COMP47700
Speech and Audio

Dr Alessandro Ragano

Office: 3rd Floor, Science East, Insight Centre
School of Computer Science
University College Dublin

alessandro.ragano@ucd.ie

Recap: Units 1 – 5 (Weeks 1 – 6)

1. Introduction / Digital Representation of Sound
2. Basic Audio Processing
3. Speech Production
   - Human speech apparatus
   - Characteristics of speech
   - Speech Understanding
4. Speech Perception and Mechanics of Hearing
   - Ear mechanics (outer, middle, inner)
   - Sound pressure wave to understanding in the brain
   - Wider brain function (other senses interacting)
5. Speech Intelligibility and Quality
   - Intelligibility vs Quality
   - Subjective and Objective Testing

Recap: PL 1–5

1. Introduction to libraries and audio processing
2. Visualisation
3. Voice and Pitch analysis
4. Audio Degradations
5. Signal analytics (Speech Intelligibility/Quality) (tomorrow)

Next 6 Weeks

- Week 7: Audio Analytics and Data Features - 05/03/2025
- Week 7: MCQ Test + Project Description - 06/03/2025
- Applied topics after Study Break: ASR, Psychoacoustics and Audio Coding, Music Information Retrieval, Sound Source Separation, Spatial Audio.
- Labs: PL6 - Machine Learning / Automatic Speech Recognition
- Labs in Week 9 and 11 will be for project development and progress review.
- Week 12: Project demos

5.1
Speech Intelligibility and Quality

Intelligibility vs Quality

Intelligibility

"A measure of how comprehensible speech is in given conditions."

Necessary but not **sufficient** for quality

More **specific** measure

Intelligibility is important in **low quality scenarios**

(e.g. very noisy or reverberant)

Intelligibility vs Quality

Intelligibility

"A measure of how comprehensible speech is in given conditions."

Necessary but not **sufficient** for quality

More **specific** measure

Intelligibility is important in **low quality scenarios** (e.g. very noisy or reverberant)

Quality

Umbrella term: pleasantness, listening effort, intelligibility, acceptability

Related to intelligibility (low intelligibility **also** poor quality)

More **subjective**

Impact of listener (context, experience, expectation, mood)

Quality of Experience is important

Cognitive Listening Effort

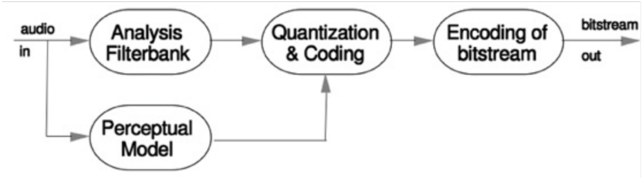
Intelligibility vs Quality

Why are they **useful** things to measure or predict? Noise removal

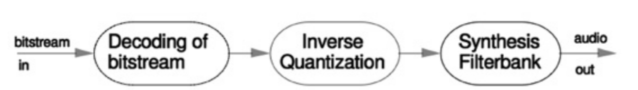
Intelligibility vs Quality

Why are they useful things to measure or predict? Audio Coding

Encoder e.g. sender in a telephone call, original lossless music recording



Decoder e.g., receiver in a telephone call, youtube or spotify audio track



Goal: Minimize bitstream size.
Constraint: Maintain acceptable audio quality.

Intelligibility vs Quality

Why are they useful things to measure or predict? Speech Synthesis

Speech Intelligibility

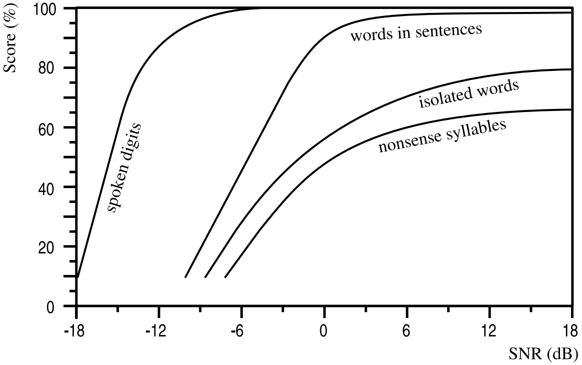
- **Diagnostic rhyme test (DRT):** distinguish between two words rhyming by initial, such as freak, leak
- **Modified rhyme test (MRT):** select one of six words, half differing by initial and half by final, such as cap, tap, rap, cat, tan, rat
- **Phonetically balanced word lists:** 50 sentences of 20 words each, and asking them to write down the words they hear;
- **ICAO spelling alphabet (ALFA, BRAVO, CHARLIE,?) test:** determine the kinds of confusions that occur under severely degraded digital voice conditions

Speech Intelligibility

Contextual information, redundancy and vocabulary size

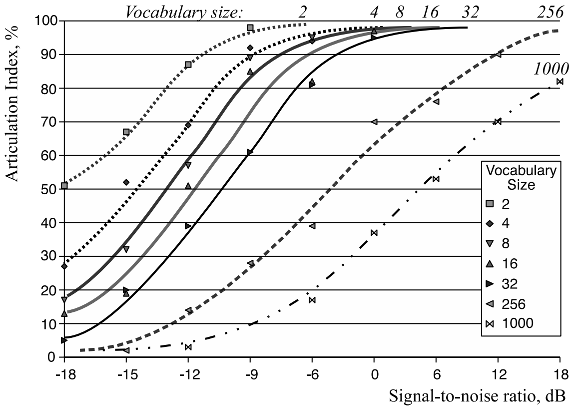
What is effect on the intelligibility of speech?

Context: Are the words familiar, isolated, in a sentence or part of a closed vocabulary



Vocabulary effect on intelligibility

Impact of Vocabulary Size on Intelligibility related to Noise



Speech Quality

How do you rate the quality of speech?

Much harder because it is more subjective

Mean Opinion Scores (MOS)

Score	Description	Impairment
5	excellent	imperceptible
4	good	perceptible but not annoying
3	fair	slightly annoying
2	poor	annoying
1	bad	very annoying

Absolute Category Rating (ACR)

MOS is an ACR range – categorical rather than continuous

Standard Protocols

Example Standards: National or International

What standards organisations can you think of?

Standard Protocols

Example Standards: National or International

What standards organisations can you think of?

Standards-based testing

- **Video:** VQEG
- **Audio (and some video)**: International Telecommunication Union
  - United Nations specialised agency for information and communication technologies.
  - P series: Terminals and subjective and objective assessment methods
  - ITU-T P.800

http://www.itu.int/ITU-T/recommendations/index.aspx?ser=P

Purpose of Standards

Why might we use standards for testing?

Purpose of Standards

Why might we use standards for testing?

For subjective testing methodology

- Reliability
- Repeatability
- Statistical Significance

Controlling variables

e.g. background noise, human variables, equipment variables

Subjective Audio Listener Tests

Speech and Audio are tested differently

Not just one standard

ACR Speech MOS tests (Rec. ITU-R P.800)
Small impairments in audio systems (Rec. ITU-R BS.1116-1)
Intermediate quality level of coding systems (Rec. ITU-R BS.1534-1)

Example Audio Test

MUSHRA (Rec. ITU-R BS.1534-1)

"**MU**lti **S**timulus test with **H**idden **R**eference and **A**nchor" Hidden Reference/Anchor(s)
Ranking relative to reference and other treatments
Continuous Quality Scale (unlike the Absolute Category Rating in MOS)

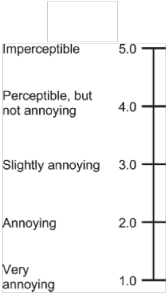


Fig. 1. ITU-R impairment scale [4], [19].

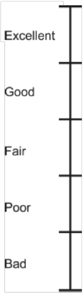
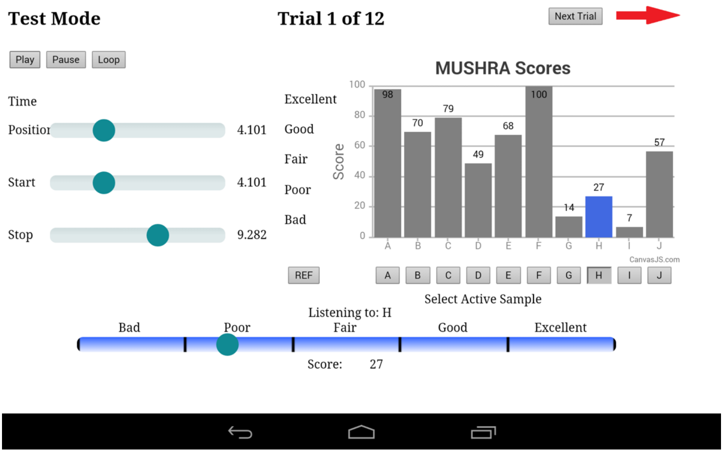


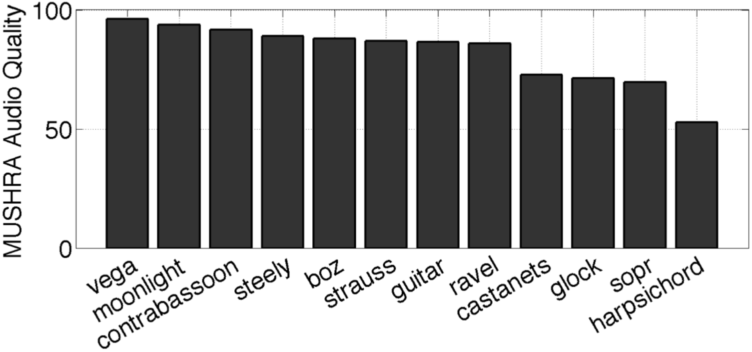
Fig. 2. ITU-R quality scale [5][19].

Source: Zielinski, S., Rumsey, F., & Bech, S. (2008). On some biases encountered in modern audio quality listening tests-a review. Journal of the Audio Engineering Society, 56(6), 427-451.

MUSHRA Test GUI



MUSHRA Test: MP3 Audio Codec at 96 kb/s



Source: A. Hines, E. Gillen, J. Skoglund, D. Kelly, A. Kokaram, N. Harte. Perceived Audio Quality for Streaming Stereo Music. In ACM Multimedia, Orlando, FL, USA, 2014

Subjective Testing Biases

Four examples of bias

1. Picking the right test data (choice of stimuli)
2. Test setup (choice of methodology)
3. Repeatability (e.g. across locations)
4. Language (unintended differences due to labels)

Stimulus Bias

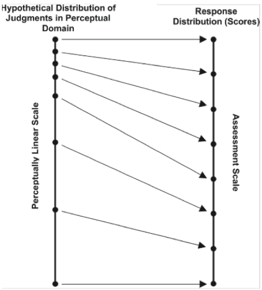


Fig. 5. Stimulus spacing bias model. (Adapted from [8].)

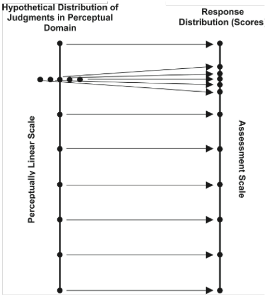


Fig. 6. Stimulus frequency bias model. (Adapted from [8].)

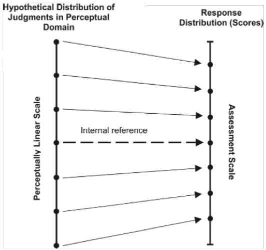
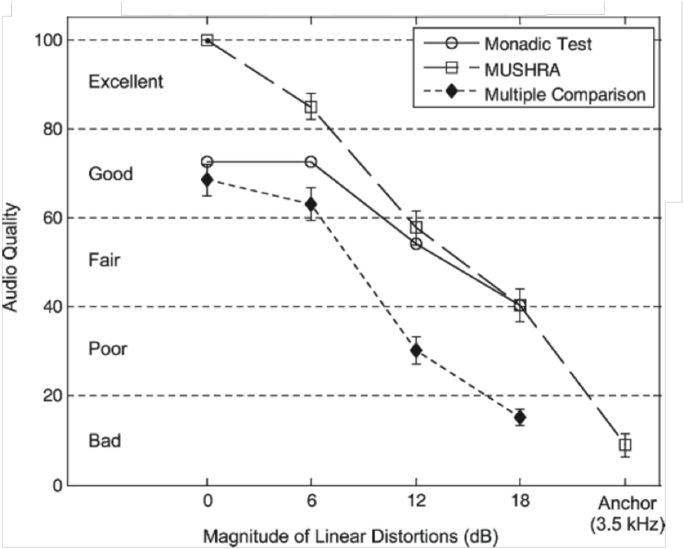


Fig. 7. Contraction bias model. (Adapted from [8].)

Source: Zielinski, S., Rumsey, F., & Bech, S. (2008). On some biases encountered in modern audio quality listening tests-a review. Journal of the Audio Engineering Society, 56(6), 427-451.

Test Method Bias



Source: Zielinski, S., Rumsey, F., & Bech, S. (2008). On some biases encountered in modern audio quality listening tests-a review. Journal of the Audio Engineering Society, 56(6), 427-451.

Language Label Bias

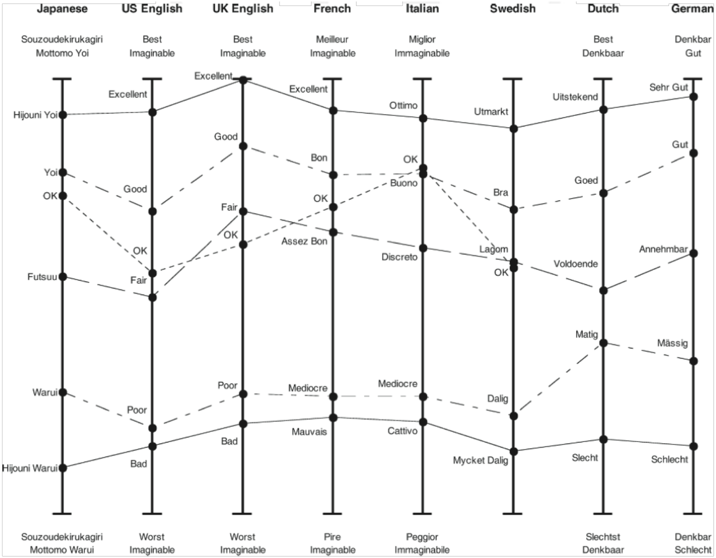


Fig. 12. Combined results of scaling of labels used in quality evaluation. (Data taken from [7], [26], [59]–[62]; see text for details.)

Source: Zielinski, S., Rumsey, F., & Bech, S. (2008). On some biases encountered in modern audio quality listening tests-a review. Journal of the Audio Engineering Society, 56(6), 427-451.

OK is not ok!

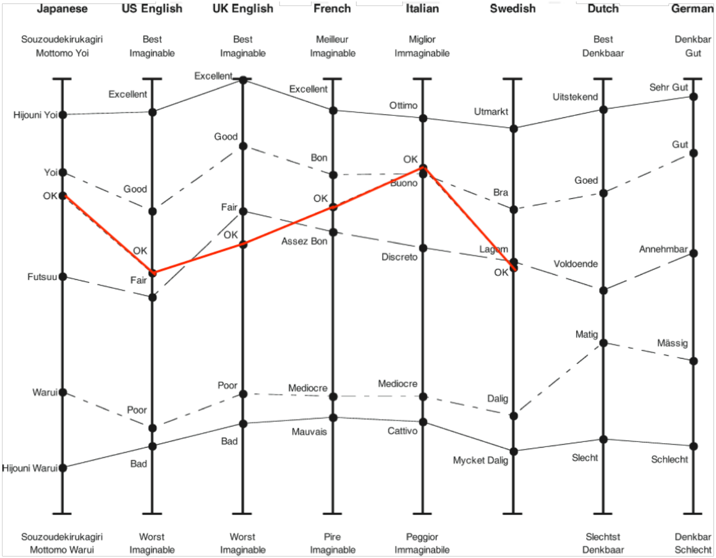


Fig. 12. Combined results of scaling of labels used in quality evaluation. (Data taken from [7], [26], [59]–[62]; see text for details.)

Subjective Testing Biases

Four examples of bias

1. Picking the right test data (choice of stimuli)
2. Test setup (choice of methodology)
3. Repeatability (e.g. across locations)
4. Language (unintended differences due to labels)

Reflecting on biases

What can we do other than be aware and try to minimise biases?

Reflecting on biases

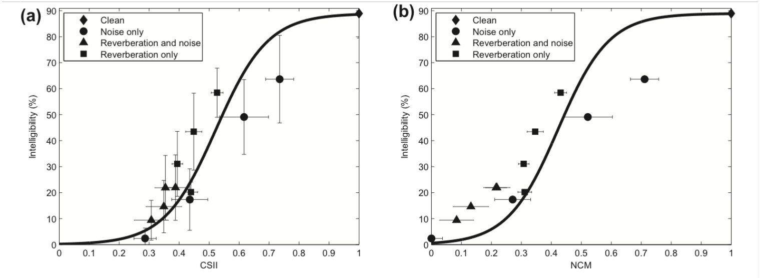
What can we do other than be aware and try to minimise biases?

Reflecting on biases

What can we do other than be aware and try to minimise biases?

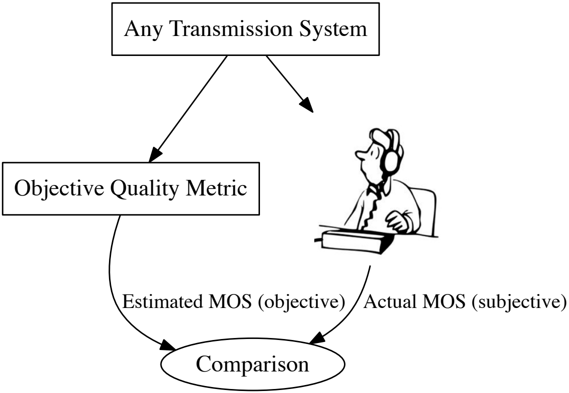
- Ground truth may be more noisy than you imagine
- If you are looking to predict ground truth, you may need to look at errorbars in x- and y- axes
- Acknowledge the impact they may have on your model (if your training data is noisy...)

Example: X or XY errorbars



Source: Santos, J. F., Cosentino, S., Hazrati, O., Loizou, P. C., & Falk, T. H. (2013). Objective speech intelligibility measurement for cochlear implant users in complex listening environments. Speech communication, 55(7), 815-824.

Objective Speech Quality Testing



Objective measures

1. **Application:** Plan, Optimise, Monitor, Maintenance
2. **Signal type:** Narrowband (NB)/Wideband (WB)/Super Wideband (SWB)
3. **Channels:** monaural/binaural/spatial
4. **Test source:** parameter, simulation, measurement
5. **Inputs:** params, ref & test sig, only test

"Horses for Courses"

Match the Application to the Model

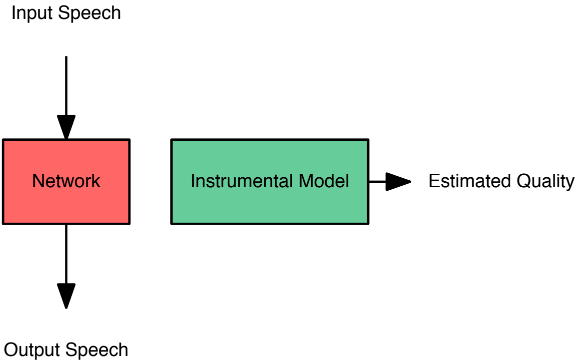
Further Reading: S. Moller, W. Y. Chan, N. Cote, T. H. Falk, A. Raake, and M. Waltermann, âSpeech quality estimation: Models and trends,â IEEE Signal Processing Magazine, vol. 28, pp. 18â28, 2011.

Objective Model Types

Network Channel and Model

Speech Signal Input and Output
Channel Degradation (Network) and Quality Model

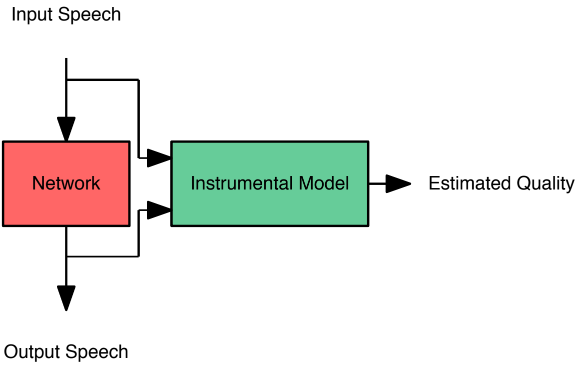
Figure: Adapted from Möller, 2011



Objective Model Types

Full-Reference Signal-Based Model

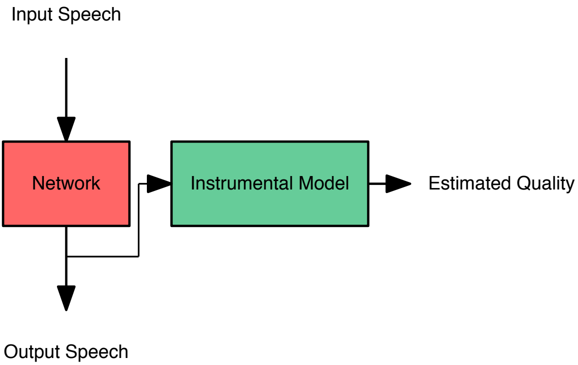
e.g. PESQ, POLQA, ViSQOL
(ITU-T, 2001; ITU-T, 2009; Hines et al. 2013)



Objective Model Types

No-Reference Signal-Based Model

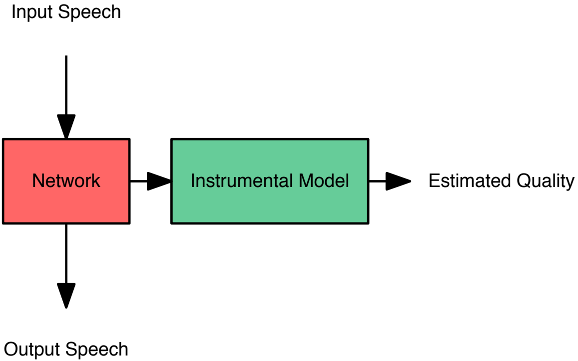
e.g. P.563, ANIQUE+, LCQA
(ITU-T, 2004; ANSI,2006; Grancharov , 2006)



Objective Model Types

Parametric Signal-Based Model

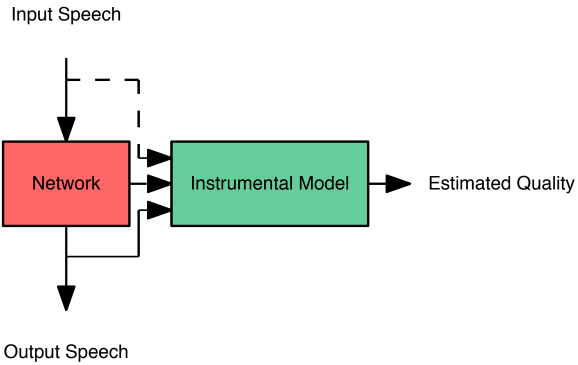
Parametric Model
e.g. E-Model (ITU-T, 2009)



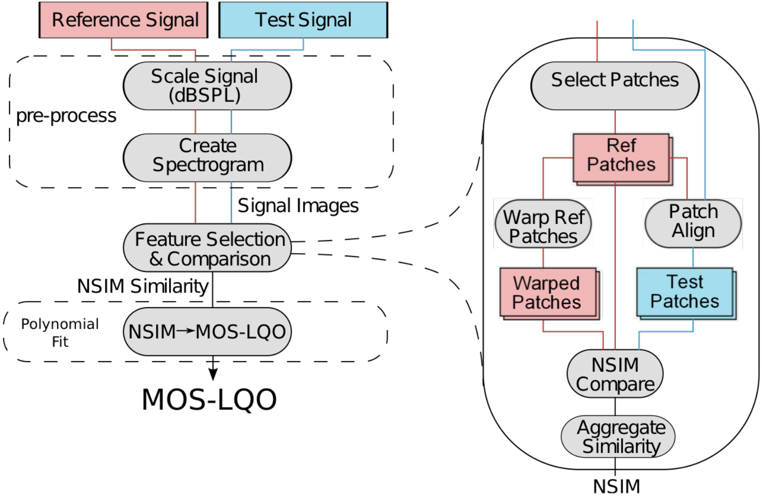
Objective Model Types

Hybrid Model

Parameter and Signal-Based (one or two) Model
Using combination of design information and real signals



Example Speech Quality Metric: ViSQOL



A. Hines, J. Skoglund, A. Kokaram, N. Harte. Robustness of Speech Quality Metrics to Background Noise and Network Degradations: Comparing ViSQOL, PESQ AND POLQA. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 3697–3701, 2013.

Simple Measures

Mean-Squared Error

A crude, time domain quality metric can be computed using the MSE of a reference and test signal.

Assuming they are the same length, N samples the difference between the reference signal, p, and the degraded signal, q, is computed for each sample, t, as:

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(p_t - q_t)^2.$$

or in Python, using numpy ndarrays for signals `P` and `Q`:

```
mse = ((P - Q) ** 2).mean(axis=ax)
```
[1]

[1] ax=0 for columnwise, ax=1 for rowwise

SNR

Signal to Noise Ratio (SNR)

- Compares the original and processed speech signals sample by sample
- Measures the distortion of a reference signal, p and a noisy or degraded signal q.

It is calculated as:

$$SNR = 10 \log_{10} \frac{\sum_{i=1}^{N} p^2(i)}{\sum_{i=1}^{N} (p(i) - q(i))^2}$$

Python Syntax

```
overall_snr = 10 * np.log10(np.divide(
np.sum(np.power(P, 2)), np.sum(np.power(P - Q +
np.spacing(1), 2))))
```

Issues with basic MSE/SNR for quality estimation

Regression to mean for long signals – errors averages out

Use a segmental approach, 20–30 ms segments (perhaps overlap and add)

Other Weaknesses

Handling: alignment, normalisation
Perceptual Relevance of changes in time domain
Simple measures do not correlate well with subjective scores for compression codecs

Codecs

Codecs making use of human auditory perceptual factors for network and signal optimisation (more on this in unit 9)
The signal may be different but does it sound different?

Better Measures?

Temporal

Segmental SNR, Segmental MSE
Break the signal into 20-30 ms segments (or overlapping segments) and calculate the SNR or MSE per segment.

Spectral

Spectral distortion, Linear Predictive Coding (LPC), Log-Likelihood Ratio (LLR)

Complex Measures

PESQ/POLQA/VISQOL etc.

Summary

Speech Intelligibility and Quality

- Intelligibility vs Quality
- Subjective Testing: why and how
- Objective Testing Models