

Lecture 23: Perceptual Measures of Quality, Similarity, and Distance

Mark Hasegawa-Johnson

ECE 537: Speech Processing Fundamentals

- 1 Subjective Measures
- 2 Objective Measures: Signal Embedding
- 3 Objective Measures: Regression or Classification of Subjective Labels
- 4 Summary

Outline

- 1 Subjective Measures
- 2 Objective Measures: Signal Embedding
- 3 Objective Measures: Regression or Classification of Subjective Labels
- 4 Summary

Subjective measures of audio quality

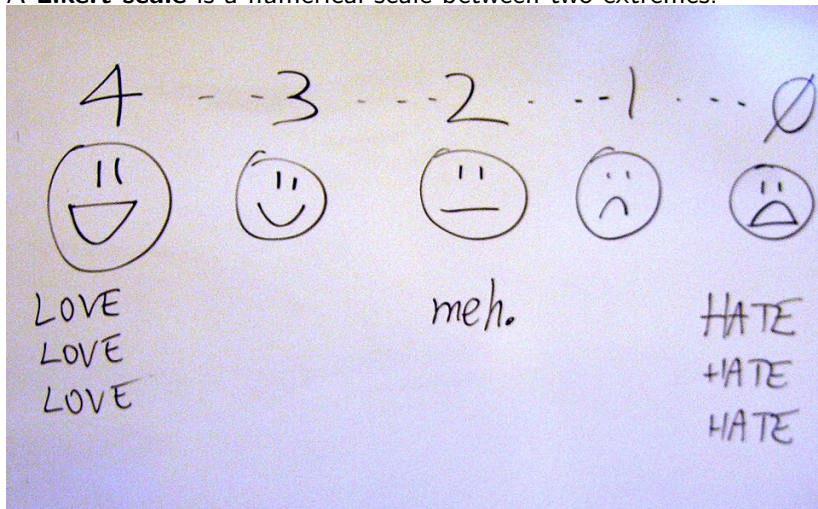
- A measure is called **subjective** if it requires responses from one or more human listeners.
- Subjective perceptual tests can be broadly classified into:
 - **Categorical** tests ask human listeners to respond with one or more categorical labels
 - **Gradient** tests ask listeners to score audio on a numerical scale between two extremes

Categorical tests

- **Same/different** and **ABX** tests determine whether listeners can hear any difference between two waveforms.
 - ABX: Play three waveforms, ask if the middle one was equal to the first or the last. Avoids person-dependent same vs. different biases of simple same/different tests.
- **Intelligibility** is most reliably measured by the word error rate of human listeners.
 - Key parameters: Do listeners know the voice or not? The language? The accent?
- **Cohen's kappa** measures the reliability with which human listeners can be trained to make a novel categorical distinction.
 - Most frequent use: You want labeled ground truth data for some perceptual distinction, but first you want to prove that humans can make the same distinction.

Gradient measures

A **Likert scale** is a numerical scale between two extremes:



Gradient measures

- Common gradient scales: Naturalness, Intelligibility, Nativeness, Speaker similarity
- Methods for improving reliability:
 - **MOS** (Mean opinion score): Play each waveform to the same set of listeners, and average their scores
 - **anchors**: Examples of good and bad stimuli are provided.
 - **MUSHRA** (Multiple stimuli with hidden reference and anchor): Two waveforms, the **reference** and **anchor**, are known to be best and worst. Listeners are not told which they are. Each listener's scores are normalized w.r.t. their reference and anchor.

Outline

- 1 Subjective Measures
- 2 Objective Measures: Signal Embedding
- 3 Objective Measures: Regression or Classification of Subjective Labels
- 4 Summary

Objective measures of audio quality

- A measure is called **objective** if it can be computed by an algorithm
- Advantage: Much cheaper if you don't have to hire humans
- Stages in most objective tests:
 - **Embedding**: convert acoustic signal into some representation in which simple distance measures (e.g., Euclidean) are more closely correlated with perceptual distance
 - **Regression** or **classification**: A neural net is trained to reproduce the numerical or categorical labels provided by human listeners on some training corpus.

Common stages in a signal-processing embedding

- ① Time-frequency representation
- ② Frequency warping
- ③ Amplitude compression
- ④ Simultaneous masking
- ⑤ Nonsimultaneous masking

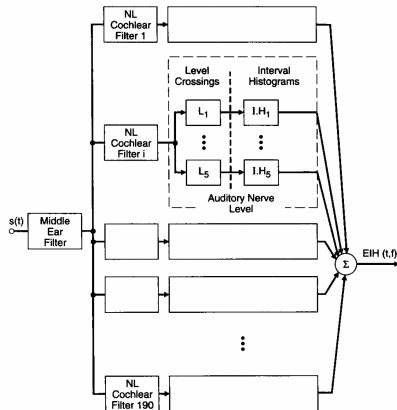
1. Time-frequency representation

Time-frequency encoding models the mechanical frequency selectivity of the basilar membrane using either sub-band filtering or a short-time transform:

$$x(t, f) = h_f(t) * x(t)$$

... or ...

$$x(t, f) = \mathcal{F} \{ w(t - \tau) x(\tau) \}$$



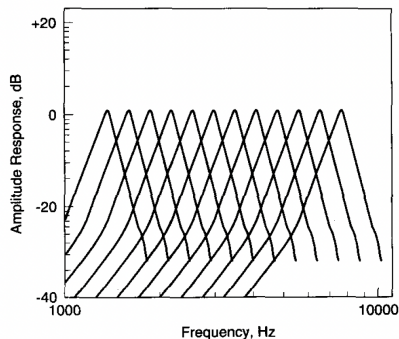
Ghitza, 1994, Fig. 3

2. Frequency warping

- Bandpass filter center frequencies are equidistant in a physiological or perceptual scale, or...
- Fourier transforms are converted from f (Hertz) to ϕ (perceptual frequency) according to

$$x(t, \phi) = \int h_{\phi}(f) J(x(t, f)) df$$

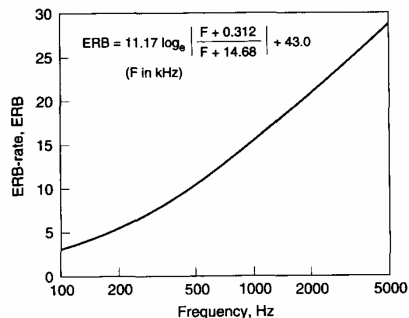
...where $J(\cdot)$ is usually amplitude or energy.



Ghitza, 1994, Fig. 6

Common frequency scales include:

- Millimeters on the basilar membrane
- Equivalent rectangular bandwidths (1 ERB=bandwidth of noise that masks a tone)
- Mel-frequency (1 mel=spacing between two tones whose pitches are as dissimilar as 1000Hz vs. 1001Hz)
- Logarithmic or Constant-Q (1 semitone = 1/12 octave)



Ghitza, 1994, Fig. 9

3. Amplitude compression

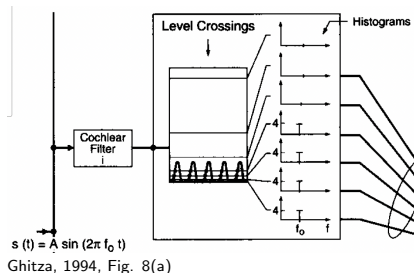
- **Timing ignored:**

Short-time energy is computed, then compressed to model loudness, i.e.,

$$y(t, \phi) = G(|x(t, \phi)|^2)$$

... where $G(\cdot)$ may be cube root or log.

- **Timing retained:** Interval histogram detects positive level crossings of nonlinearly-spaced thresholds

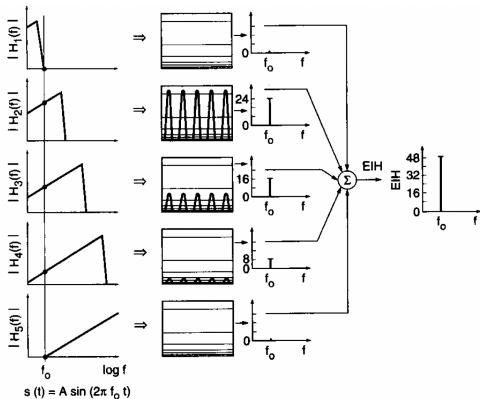


4. Simultaneous Masking

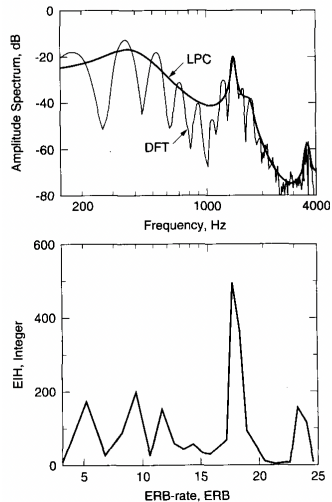
Loud sounds mask quiet sounds at nearby frequencies if they occur at the same time. Some implementations:

- **Perceptual warping filter** is used by multipulse LPC, CELP, and other low-bit-rate speech coders
- **Perceptual LPC (PLP)** models the warped-frequency, warped-amplitude spectrum $y(t, \phi)$ using 5th-order LPC. Thus only two peaks are modeled; all smaller peaks are ignored.
- **The ensemble-interval histogram (EIH)** accumulates a histogram of inter-spike intervals, $1/f$, across cochlear bands ϕ and threshold levels k . In this representation, a strong tone will dominate all cochlear bands in which it appears, thus masking others:

$$\text{EIH}(t, f) = \sum_{\phi} \sum_k y_k(t, \phi) \times y_k\left(t - \frac{1}{f}, \phi\right)$$



Ghitza, 1994, Fig. 8(b): EIH algorithm, showing the dominance of many neighboring cochlear bands by the same high-energy pure tone.



Ghitza, 1994, Fig. 15: EIH result, showing masking of low-amplitude spectral components by higher-amplitude components at nearby frequencies.

5. Nonsimultaneous Masking

- Loud sounds at time t can mask quiet sounds at $t + \epsilon$, or even, if the amplitude difference is large and ϵ is small, at $t - \epsilon$.
- Few DSP models represent this effect (unlike simultaneous masking), but one exception is. . .
- **RASTA Processing of Speech** (RelAtive SpecTral Amplitude, Hermansky and Morgan, 1994): The nonlinear-frequency, amplitude-compressed spectral representation $y(t, \phi)$ is filtered as

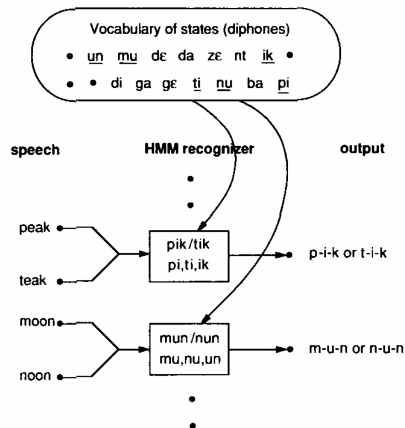
$$z(t, \phi) = r(t) * y(t, \phi)$$

... where $r(t)$ is a bandpass filter with a lowpass cutoff of 0.26Hz (unchanging background=inaudible) and a highpass cutoff of 12.8Hz (high-frequency flutter=inaudible).

Outline

- 1 Subjective Measures
- 2 Objective Measures: Signal Embedding
- 3 Objective Measures: Regression or Classification of Subjective Labels
- 4 Summary

Modeling of subjective classifications: Ghitza (1994) validated the EIH model by fitting a number of subjective test results, e.g., two-alternative-forced-choice (2AFC) phoneme classifications such as the diagnostic rhyme test (DRT).



Ghitza, 1994, Fig. 11

Modeling of Likert scales: Predictions of MOS by EIH (Fig. 13) are more accurate than those using cepstral distance (Fig. 14) for low-delay speech coders (PCM, ADPCM, LDCELP), but not for high-delay coders (CELP).

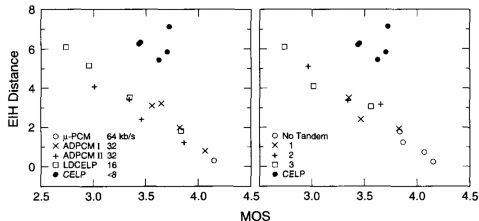


Fig. 13. EIH-distance versus MOS. The results are grouped according to coder identity (left) or according to tandem configuration (right). A particular coding condition can be located by a cross-alignment of the left and the right plots.

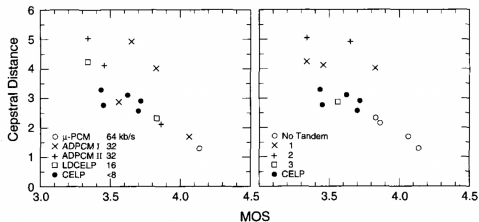


Fig. 14. Cepstral-distance versus MOS. Figure legend is as in Fig. 13.
Ghitza, 1994, Figs. 13-14

Objective modeling of subjective tests



<https://sites.google.com/view/voicemos-challenge>

- The VoiceMOS challenge (2022, 2024, 2026) benchmarks objective models of subjective auditory tests
- UTMOS (University of Tokyo MOS) is the heavily-cited open-source winner of VoiceMOS 2022 (doi://10.21437/Interspeech.2022-439)

Outline

- 1 Subjective Measures
- 2 Objective Measures: Signal Embedding
- 3 Objective Measures: Regression or Classification of Subjective Labels
- 4 Summary

Summary

- **Subjective** tests (classification like ABX or gradient like MOS) require a human listener.
- **Objective** tests usually include a signal processing embedding that models aspects of the auditory system including perhaps time-frequency encoding, perceptual frequency warping, amplitude compression, and masking
- Recent research on **objective modeling of subjective scores** includes VoiceMOS Challenge, UTMOS algorithm