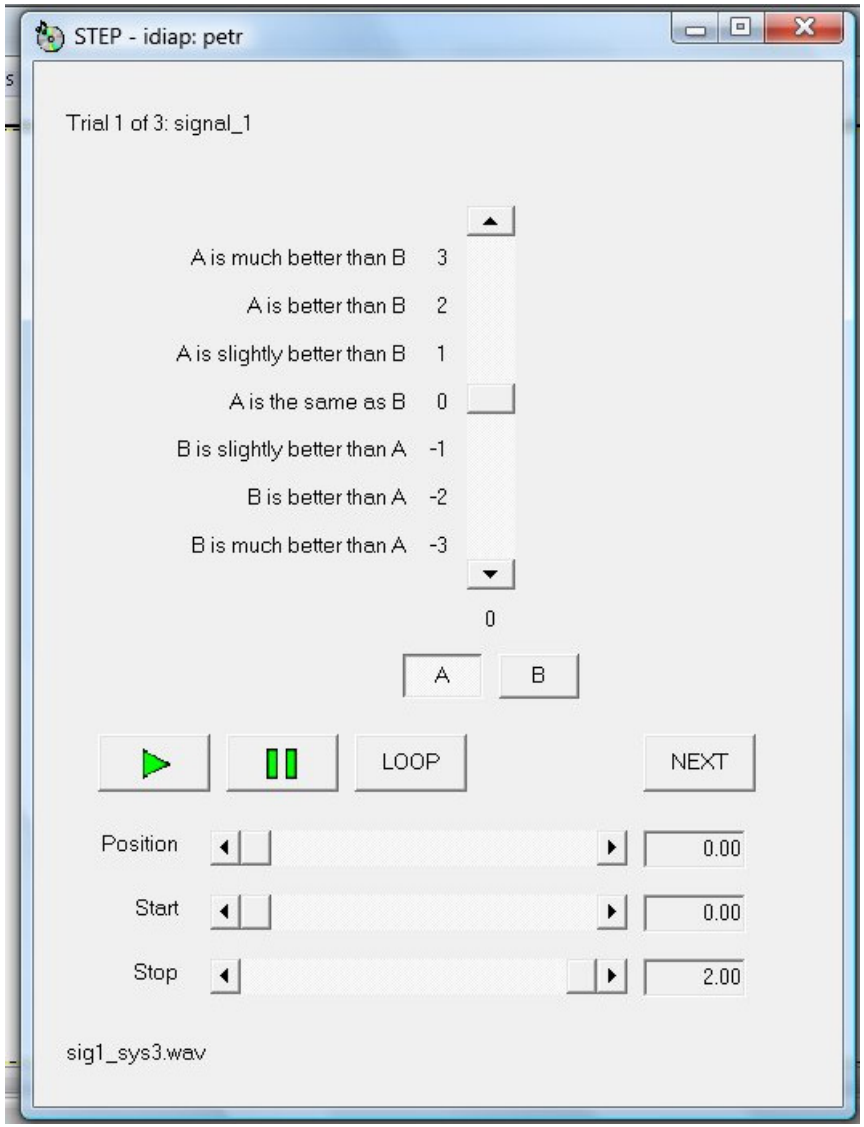
http://www.itu.int/publ/R-REC/e

- ITU-R recommendation: ITU-R BS. 116-1, method for the subjective assessment of small impairments in audio systems including multichannel sound systems:
  - called double blind triple-stimulus
  - only two examples (one of them is hidden reference)
  - Example of DCR test

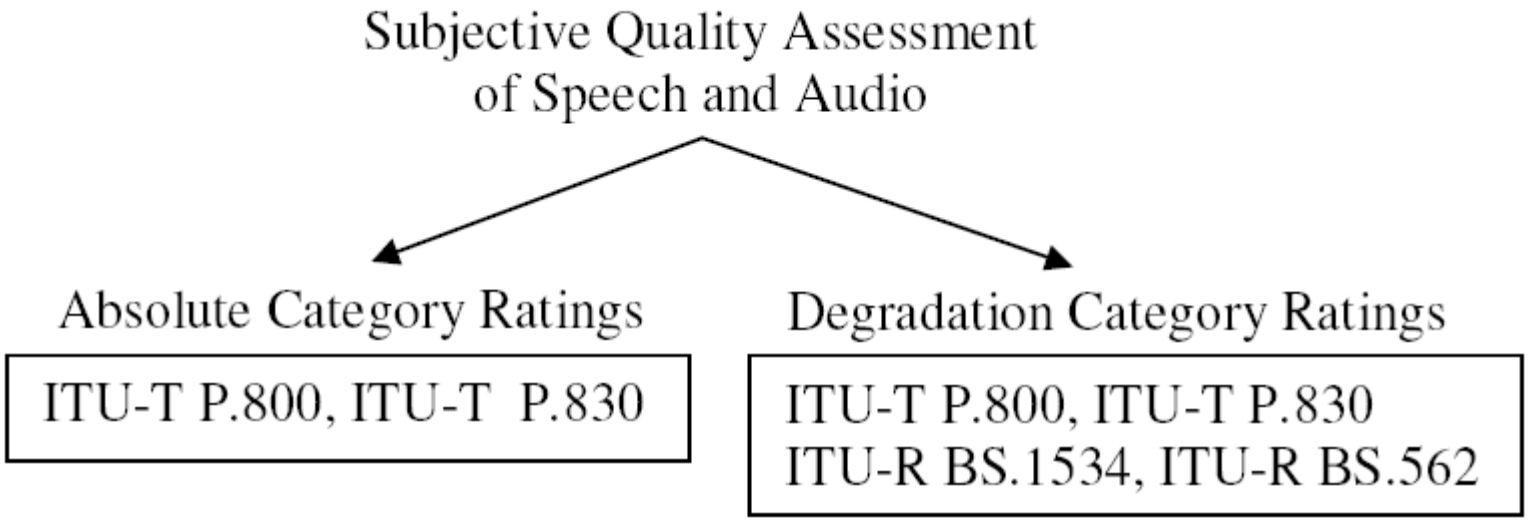
ITU-T BS.1116: Methods for the subjective assessment of small impairments in audio systems including multi channel sound systems



A-B comparison: example of CCR test



Conclusions: the two major types of subjective quality assessment methods and related ITU standards and recommendations.



- A classification of the most popular ACR and DCR tests - standardized by the ITU.
- Major conceptual differences between the two tests:
  - in ACR, even an original signal can receive low grade, since listeners compare with their internal model of "clean speech".
  - DCR tests provide a quality scale of higher resolution, due to comparison of the distorted signal with one or more reference/anchor signals.

Cont. (MOS terminology)

- **Defined in ITU-T Rec. P.800.1**
- The mean of opinion scores, i.e., of the values on a predefined scale that subjects assign to their opinion of the performance of the telephone transmission system used either for conversation or for listening to spoken material.
- To distinguish the area of application:
  - LQ refers to Listening Quality, CQ refers to Conversational Quality, S refers to Subjective, O refers to Objective, and E refers to Estimated.

	Listening-only	**Conversational**
Subjective	MOS-LQS	MOS-CQS
Objective	MOS-LQO	MOS-CQO
Estimated	MOS-LQE	MOS-CQE

ITU Standards

- http://www.itu.int/publications/template.aspx?oas=y&target=/publications/Subscription.html

Objective Measures

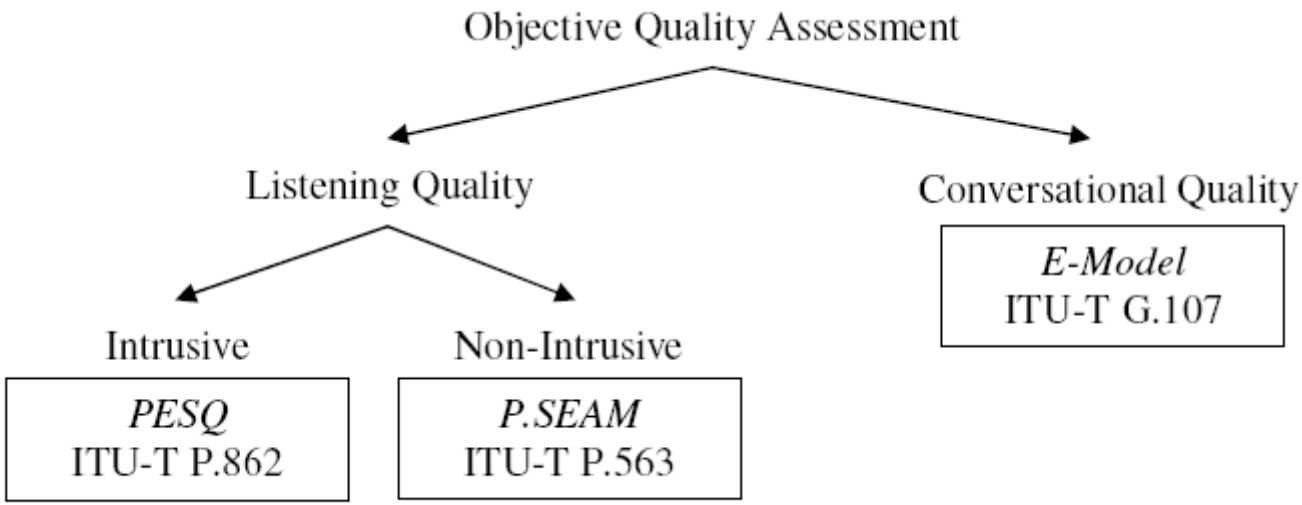
- Subjective tests – expensive, time-consuming, labor-intensive.
- Objective quality algorithms can be used instead, but they have to be properly "calibrated" to the output of subjective quality tests.
- Accuracy of the objective testing is determined by:
  - its correlation with MOS scores for a set of data
  - The estimation performance is assessed using:
    - Correlation coeff. And RMSE (between the predicted quality ~~Q~~ and the measured subjective quality Q)
    - Evaluation done over a large multi-language database with wide range of distortions.
    - N – number of MOS labeled utterances used;

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^{N}(Q_i - \hat{Q}_i)^2}{N}},$$

$$R = \frac{\sum_{i=1}^{N}(\hat{Q}_i - \mu_{\hat{Q}})(Q_i - \mu_Q)}{\sqrt{\sum_{i=1}^{N}(\hat{Q}_i - \mu_{\hat{Q}})^2}\sqrt{\sum_{i=1}^{N}(Q_i - \mu_Q)^2}},$$

Cont.

- Estimation of listening/conversational subjective quality.
- Intrusive/non-intrusive – according to the input information they require



Non-Intrusive Listening Quality Measures

- the original speech signal may not be available, or it may be difficult to align it to the processed speech signal.

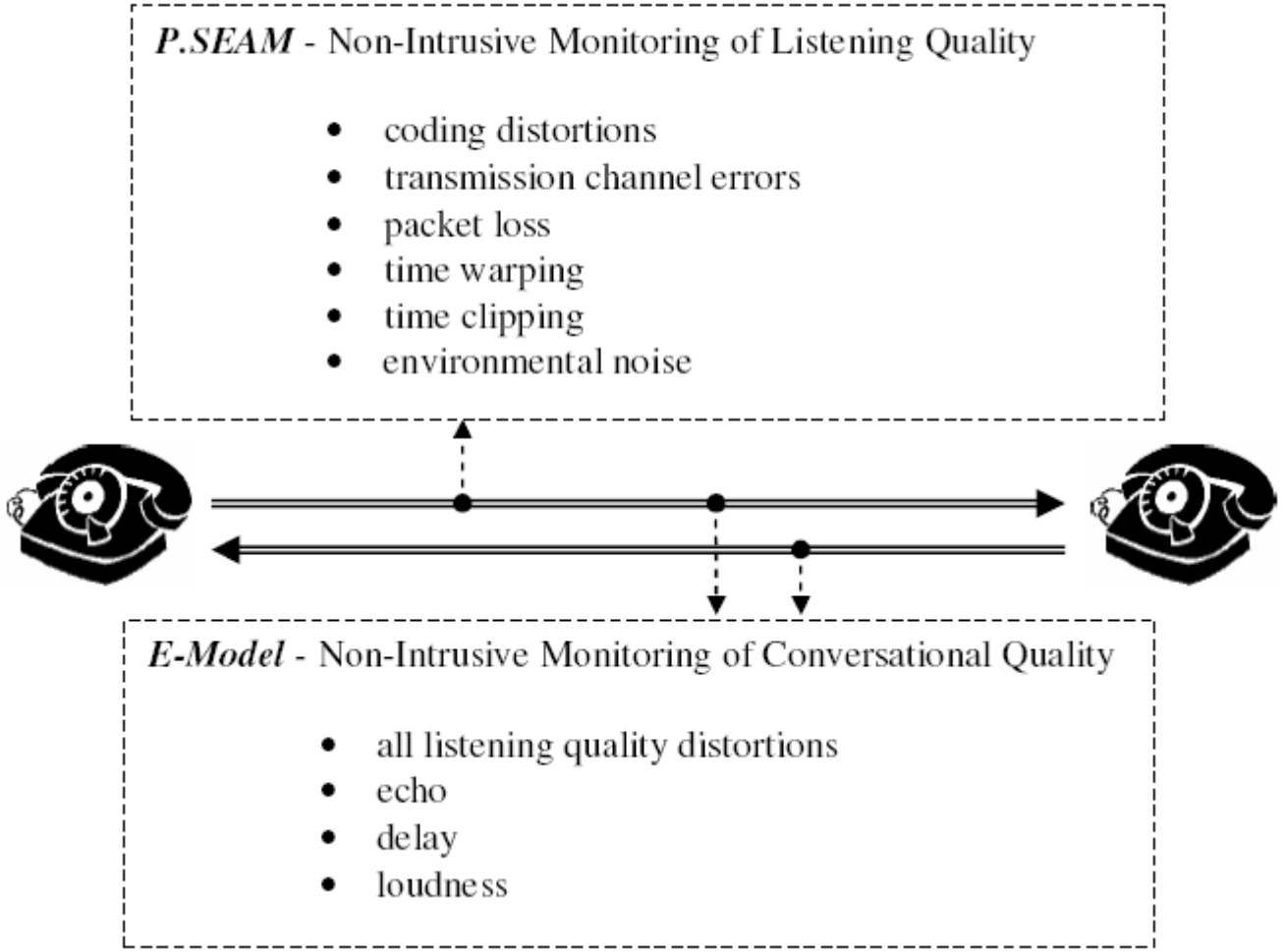
  => to predict the speech quality from the processed signal only.
- Usually algorithms perform a perceptual transform on the input signal,
- BUT offer a large variety of mapping schemes, such as Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), Neural Networks, etc.

  ➔ important in monitoring of communication systems, such as wireless communications and VoIP

Objective Measures for Assessment of Conversational Quality

- E-model: an estimate of the conversational subjective quality
  - a purely parametric model
  - determines a transmission rating model that monitors many different parameters and combines their values into an end-performance factor.
- The objective is to determine a transmission quality rating R factor, with range typically between 0 and 120:
  - R can be converted to estimated listening and conversational quality MOS scores.
- Nowadays used non-intrusively over the network as a passive monitoring tool.

Non-intrusive monitoring of listening and conversational quality over the network



Intrusive listening quality measures

SNR, SSNR

- SNR and SSNR – simplest and historically most common techniques
  - s, y are original and distorted speech vectors
- SNR is a term for the power ratio between a signal and the back-ground noise
- SNR and SSNR show little correlation to perceived speech quality.

$$SNR = 10\log_{10}\frac{\sum_n s^2(n)}{\sum_n (s(n)-\hat{s}(n))^2}$$

$$SNR_{Seg} = \frac{1}{N}\sum_{k=1}^{N} 10\log_{10}\left[\frac{\sum_{n\in \text{frame}_k} |s(n)|^2}{\sum_{n\in \text{frame}_k} |\hat{s}(n)-s(n)|^2}\right]$$

$$\mathbf{e} = \mathbf{s} - \mathbf{y}.$$

$$\mathrm{d}_{SNR}(\mathbf{s},\mathbf{y}) = 10\log_{10}\left(\frac{\mathbf{s}^T\mathbf{s}}{\mathbf{e}^T\mathbf{e}}\right),$$

$$\mathbf{e}_n = \mathbf{s}_n - \mathbf{y}_n$$

$$\mathrm{d}_{SSNR}(\mathbf{s},\mathbf{y}) = \frac{1}{N}\sum_{n=1}^{N} 10\log_{10}\left(\frac{\mathbf{s}_n^T\mathbf{s}_n}{\mathbf{e}_n^T\mathbf{e}_n}\right),$$

Frequency domain measures

- Frequency weighted segmental SNR
- Weighted spectral slope measure

$$SNR_{FWS}=\frac{1}{N}\sum_{k=1}^{N}\frac{1}{W_k}\sum_{j=1}^{F}10\log_{10}\left[\frac{w_{j,k}\cdot\sum|s(n)|^2}{\sum|\hat{s}(n)-s(n)|^2}\right]$$

$$W_k=\sum_{j=1}^{F}(w_{j,k})$$

Cont.

- Frequency domain measures:
  - Known to be significantly better correlated with human perception.
  - Less sensitivity to signal misalignment.
  - Gain normalized approach – most popular.
  - Popular frequency domain measures: Log-spectral distance, Itakura-Saito, Log-Likelihood, and Log-Area-Ratio measures
  - Log-spectral distance: The **log-spectral distance (LSD)**, also referred to as **log-spectral distortion**, is a distance measure (expressed in dB) between two spectra. The log-spectral distance between spectra *P*(ω) and is defined as

$$\mathrm{d}_{SD}(\mathbf{s},\mathbf{y}) = \frac{1}{N}\sqrt{\sum_{n=1}^{N}\int_{-\pi}^{\pi}\left(10\log_{10}\left(\frac{P_{\mathbf{s}}(\omega,n)}{P_{\mathbf{y}}(\omega,n)}\right)\right)^2\frac{d\omega}{2\pi}},$$

Itakura-Saito divergence

- The **Itakura–Saito distance** is a measure of the perceptual difference between an original spectrum *P*(ω) and an approximation of that spectrum. It was proposed by Fumitada Itakura and Shuzo Saito in the 1970s while they were with NTT.
- Alpha and beta denote the LPCs for the frame k of the signal s and y=s^
- M – prediction order

$$D_{IS}(P(\omega), \hat{P}(\omega)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{P(\omega)}{\hat{P}(\omega)} - \log \frac{P(\omega)}{\hat{P}(\omega)} - 1 \right] d\omega$$

$$\alpha^T(k) = [1 \ -a_{s,1} \ \cdots \ -a_{s,M}]$$

$$\beta^T(k) = [1 \ -a_{\hat{s},1} \ \cdots \ -a_{\hat{s},M}]$$

$$d_{IS} = \frac{1}{N} \sum_{k=1}^{N} \log \frac{\beta^T(k) R_s(k) \beta(k)}{\alpha^T(k) R_s(k) \alpha(k)}$$

Log-Area ratio

Log area ratios (LAR) can be used to represent reflection coefficients (another form for linear prediction coefficients) for transmission over a channel. While not as efficient as line spectral pairs (LSPs), log area ratios are much simpler to compute. Let r_k be the kth reflection coefficient of a filter, the kth LAR is:

$$A_k = \log \frac{1 + r_k}{1 - r_k}$$

$$E_0 = R(0) \tag{38a}$$

$$k_i = -\left[R(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)\right] \Big/ E_{i-1} \tag{38b}$$

$$a_i^{(i)} = k_i$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)}, \quad 1 \leqslant j \leqslant i-1 \tag{38c}$$

$$E_i = (1 - k_i^2) E_{i-1}. \tag{38d}$$

Equations (38b)–(38d) are solved recursively for $i = 1, 2, \cdots, p$. The final solution is given by

$$a_j = a_j^{(p)}, \quad 1 \leqslant j \leqslant p. \tag{38e}$$

$$\frac{Z_{i+1}}{Z_i} = \frac{1 + k_i}{1 - k_i}, \quad 1 \leqslant i \leqslant p.$$

Perceptual domain measures

- Based on the research done by Zwicker, Schröder, Brandenburg et al. in 80's.
- The Bark Spectral Distortion – one of the first methods based on models of human auditory perception:
  - Average Euclidean distance between orig. and distorted speech signals in the Bark domain.
- Perceptual Speech Quality (PSQM): ITU-T Rec. P.861 – objective analysis of speech codecs
  - PSQM correlated up to 98 percent with the scores of subjective listening tests (selected out of 12 algorithms).
  - Designed to assess the performance of speech codecs and impairments encountered in network.
  - Accuracy not sufficient.
  - With the ongoing development of speech coding, especially for packet transmission, also newer algorithms for speech quality measurement were developed, like PSQM+, PSQM99, MNB, PAMS, TOSQA, PACE and VQI. Verification tests performed by the ITU showed that far the best of these was PSQM99. The second best was PAMS, but none of these proposals was good enough for a revision of the P.861 standard.
- The most successful ITU measures in 1990s combined into PESQ (PSQM99 with an improved delay compensation)
  - PESQ (ITU-T Draft Rec. P.862):
    - Intended for measuring of narrow band quality telephone signals (like PSQM)
    - quality estimate in the following environments: speech codecs, transmission channel errors, speech input level at the codec, noise added by the system, time warping, packet loss, and time clipping.
    - Currently also wide-band extension exists

Wideband audio quality

- PAQM, PSQM, NMR [4], PERCEVAL, DIX, OASE, POM, ... , all of them developed for wideband audio codecs.
  - REASON: perceptual codecs started earlier in the broadcast environment, than it did in telecommunications.
- Perceptual Evaluation of Audio Quality (PEAQ) (ITU-R BS.1387)
  - Developed from the above systems and standardized in 1998.

Review: PSQM, PESQ, PEAQ

- All three standards, ITU-T P.861, ITU-T P.862 and ITU-R BS.1387:
  - today represent the state-of-the-art technique for the objective evaluation of the perceived speech/audio quality.
  - however, all of these techniques were derived from modeling the corresponding subjective experiment by an algorithm based approach. Thus it is essential to understand the scope of the modeled subjective experiment when trying to interpret the calculated results.

Cont.

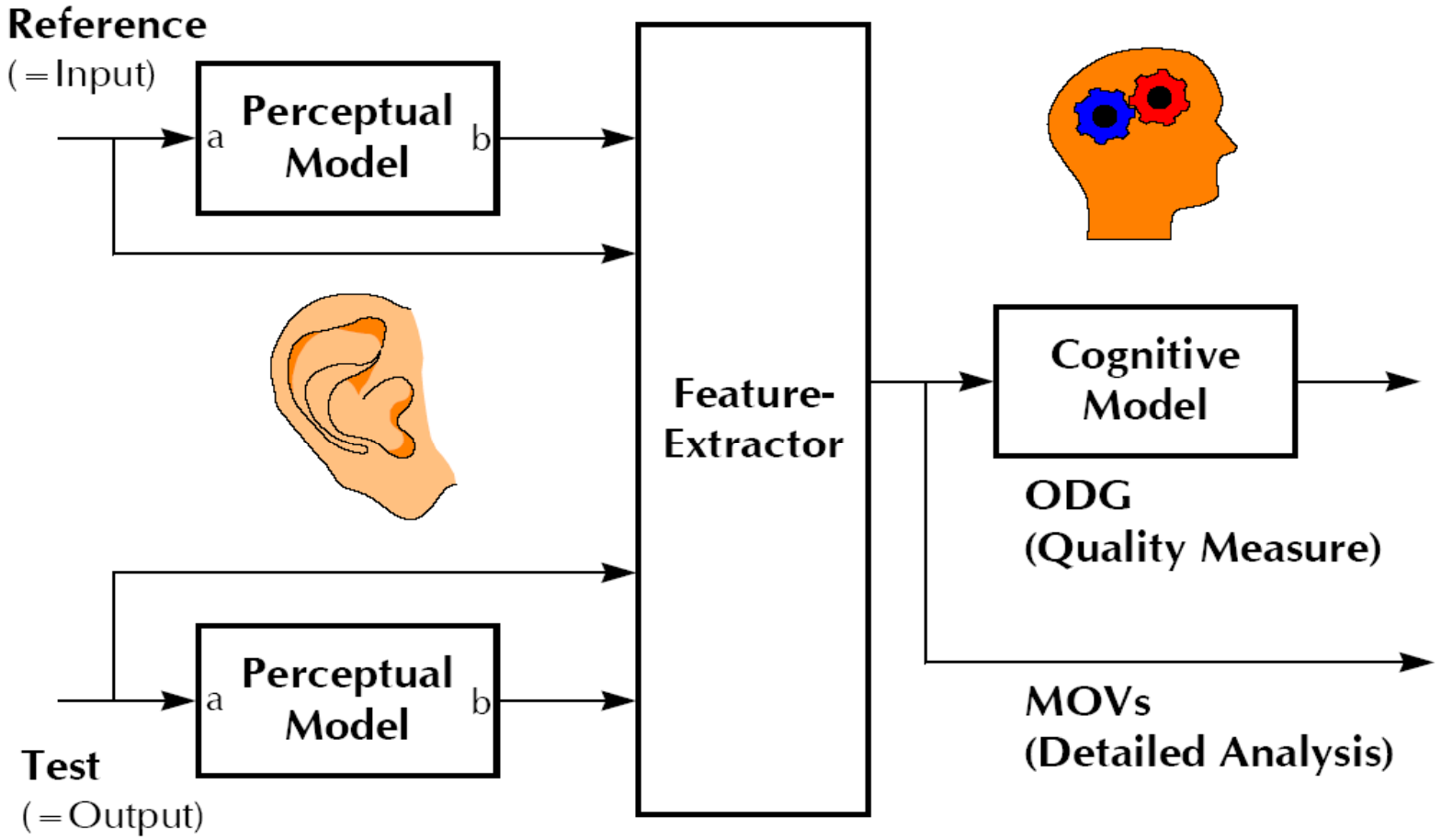
- Comparison of all of the relevant measurement algorithms (pesq, peaq):
  - they can be broken down to a block diagram.
  - Although they significantly differ in the way they try to model human perception, they also show a very high degree of similarity in their basic structure.
  - In general - wideband audio signals this part of the algorithm is more important than for speech quality measures, and therefore it is modeled more accurately in e.g. PEAQ.
  - The algorithm models the audible distortion present in the signal under test by comparing the outputs of the ear models. The information obtained by this process is called MOVs ("Model Output Variables"), and may be useful for a detailed analysis of the signal.

➔ The final goal instead is deriving a quality measure, consisting of a single number that indicates the audibility of the distortions present in the signal under test.

➔ some further processing of the MOVs is required, which simulates the cognitive part of the human auditory system.

Various proposals exist for this step. They range from algorithmic descriptions (e.g. PESQ) to artificial neural networks (e.g. PEAQ).

→ most algorithms require time aligned input signals - the process how to achieve this is usually not part of the model description.


Reference
(=Input)
a
Perceptual Model
b
Feature-Extractor
Cognitive Model
ODG
(Quality Measure)
MOVs
(Detailed Analysis)
a
Perceptual Model
b
Test
(=Output)

Algorithmic blocks for PSQM, PESQ, PEAQ:

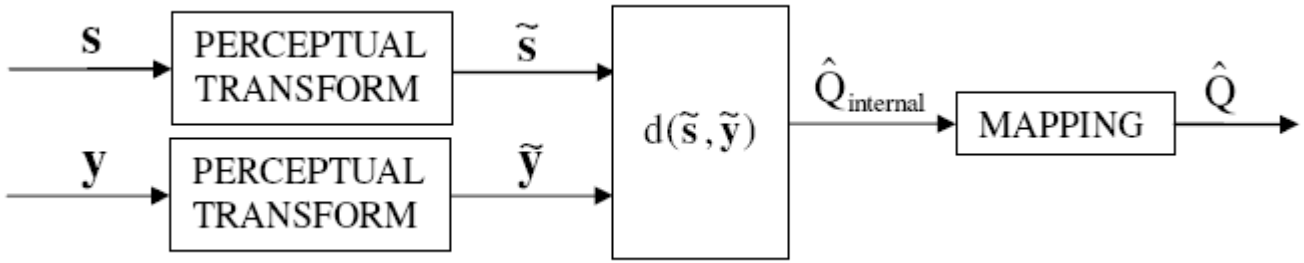
1) the signals are processed by filter that simulates the frequency response of a typical telephone headset.

2) a “Hoth” noise is injected to model a typical listening environment.

3) an intensity warping is performed, to model the relationship between signal power and perceived loudness.

4) a loudness scaling is performed to equalize the momentary compressed loudness of the two signals.

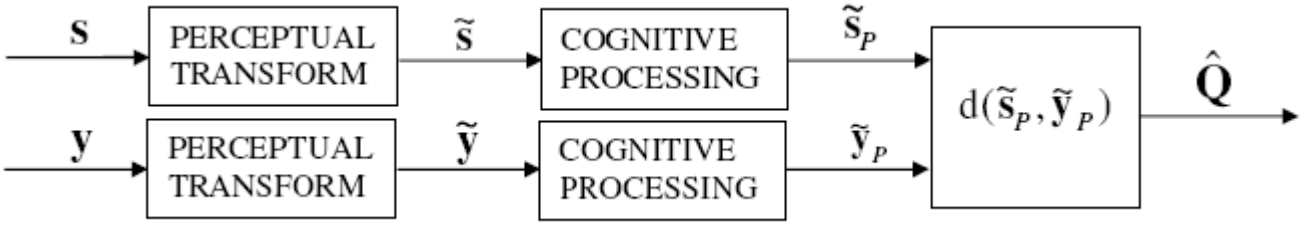
5) the distance between the transformed signals is calculated and mapped to an estimate of MOS value.



Cont.

- The final part of the human judgment process entails cognitive processing in the brain, where compact features are extracted from auditory excitations.
- The algorithms incorporate knowledge of the low-level auditory processing,
  - but neglect the high-level cognitive processing, performed by the brain.
  - Few examples exist:
    - Measuring Normalizing Blocks (MNB) – relatively simple perceptual transform, but a sophisticated error pooling system.
    - Statistical data mining - a large pool of candidate features is created and the ones that lead to the most accurate prediction of perceived quality are selected.

Desired scheme of perceptually motivated speech quality assessment measure



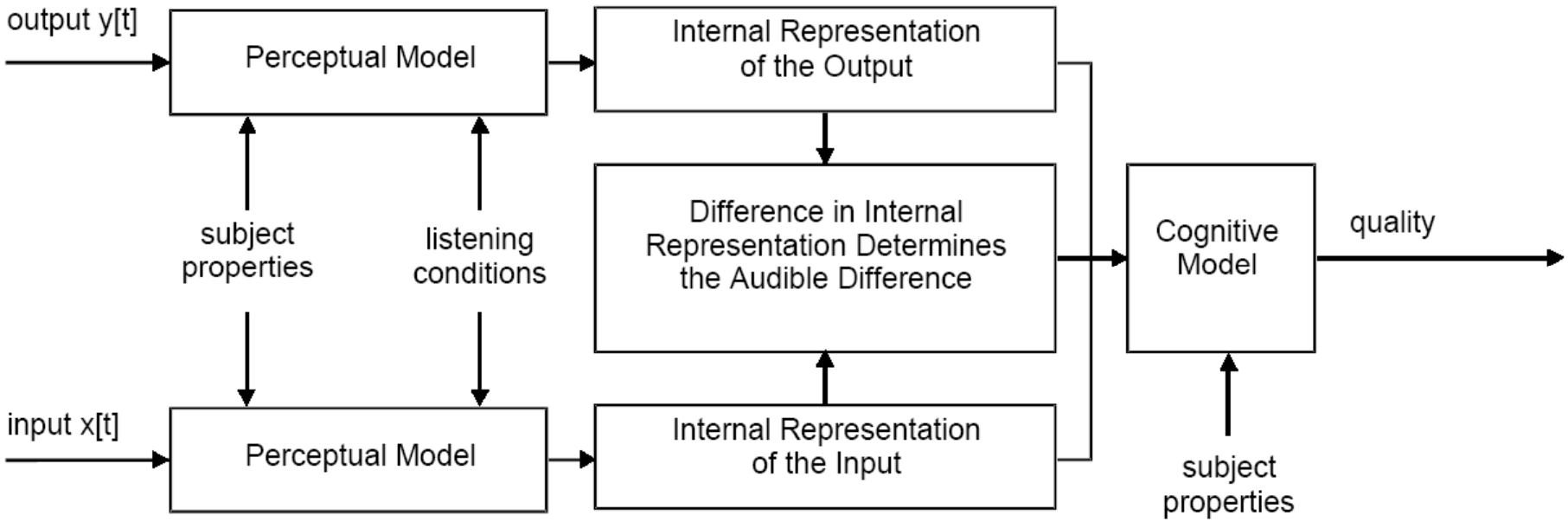
The weakness of the majority of existing perceptually motivated speech quality measures:

- exploit the knowledge of the human auditory system to weight more the error signal in regions where it is more audible.
- BUT - more audible does not necessarily mean more objectionable, since the latter is dependent of the a-priori information in the human brain.
- NO guarantee that less audible parts of the signal may not be of higher importance for the pattern extraction and comparison process performed by the human brain, after the signal has been perceptually transformed.

a) PSQM, PSQM+

- the perceptual speech quality measure
- the psychoacoustic effects known from masking experiments seem to differ in significance, when comparing the perception of speech and music signals.
  - human brain possibly recalls the reference sound of familiar voices more accurately from the daily life experience, compared to music sounds.
  - Up to now, no single homogeneous approach has been presented that would allow for high correlation with both, speech, and music signals without adapting algorithm parameters
- mapped onto psychophysical representations that match the internal representations of the speech signals (the representations inside our heads) as closely as possible.
- This difference is used for the calculation of the noise disturbance as a function of time and frequency. In PSQM, the average noise disturbance is directly related to the quality of coded speech.
- The standard version of PSQM as defined by P.861 has three major drawbacks:
  - The time alignment
  - The asymmetry processing of PSQM weights loud distortions much stronger than a human listener would do.
  - On time clipped passages (e.g. caused by dropouts or packet loss) the opposite effect shows up.

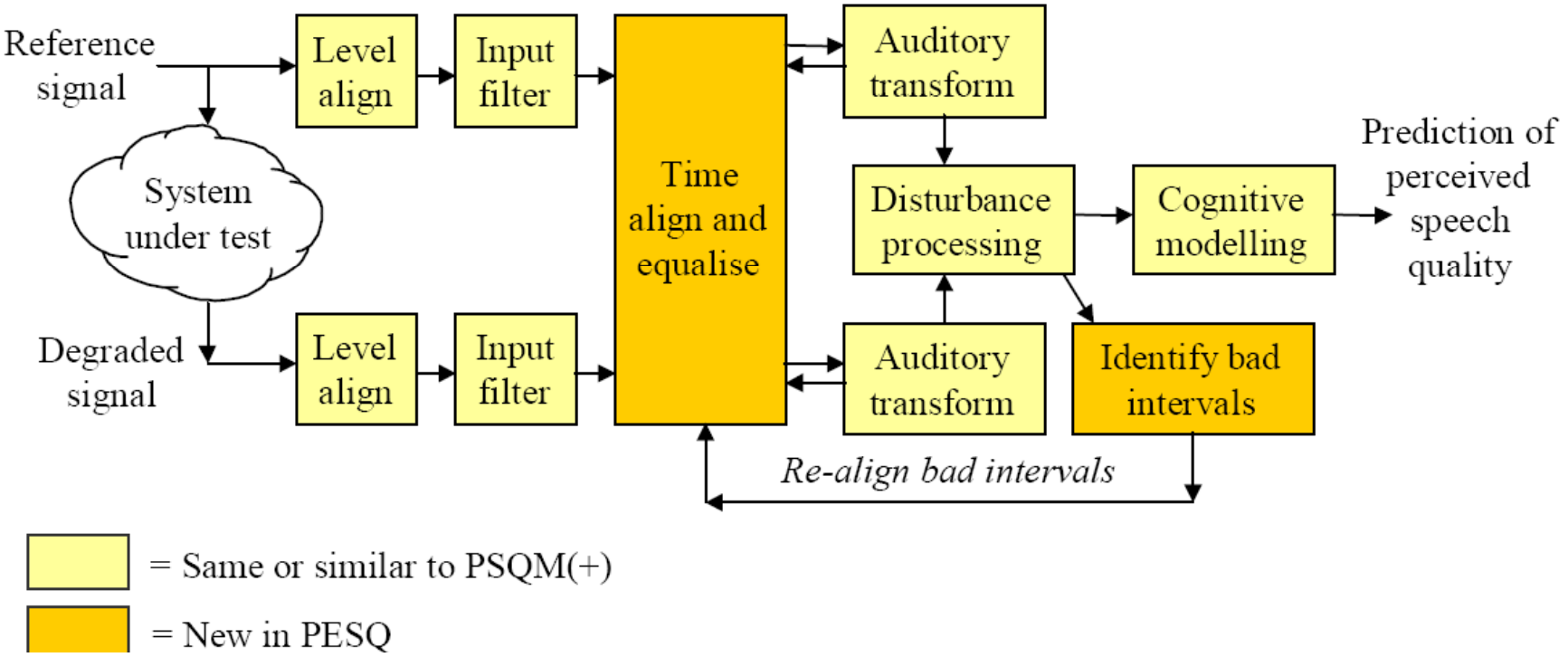
Cont.


output y[t]
Perceptual Model
Internal Representation of the Output
subject properties
listening conditions
Difference in Internal Representation Determines the Audible Difference
Cognitive Model
quality
input x[t]
Perceptual Model
Internal Representation of the Input
subject properties

b) PESQ – P.862

- PSQM developed for codecs used in mobile transmission, like GSM, …
- With modern networks, such VoIP:
  - dealing with much higher distortions as with GSM codecs
  - the delay between the reference and the test signal is not constant anymore.
- It combines the excellent psychoacoustic and cognitive model of PSQM+ with a time alignment algorithm that perfectly handles varying delays.
- The only drawback of PESQ:
  - not designed for real-time applications. This is in turn why it cannot fully replace PSQM+.
- Wide-band extension (only for 16kHz) signal – P. 862.2

PESQ


Reference signal
Level align
Input filter
Time align and equalise
Auditory transform
Prediction of perceived speech quality
System under test
Disturbance processing
Cognitive modelling
Degraded signal
Level align
Input filter
Auditory transform
Identify bad intervals
Re-align bad intervals
= Same or similar to PSQM(+)
= New in PESQ

c) PEAQ – BS. 1387

- Nowadays - most accurate and most detailed perceptual model
- two options: a Basic version and an Advanced version.
  - The Basic version uses a FFT based ear model,
  - the Advanced version uses that model as well as a filter bank based ear model.
  - In both cases, model output variables are combined using a trained neural network to give a single metric, the Objective Difference Grade (ODG) which measures the degradation of a test input relative to a reference input.

Time to Frequency Domain (FFT-based Ear Model)

- Intended for 48 kHz sampled signal.
- Frames of 43ms/50% overlap.
- Hann windowing.
- DFT.
- Calibration of equal loudness: A calibration step is needed to fix the mapping from input signal levels to loudness.
- Scaling factor corresponding to a full-scale test sine.
- **Outer and Middle Ear Modelling.**
- **Critical Band Decomposition:**
  - The grouping into critical bands uses a frequency to Bark scale conversion
  - For the Basic version, there are 109 filter bands; for the Advanced version there are 55 bands. The band edges in Hz are given to 3 decimal places in tables in BS.1387.
- **Internal Noise:**
  - An offset is added to the band energies to compensate for internal noise generated in the ear
- **Frequency Spreading:**
  - The spreading function is level and frequency dependent.
- **Time Domain Spreading:**
  - depends on multiple frames.

Outer and Middle Ear Modeling

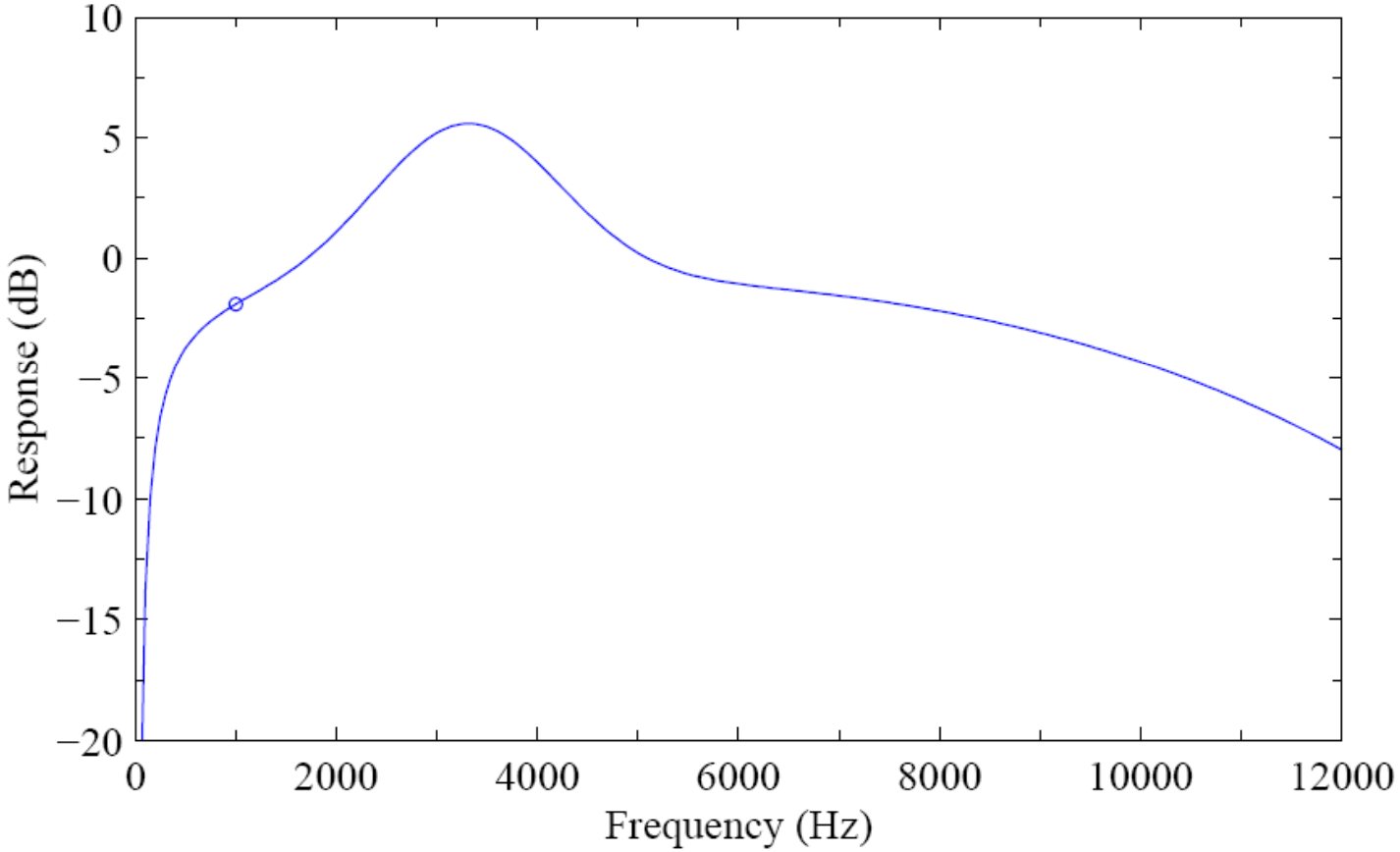


Fig. 1 Outer and middle ear response. A marker appears at 1 kHz.

Internal Noise

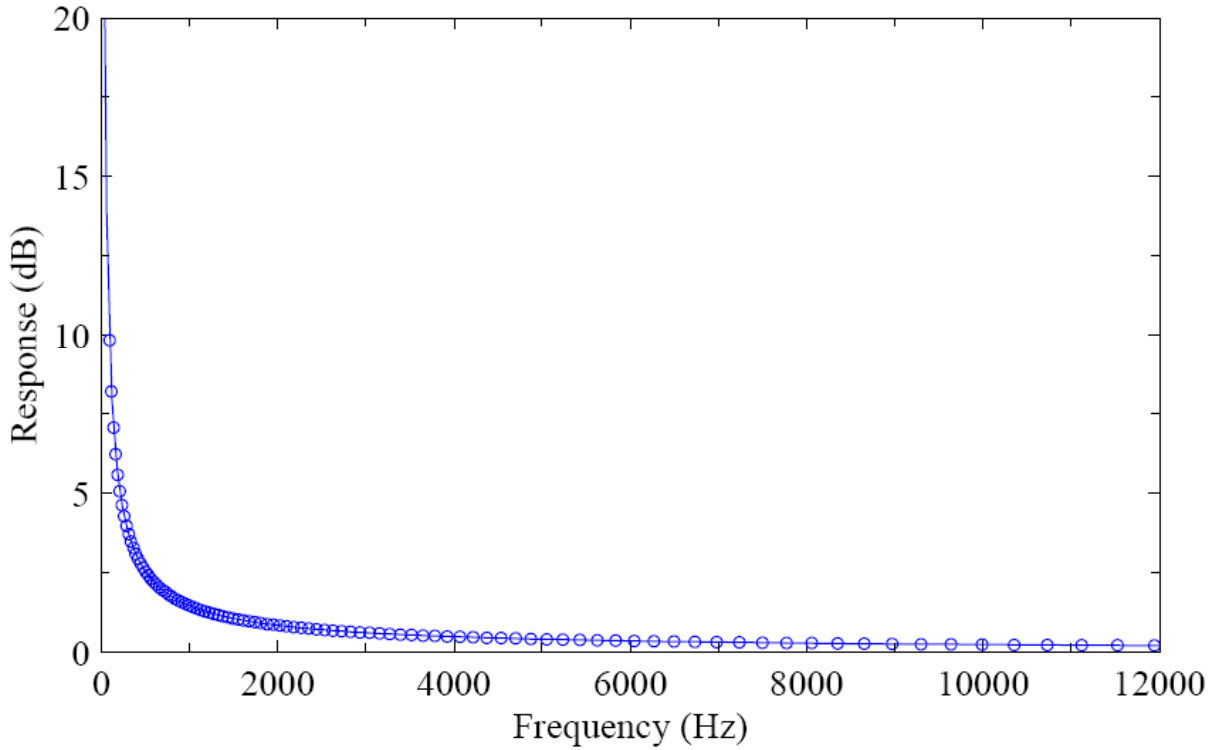


Fig. 2 Internal noise contribution. The markers indicate the centres of the frequency bands for Basic version of PEAQ.

Frequency Spreading

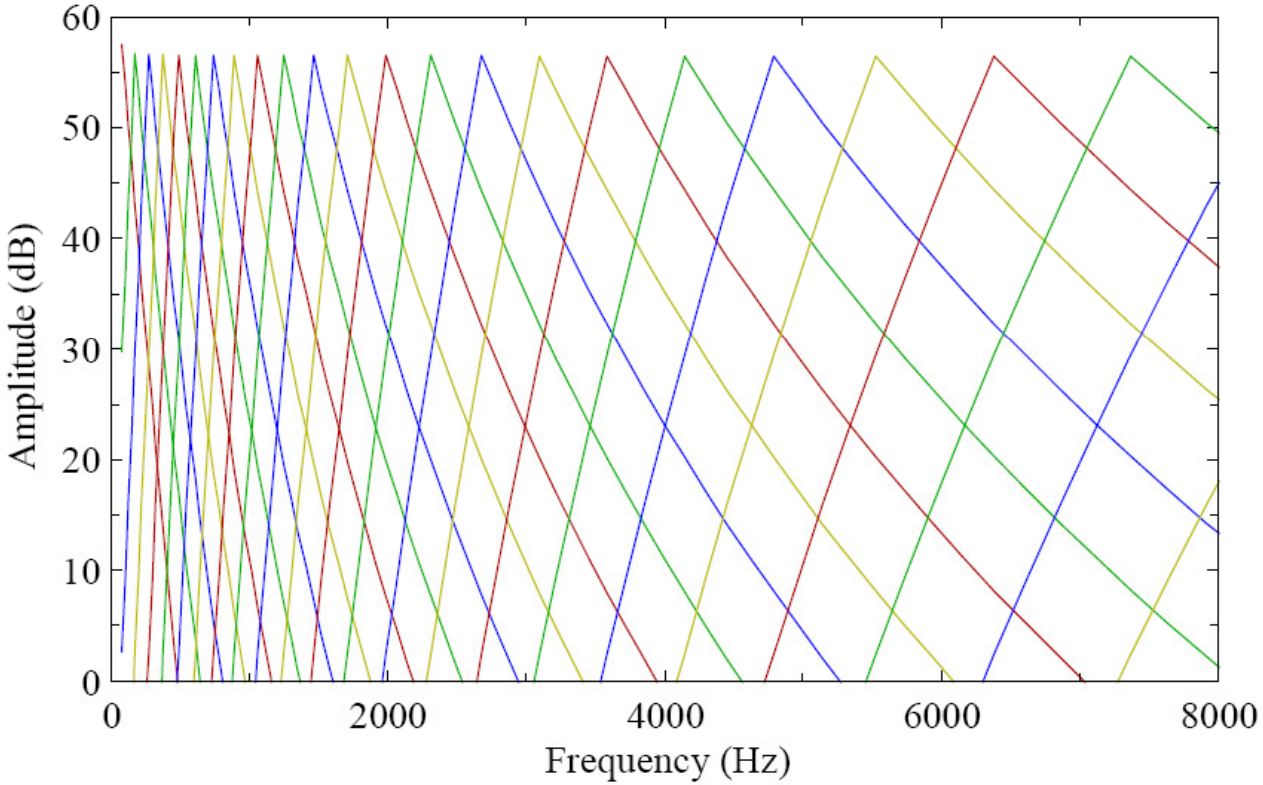


Fig. 4 Spreading functions (60 dB signal level). Every fourth spreading function is plotted for the Basic version of PEAQ.

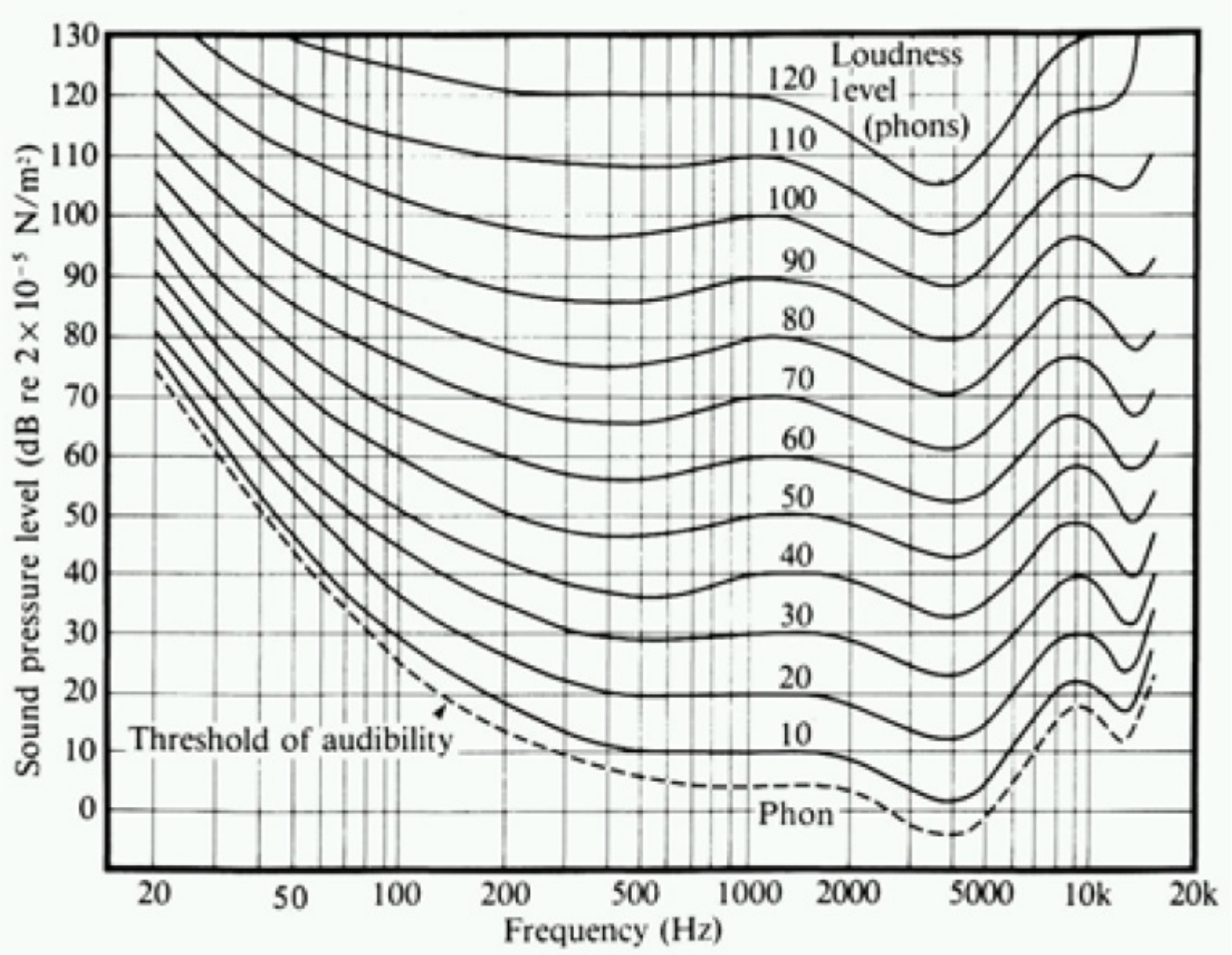


Figure 4: Equal loudness contour diagram.

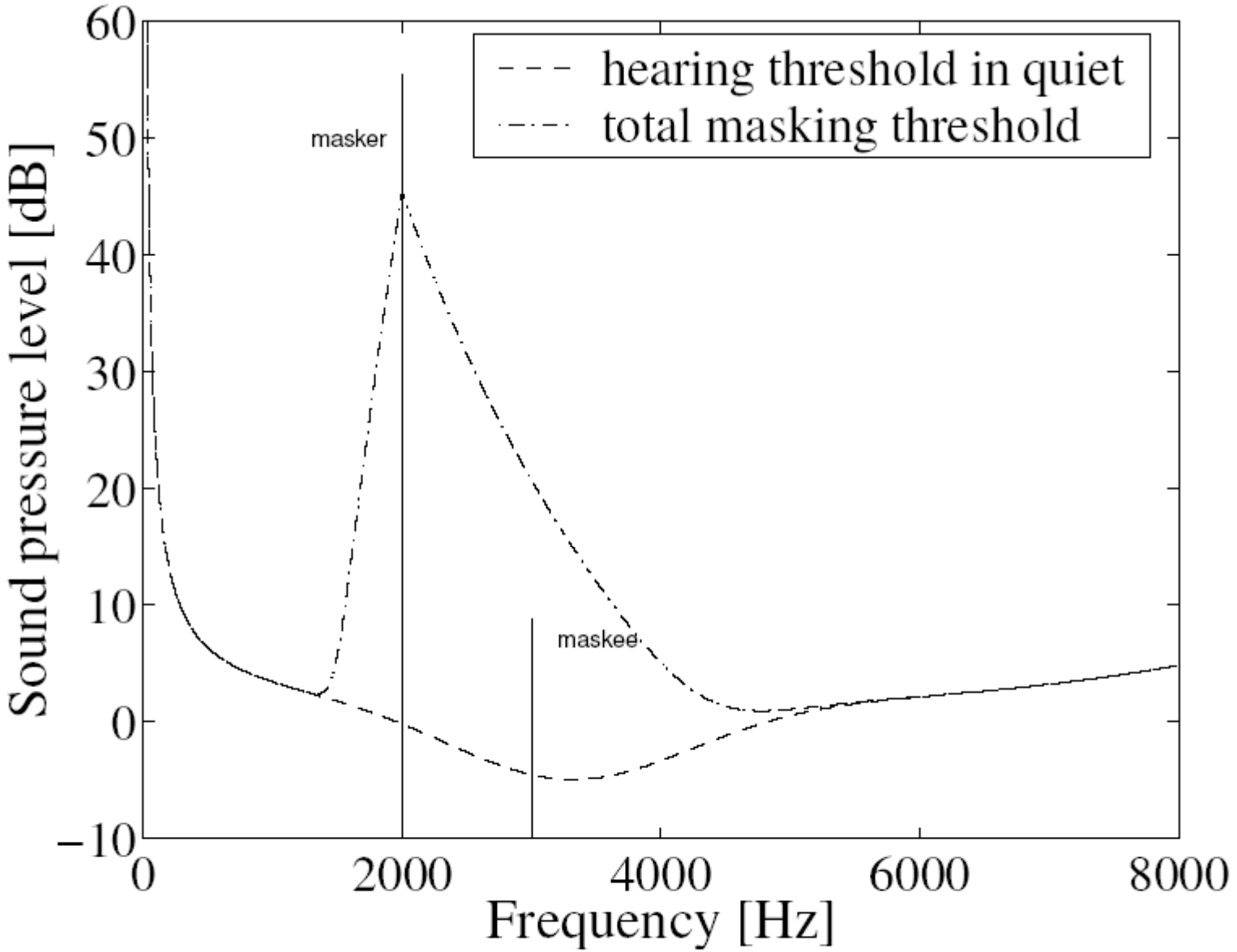


Figure 6: Simultaneous masking occurs when a strong tone makes the nearby tone inaudible.

The Filter Bank Ear Model

- The Advanced version of PEAQ uses a Filter bank ear model as well as the FFT-based model.
- **DC Rejection Filter**
  - to remove subsonic signal components.
- **Filter Bank**
  - bank uses bandpass filters at 40 centre frequencies ranging from 50 *to* 18000.02 *Hz. The centre frequencies are equally-spaced on the Bark scale*
- **Outer and Middle Ear Modeling**
- **Frequency Domain Spreading**
- **Backward Masking**
  - The frequency-spread energies are time-smeared with an FIR filter.
- **Internal Noise**
  - Internal noise is added to each band.
- **Forward Masking**

Filter Bank responses

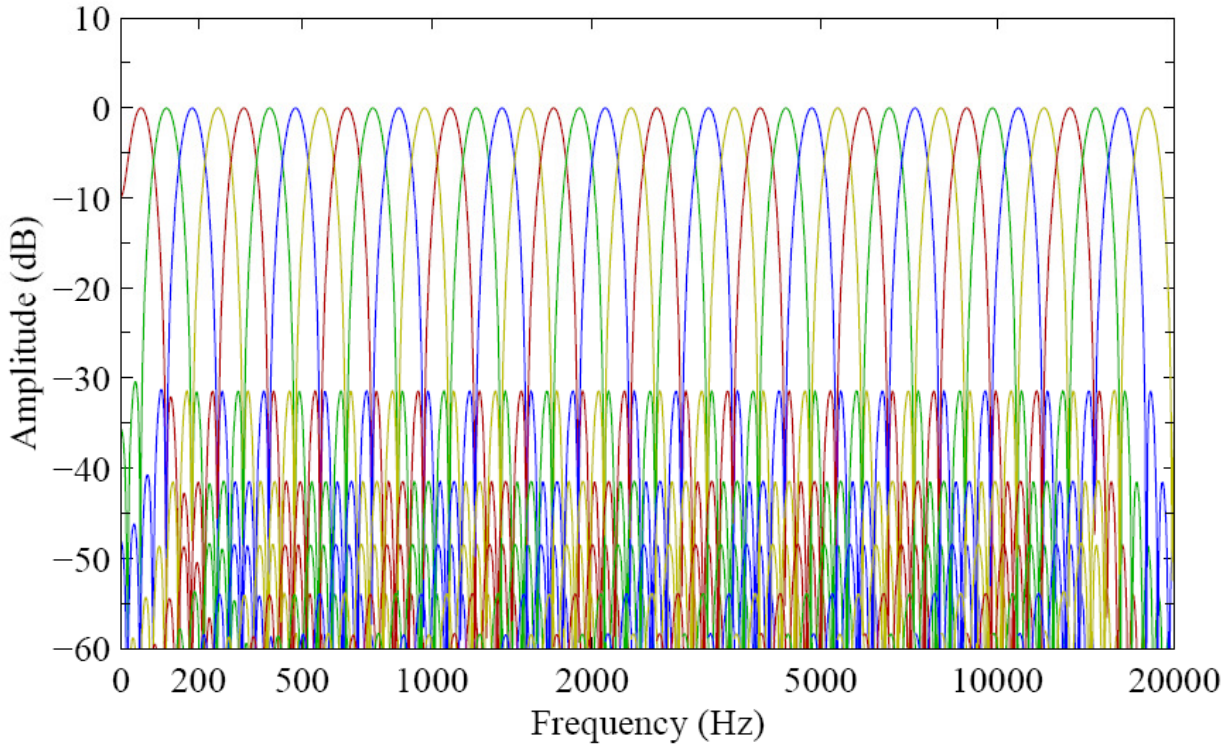


Fig. 7 Superimposed filter bank responses. The frequency axis is linear on a Bark scale.

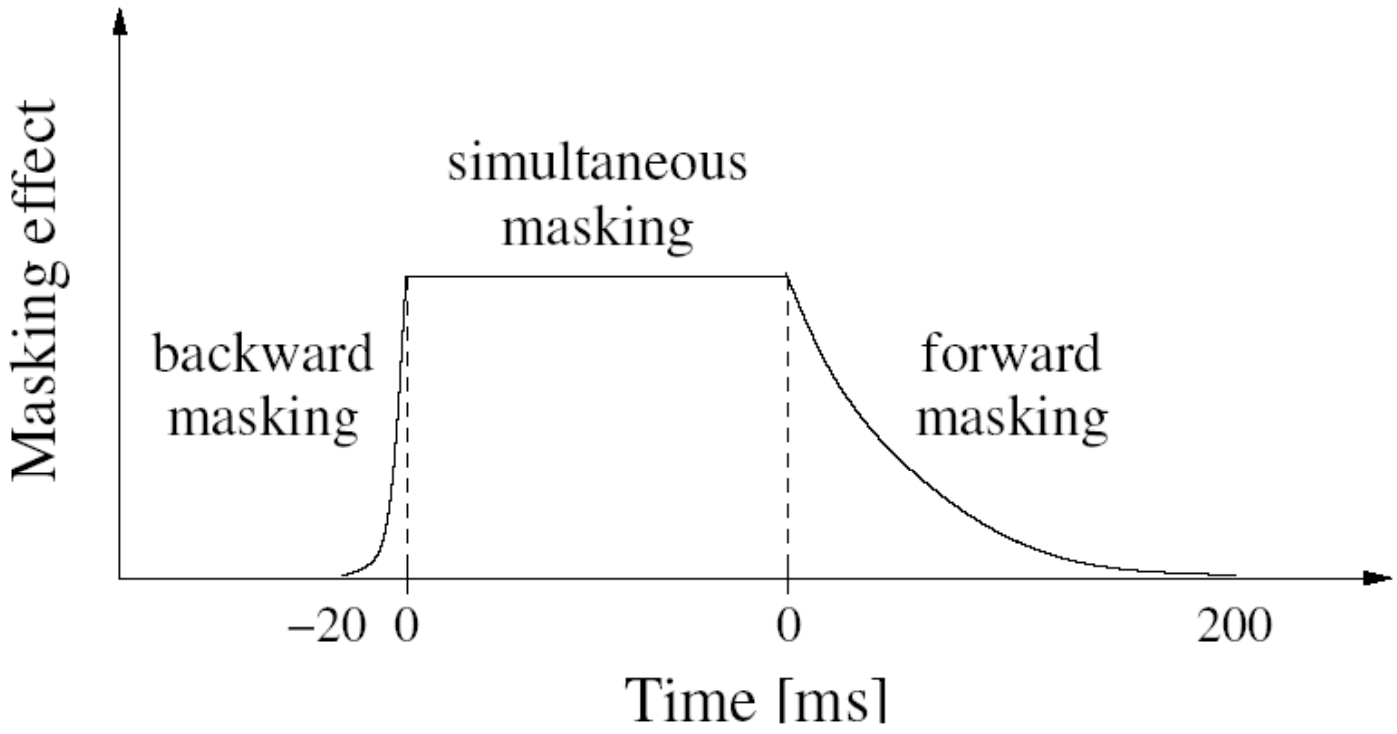


Figure 5: Non-simultaneous masking occurs before and after the masker.

Table 1 Processing parameters

PEAQ Version	Model	Sampling Rate F_{ss}	No. Centre Frequencies N_c
Basic	FFT	F_s /1024	109
Advanced	FFT	F_s /1024	55
	Filter Bank	F_s /192	40

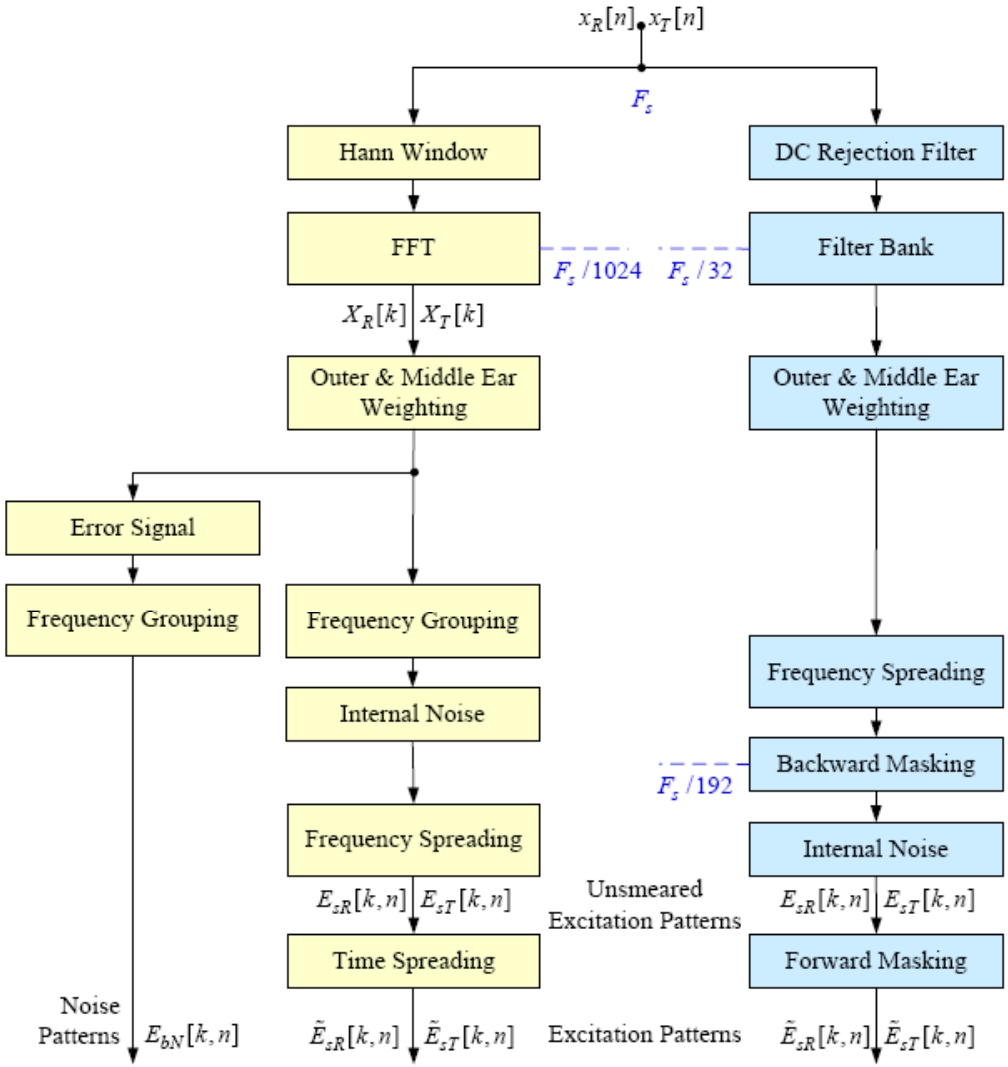


Fig. 8 Output signals from the FFT model and the filter bank model. Subscripts R and T will refer to signals derived from the reference and test signals, respectively.

Pattern Processing

- The outputs of the FFT and filter bank blocks are further processed.
- Let us consider only mono signals.
- **Excitation Pattern Processing:**
  - *Time Domain Spreading*
  - *Pattern Adaptation*
- **Modulation Pattern Processing**
  - to compute averages and average differences in an approximate loudness domain (0.3 power domain)
- **Loudness Calculation**

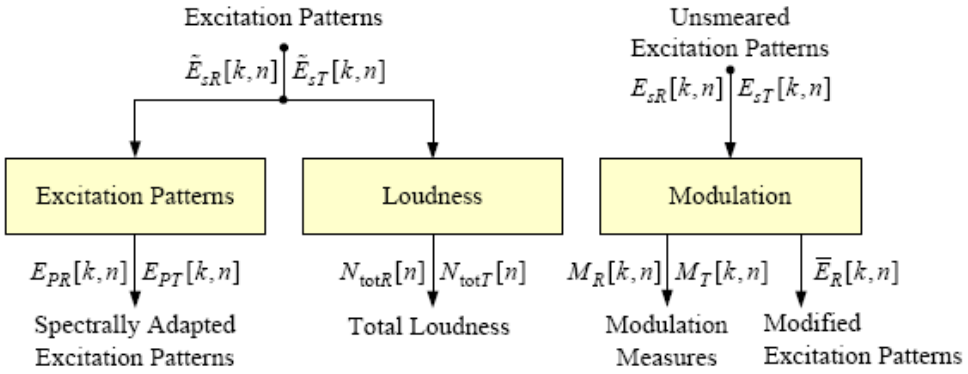


Fig. 10 Inputs and outputs from the pattern processing step.

Calculation of the Model Output Variables

- The outputs of the previous steps are generally functions of time and frequency for the reference signal and the test signal.
- These functions are distilled into functions of time.
- These functions of time are averaged to give a single value, the model output variable (MOV).

Table 4 Model Output Variables – PEAQ Basic

Model Output Variable	Model	Description
$BandwidthRef_B$	FFT	Bandwidth of the reference signal
$BandwidthTest_B$	FFT	Bandwidth of the test signal
$Total\ NMR_B$	FFT	Noise-to-mask ratio
$WinModDiff1_B$	FFT	Windowed modulation difference
ADB_B	FFT	Average block distortion
EHS_B	FFT	Harmonic structure of the error
$AvgModDiff1_B$	FFT	Average modulation difference
$AvgModDiff2_B$	FFT	Average modulation difference
$RmsNoiseLoud_B$	FFT	Distortion loudness
$MFPD_B$	FFT	Maximum filtered probability of detection
$RelDistFrames_B$	FFT	Relatively disturbed frames

Table 5 Model Output Variables – PEAQ Advanced

Model Output Variable	Model	Description
$RmsModDiff_A$	Filter Bank	Modulation changes
$RmsNoiseLoudAsym_A$	Filter Bank	Distortion loudness
$Segmental\ NMR_B$	FFT	Noise-to-mask ratio
EHS_B	FFT	Harmonic structure of the error
$AvgLinDist_A$	Filter Bank	Linear distortions

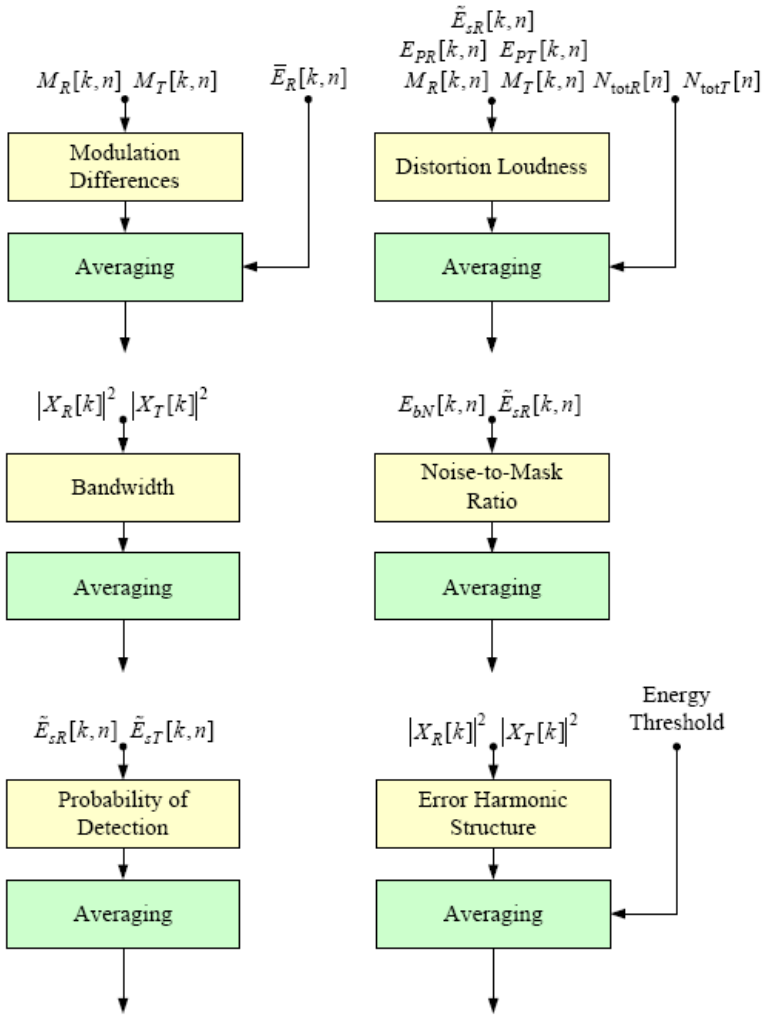


Fig. 11 Inputs to the model output variable calculations.

Calculation of the Objective Difference Grade

- MOV's will be combined using a neural network to give an objective difference grade (ODG)
  - ODG measures the degradation of the test signal with respect to the reference.
- The neural network has been trained to give good matches to the subjective impairment scale shown in the table.

Table 6 Model Output Variables – PEAQ Advanced

Difference Grade	Description of Impairments
0	Imperceptible
-1	Perceptible but not annoying
-2	Slightly annoying
-3	Annoying
-4	Very annoying

Neural Network – Basic Version

- neural network with 11 input nodes, 1 hidden layer with 3 nodes and a single output, the distortion index D.
  - I is the number of MOV's (11 for the Basic version) and J is the number of nodes in the hidden layer. The terms w are bias terms.

$$D_I = w_{yb} + \sum_{j=0}^{J-1}\left(w_y[j]\,\mathrm{sig}\left(w_{xb}[j] + \sum_{i=0}^{I-1} w_x[i,j]M_v'[i]\right)\right),$$

MUSHRA test: 32kbps

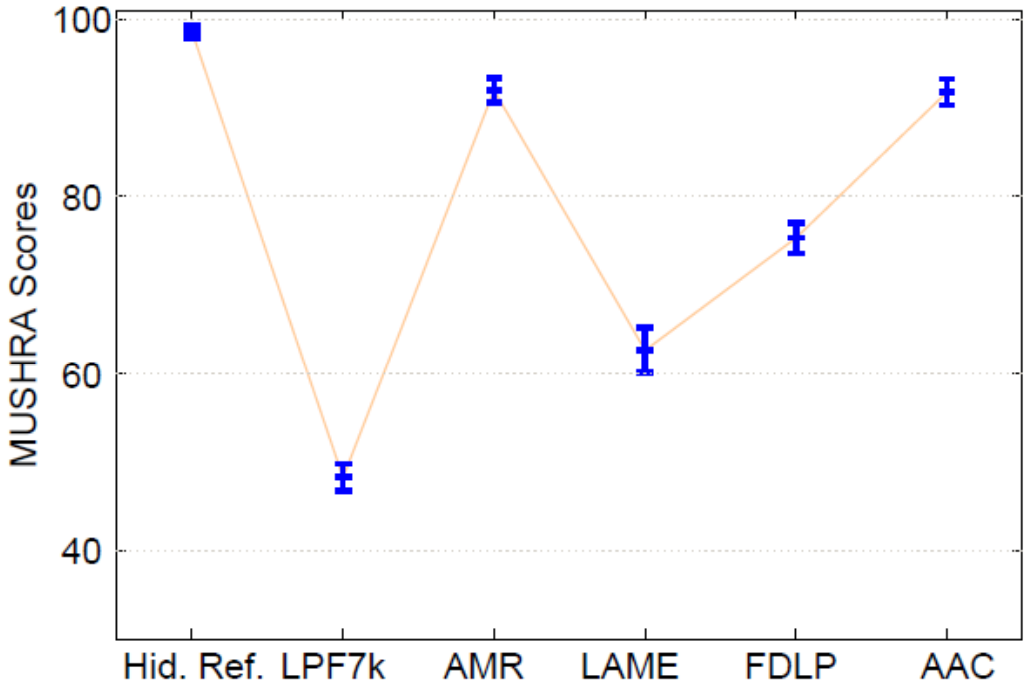
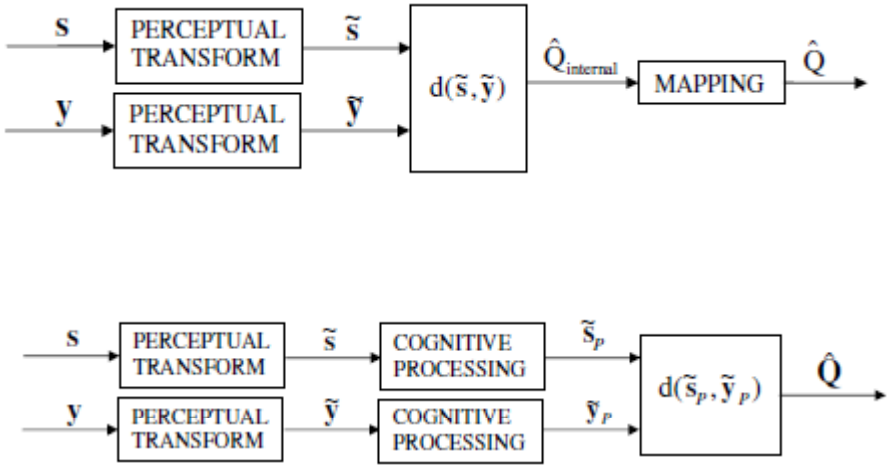


Fig. 10. MUSHRA results for 6 speech/audio samples using four coded versions at 32 kbps (AMR-WB+ (AMR), FDLP-MDCT (FDLP), MPEG-4 HE-AAC (AAC) and LAME-MP3 (LAME)), hidden reference (Hid. Ref.) and 7 kHz low-pass filtered anchor (LPF7k).

Cont.



• Desired and current approaches: exploit knowledge of the human auditory system To weight more the error signal in regions where it is more audible
• more audible does not mean more objectionable, since the latter is dependent on human brain processing