University of Colorado
Boulder

Applying Intelligibility and Quality Metrics to Noisy Speech, Noise Suppression, and Hearing Aids

James M. Kates
James.Kates@colorado.edu

Kathryn H. Arehart
Kathryn.Arehart@colorado.edu
University of Colorado, Boulder, CO 80309

Clarity-2021, 16 September 2021

Colleagues and Collaborators

- Melinda Anderson
- Ramesh Kumar Muralimanohar
- Emily Lundberg
- Naomi Croghan
- In-Ki Jin
- Kristen Sommerfeldt
- Song-Hui Chon
- Lewis O. Harvey, Jr.
- Pam Souza (Northwestern U.)
- Varsha Rallapalli (Northwestern U.)

Hearing-Aid Processing

- Typical hearing-aid design
  - Multichannel filterbank
  - Time-varying gain adjustments in each frequency band
  - Gain can improve audibility, but amplitude modulation introduces nonlinear distortion
- Metrics measure signal changes
  - Envelope important for speech intelligibility and quality
  - Determine degree to which envelope is modified
  - Can also add other signal features: TFS, spectral change
  - Want interaction of all hearing-aid signal processing algorithms
  - Context of the auditory periphery and hearing loss

Quantitative Metrics

- Intrusive
  - Compare degraded signal to clean reference
  - Any change in the degraded signal is considered detrimental
  - Degradation includes effects of processing and auditory threshold
- Non-intrusive
  - Uses degraded signal alone
  - Requires machine model of perception
- This presentation deals with intrusive metrics
  - Hearing Aid Speech Perception Index (HASPI): intelligibility
  - Hearing Aid Speech Quality Index (HASQI): speech quality

Metric Construction

- Components
  - Model of the auditory periphery
  - Speech feature extraction
  - Map features to human subject data
- Training data
  - Metric tied to the speech materials and signal degradations used to train it
  - Sentences different mapping than keywords correct
  - Low data-rate codecs differ from additive noise
  - Extrapolating beyond the training data may be inaccurate

University of Colorado
Boulder

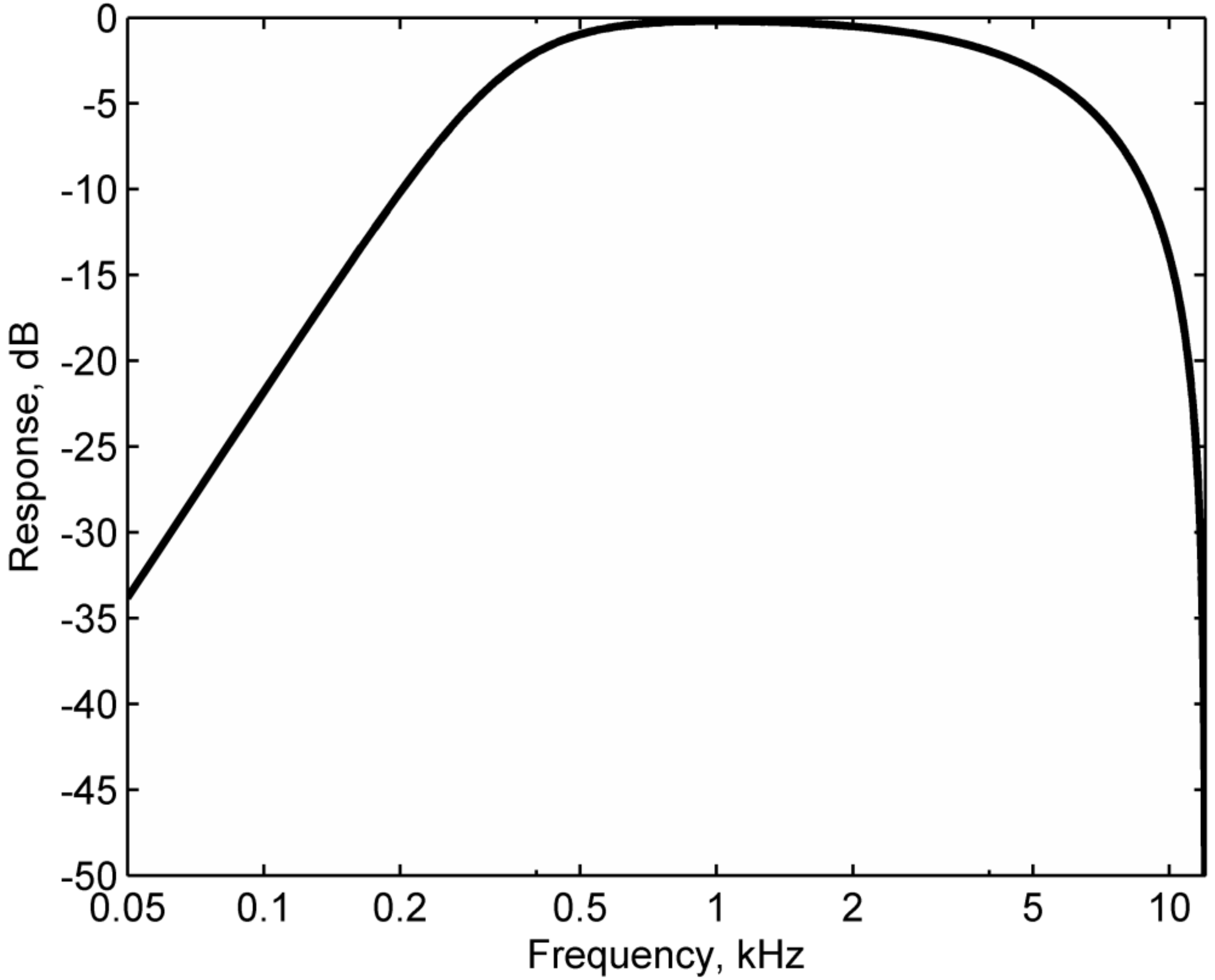
HASPI and HASQI Auditory Model

Model of the Auditory Periphery

- Resample signal at 24 kHz
- Middle ear bandpass filter 350 to 5000 Hz
- Auditory filterbank
  - 32 gammatone filters from 80 to 8000 Hz
  - Bandwidth depends on hearing loss and signal level
- Outer hair cell (OHC) dynamic-range compression
  - Compression ratio decreases with increasing OHC damage
  - Shift in auditory threshold
- Inner hair cell (IHC) neural firing rate adaptation
  - Rapid (2 ms) and short-term (60 ms) adaptation
  - IHC damage gives additional threshold shift

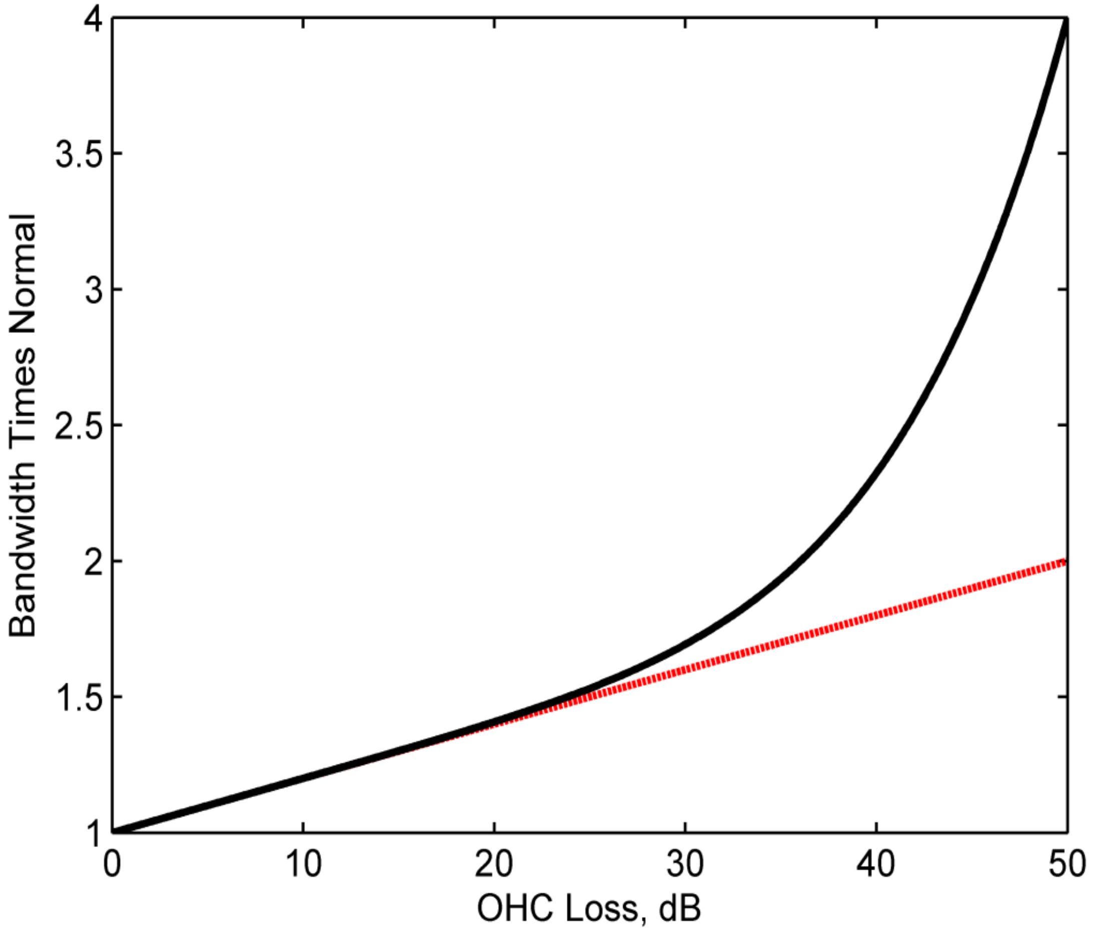
(Kates, 2013)

Middle Ear Filter

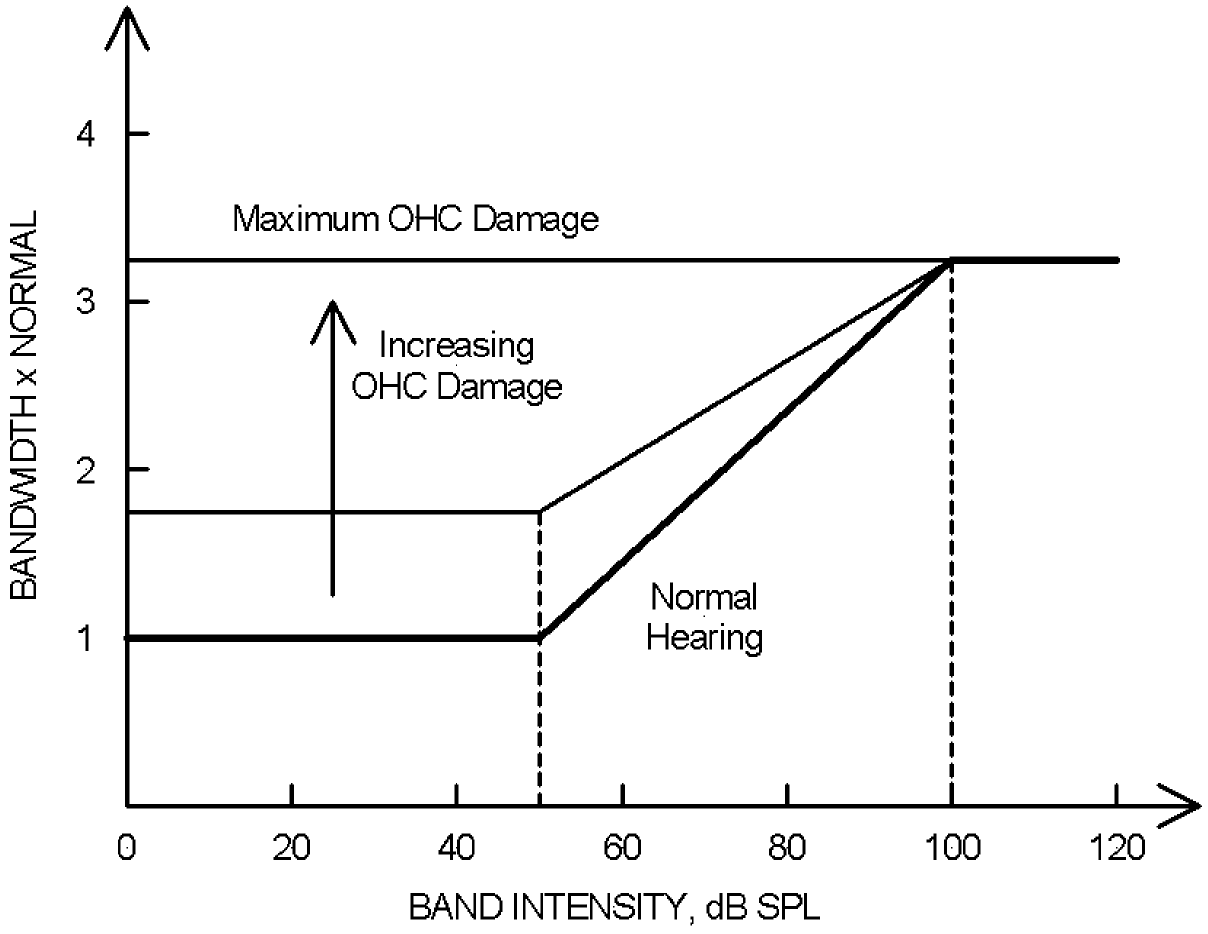


University of Colorado Boulder

Auditory Filter Bandwidth

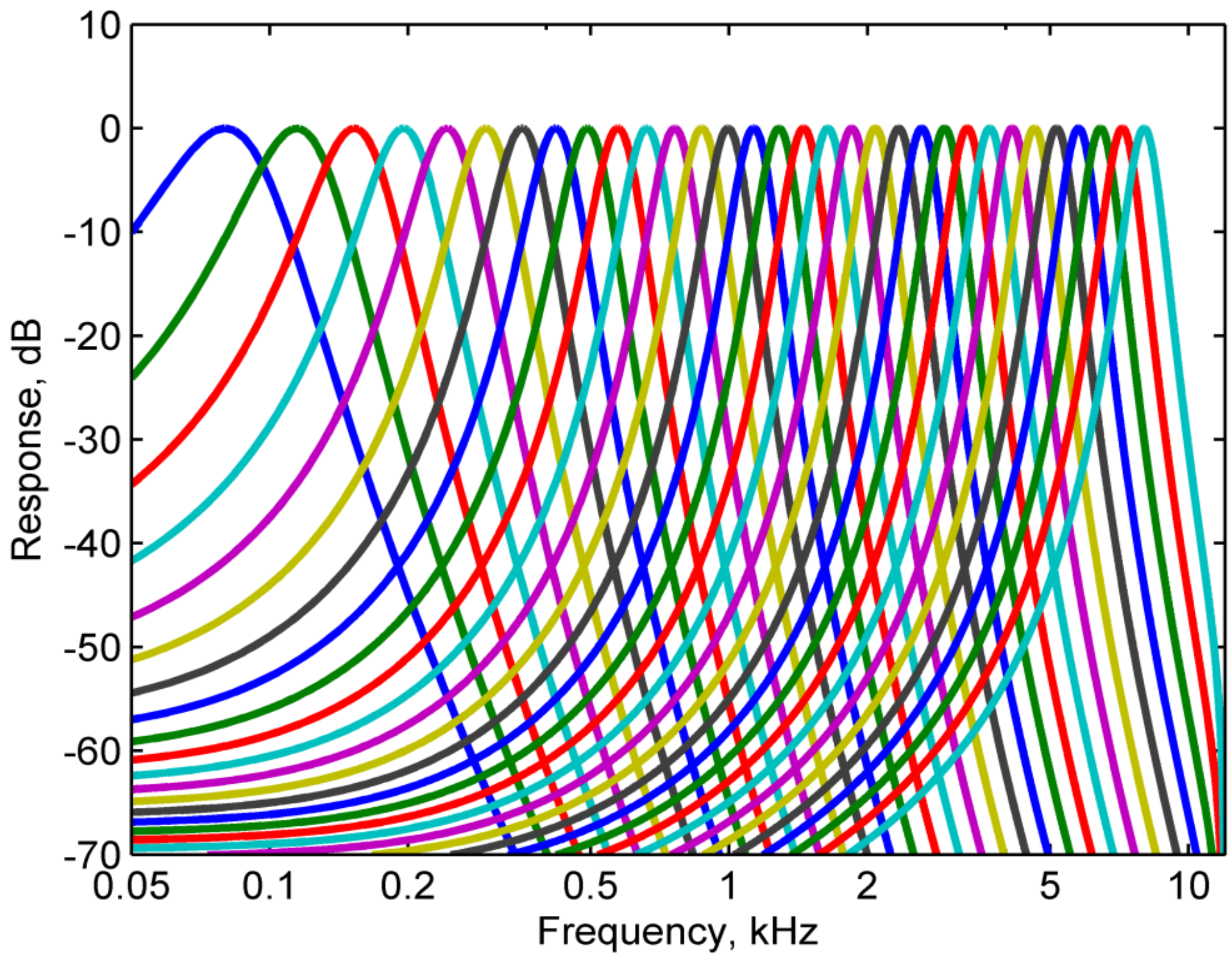


Signal Intensity

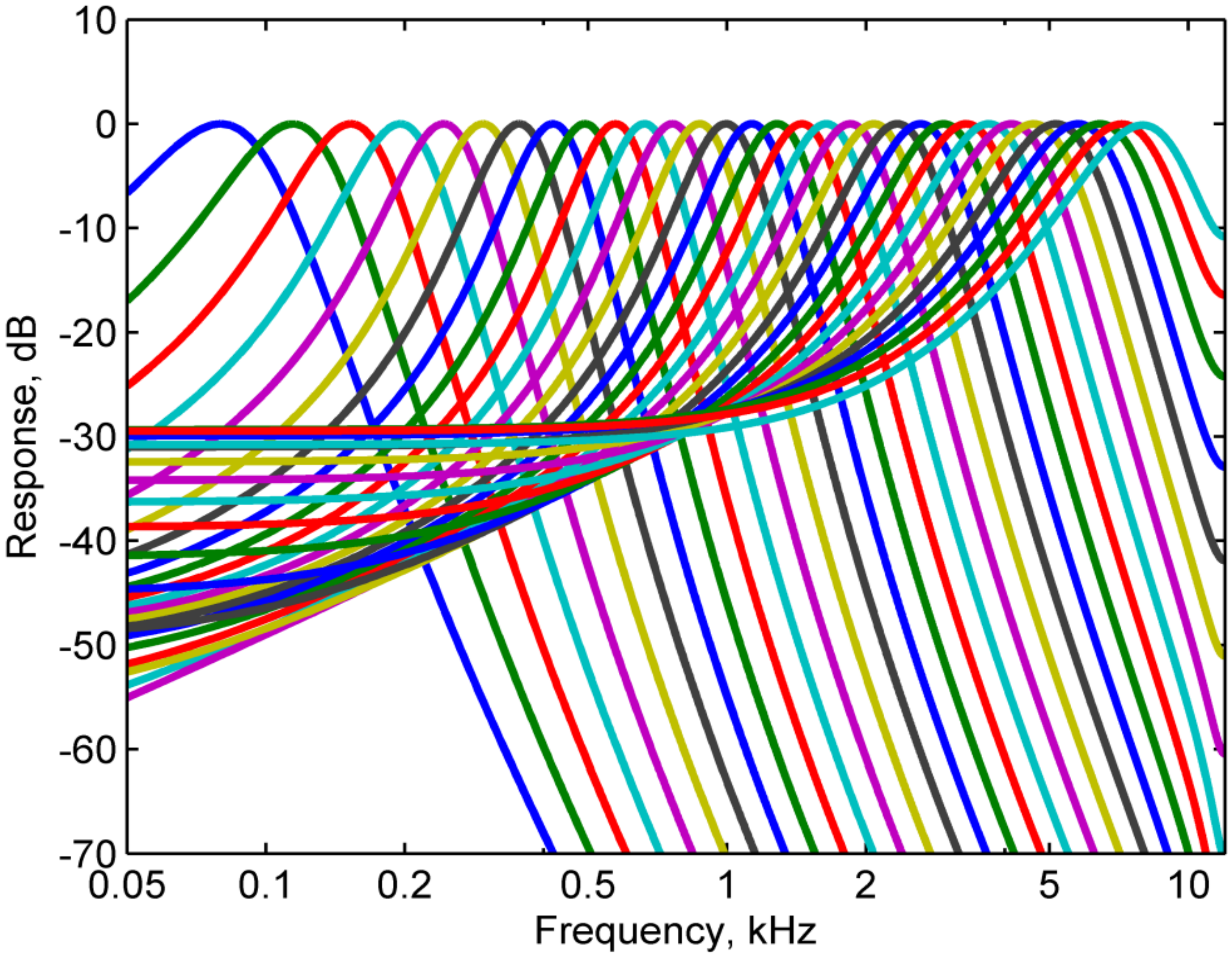


University of Colorado Boulder

Gammatone Filters: Normal Hearing

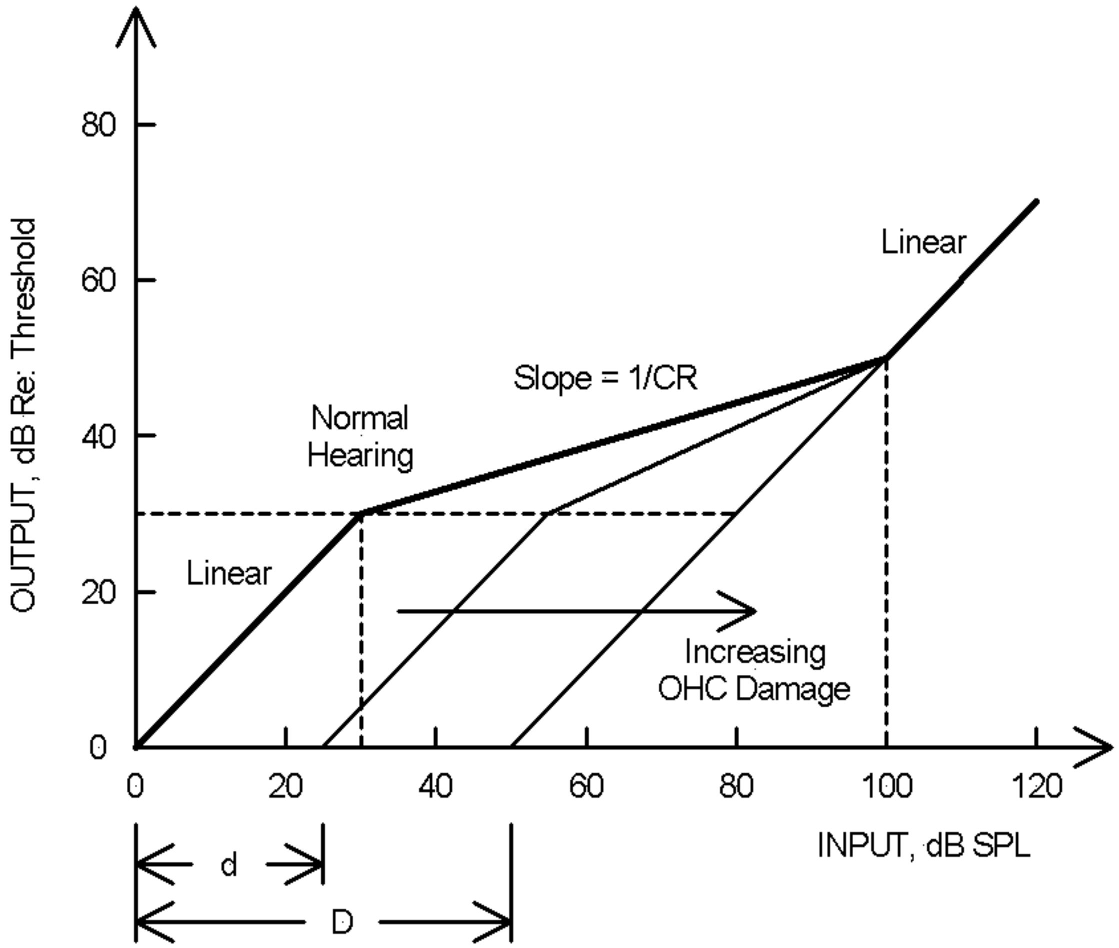


Gammatone Filters: Max OHC Damage

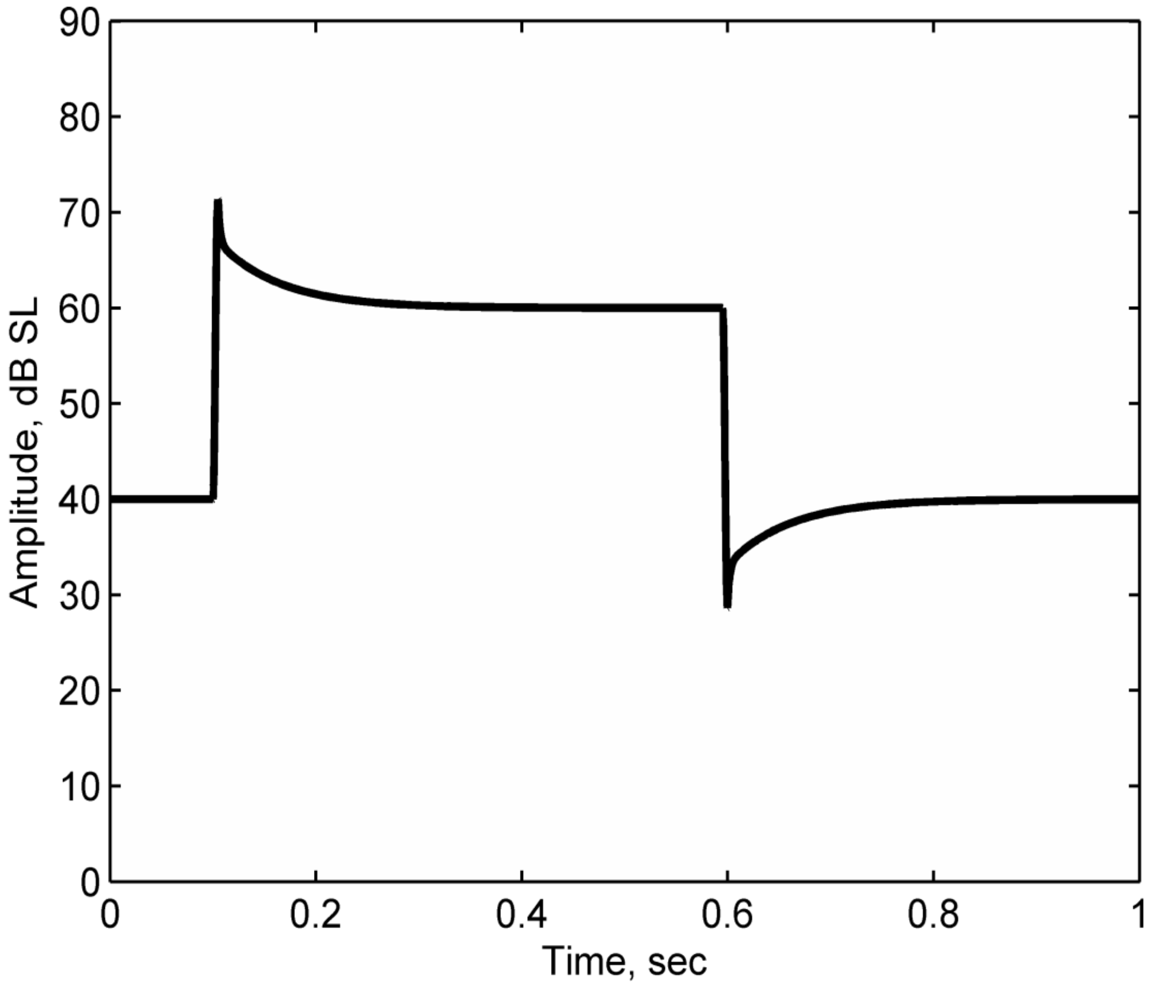


University of Colorado Boulder

OHC Dynamic-Range Compression



IHC Adaptation



Auditory Model Summary

- One peripheral model for all applications
  - Intelligibility and quality
  - Normal and impaired hearing
- Representation of hearing loss based on audiogram
  - Elevated auditory threshold
  - Increased auditory filter bandwidth
  - Reduced OHC dynamic-range compression
  - Reduced amount of two-tone suppression
- Model outputs
  - Envelope: Modulated envelope in each band in dB re: threshold
  - BM Vibration: Modulated envelope in dB applied to bandpass signal, includes temporal fine structure in each band

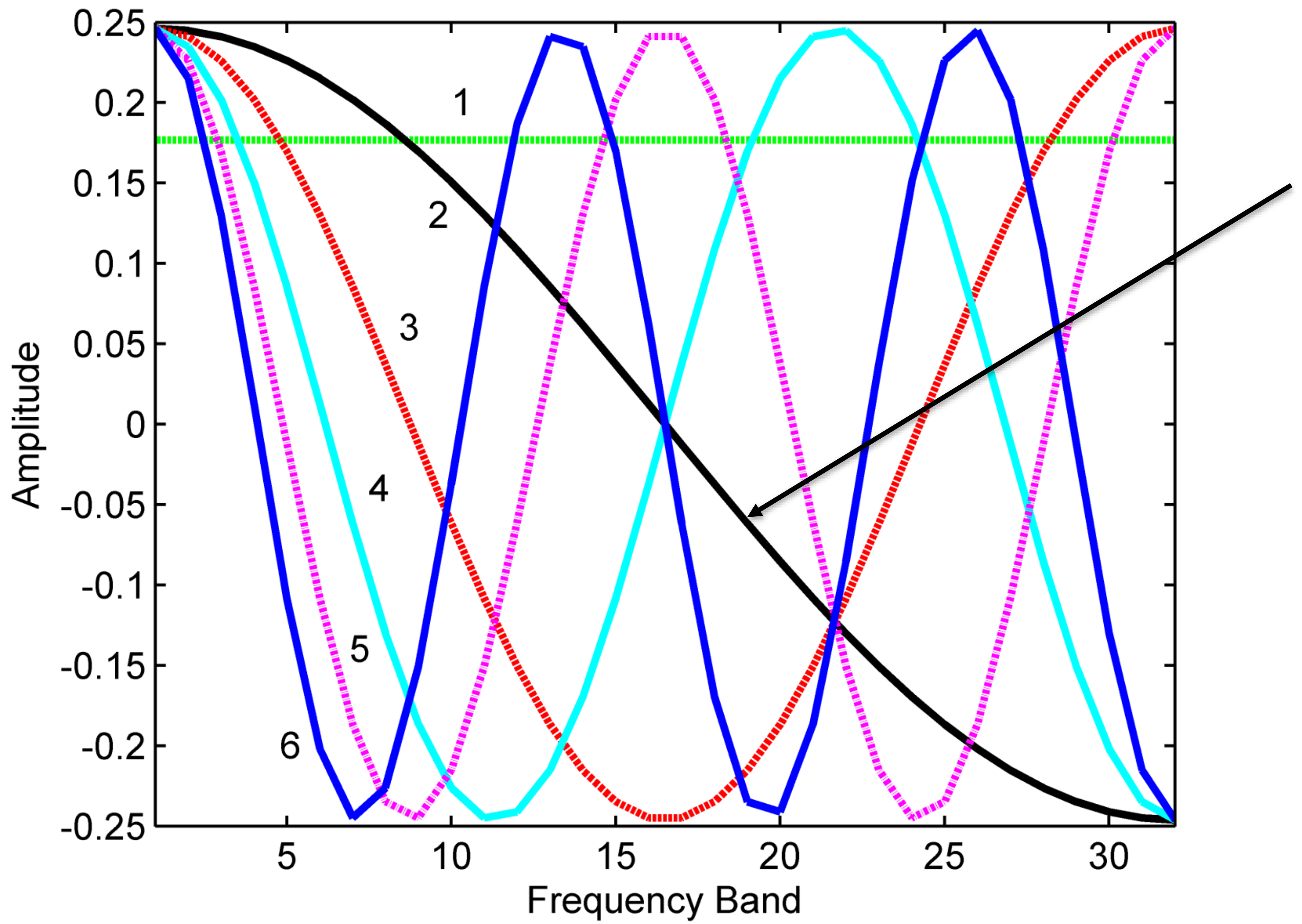
HASPI and HASQI Calculation

HASPI version 2

- Reference is clean speech, normal-hearing periphery
- Envelope time-frequency modulation analysis
  - Peripheral model envelope outputs in dB above threshold
  - Lowpass filter at 320 Hz, resample at 2560 Hz (8 x cutoff)
  - Remove samples identified as silences in reference
  - Fit short-time log spectra with 5 half-cosine basis functions from ½ to 2½ cycles per spectrum => mel-freq cepstral coefficients
  - Each coefficient sequence passed through modulation filterbank, 10 bands with center frequencies from 2 to 256 Hz
  - 5 basis functions vs time x 10 modulation filters = 50 sequences
  - Cross-correlate processed signal with reference for all 50
  - Average over 5 basis functions at each modulation rate to get 10 correlation values

(Kates and Arehart, 2014b; Kates and Arehart, 2015; Kates and Arehart, 2021)

Log Spectrum Basis Functions



Basis function 2 applied to the short-time spectra gives spectral tilt as a function of time.

HASPI v2 Neural Network

- Fit to HINT or IEEE sentences correct (all 5 keywords)
- Conditions: Noise and 6-talker babble, noise suppression, WDRC, frequency lowering, reverberation and reverb processing
- Neural network structure
  - Inputs are the 10 averaged modulation rate values
  - Single hidden layer, 4 neurons
  - Output layer with 1 neuron
  - Sigmoid activation function
- Ensemble of 10 networks: Bootstrap aggregation (“bagging”)
  - Networks fit to 63% of data randomly selected with replacement
  - Average outputs of the 10 networks
  - Reduced error variance and improved immunity to overfitting

HASQI version 2

- Reference is clean speech NAL-R, impaired-hearing periphery
- Fit to HINT pair (1 M + 1 F) in noise and babble, linear, nonlinear proc
- Nonlinear term
  - Envelope modulation
    - Low-pass filter dB envelope in each band
    - Measures time-frequency envelope fidelity
    - Cepstral correlation
  - Temporal fine structure
    - Short-time normalized cross-correlation
    - Loss of neural firing rate synchronization at high frequencies
    - Vibration correlation
- Linear term: differences in long-term spectrum and slope

(Kates and Arehart, 2014a)

HASQI Cepstral Correlation

- Input is envelope in dB
- Segment 16-ms windows, 50% overlap => lowpass filter
- Remove segments identified as silences in reference
- Fit each remaining segment with half-cosine basis functions
- Sequences for amplitude of each basis function over time
- Cross-correlation of processed with clean reference sequences
- c = average over basis functions 2 - 6

HASQI Vibration Correlation

- Vibration correlation v
  - Input is BM vibration
  - Segment 16-ms windows, 50% overlap
  - Remove silent segments found in reference
  - Short-time correlations of processed with reference segments
  - Loss of IHC synchronization 5-pole LP at 3.5 kHz
  - Weighting reduces importance of TFS at high frequencies
  - Normalize, weight with loss of synchronization, and average over segments and frequency bands
- Nonlinear term $Q_{Nonlin} = c^2 v$

University of Colorado Boulder

HASQI Linear Term

- RMS average envelope outputs in each frequency band
- Sum over bands and adjust overall levels to remove loudness difference between reference and processed signals
- Standard deviation of the spectral differences σ_1
- Spectral slope from 1st differences of adjacent bands
- Standard deviation of the spectral slope differences σ_2
- Linear model $Q_{Linear} = 1 - 0.579\sigma_1 - 0.421\sigma_2$

HASQI v2 Model

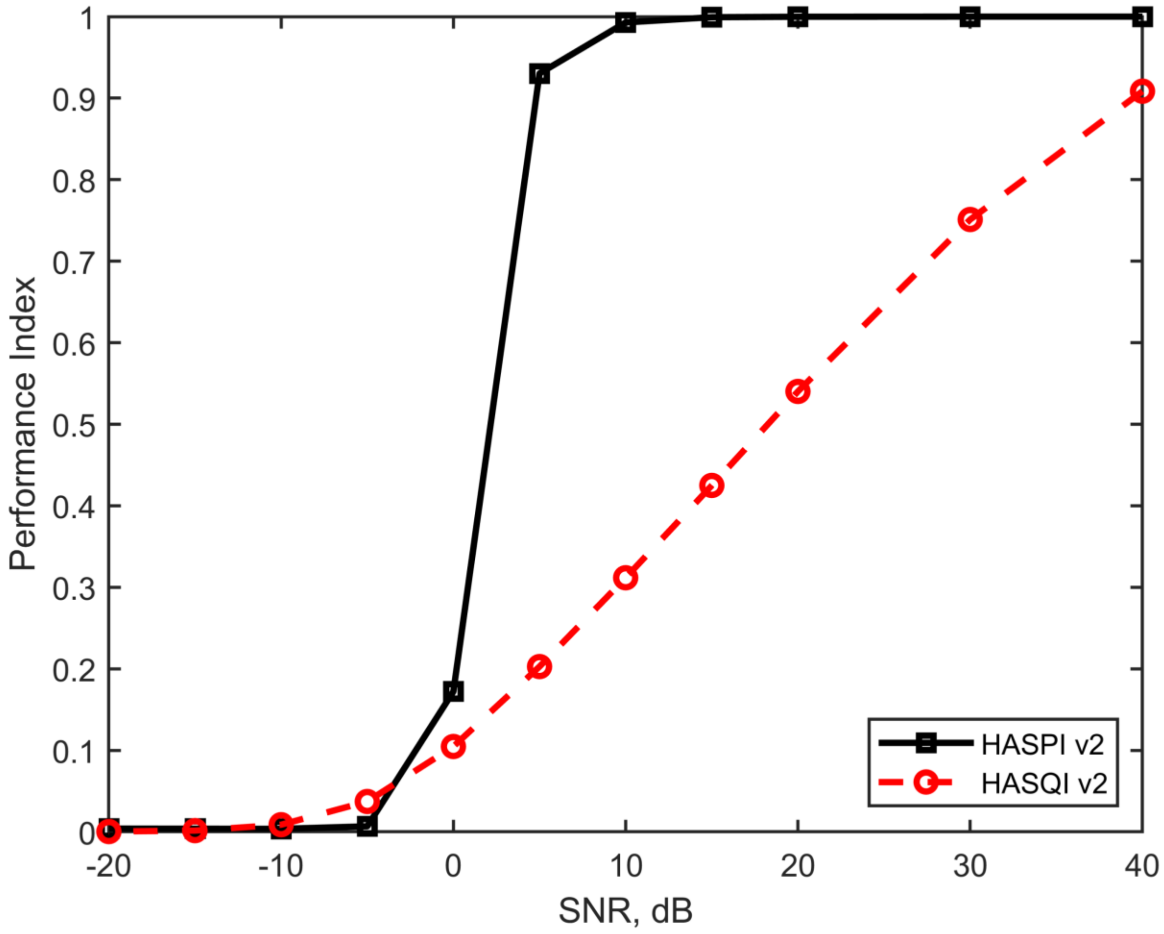
- Product of nonlinear and linear terms

$$Q_{Combined} = Q_{Nonlin} \times Q_{Linear}$$

- Nonlinear term: Envelope dominates, but TFS also important
- Linear term: Spectrum and spectral slope both important
- Product: Change in either nonlinear or linear will reduce quality

University of Colorado Boulder

HASPI and HASQI for LTASS Noise, NH



Applications

University of Colorado Boulder

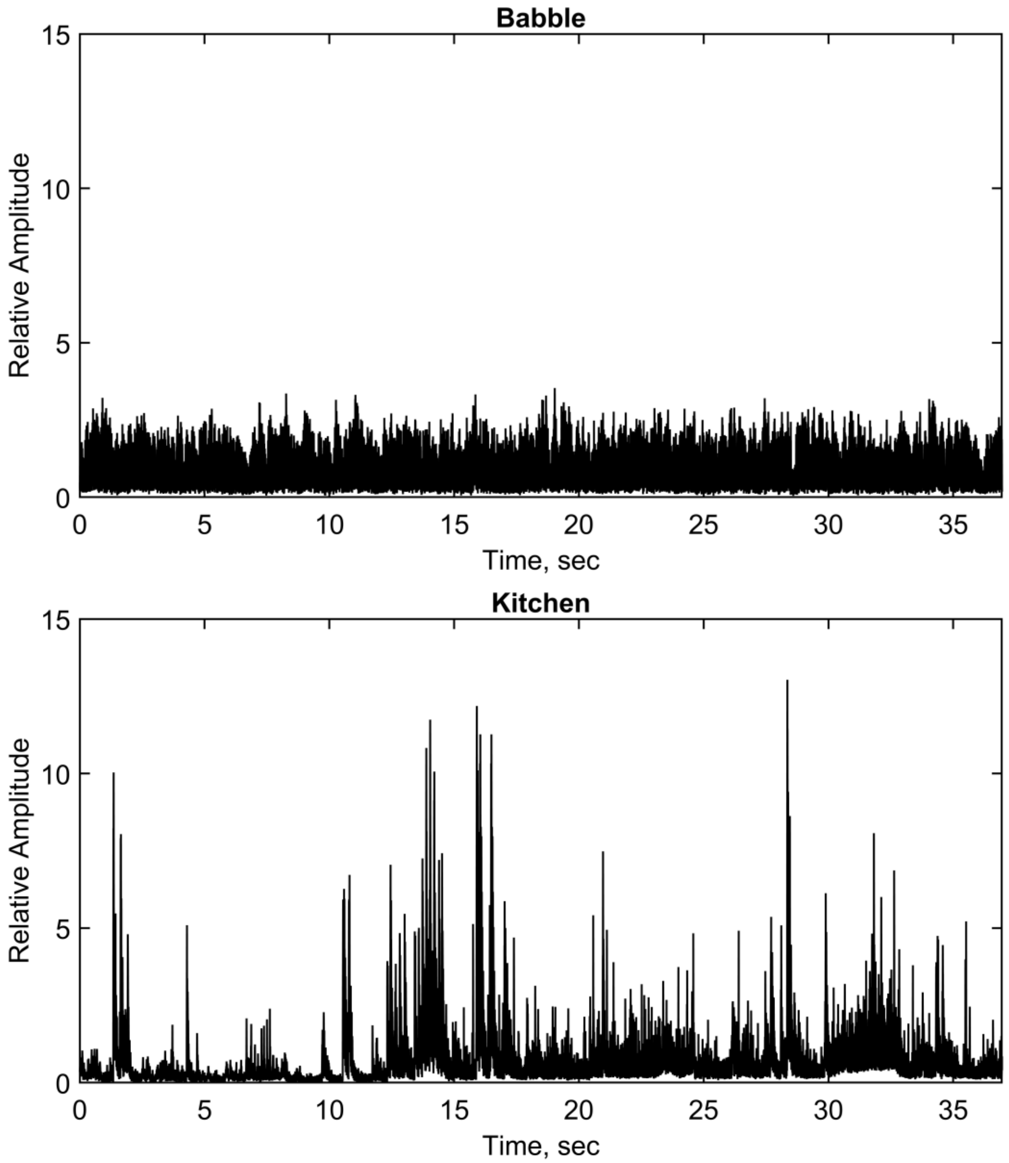
1. Quality Ratings for Noisy Speech

- Relationship between ratings and envelope modulation
- Ten older adult HI listeners, mild-moderate loss
- Five different noise types: 6-talker babble, 3-talker conversation, street traffic, kitchen, and fast-food restaurant
- Nine segments for each noise type
- Four SNRs: 3, 8, 12, and 20 dB
- Bilateral hearing-aid simulation: 2 settings x 2 vents per subject
- Quality ratings for HINT M + F sentence pair
  - 4 repetitions per processing condition with random noise segment
  - Rate 320 stimuli on scale from 0 – 10, converted to 0 – 100 for analysis

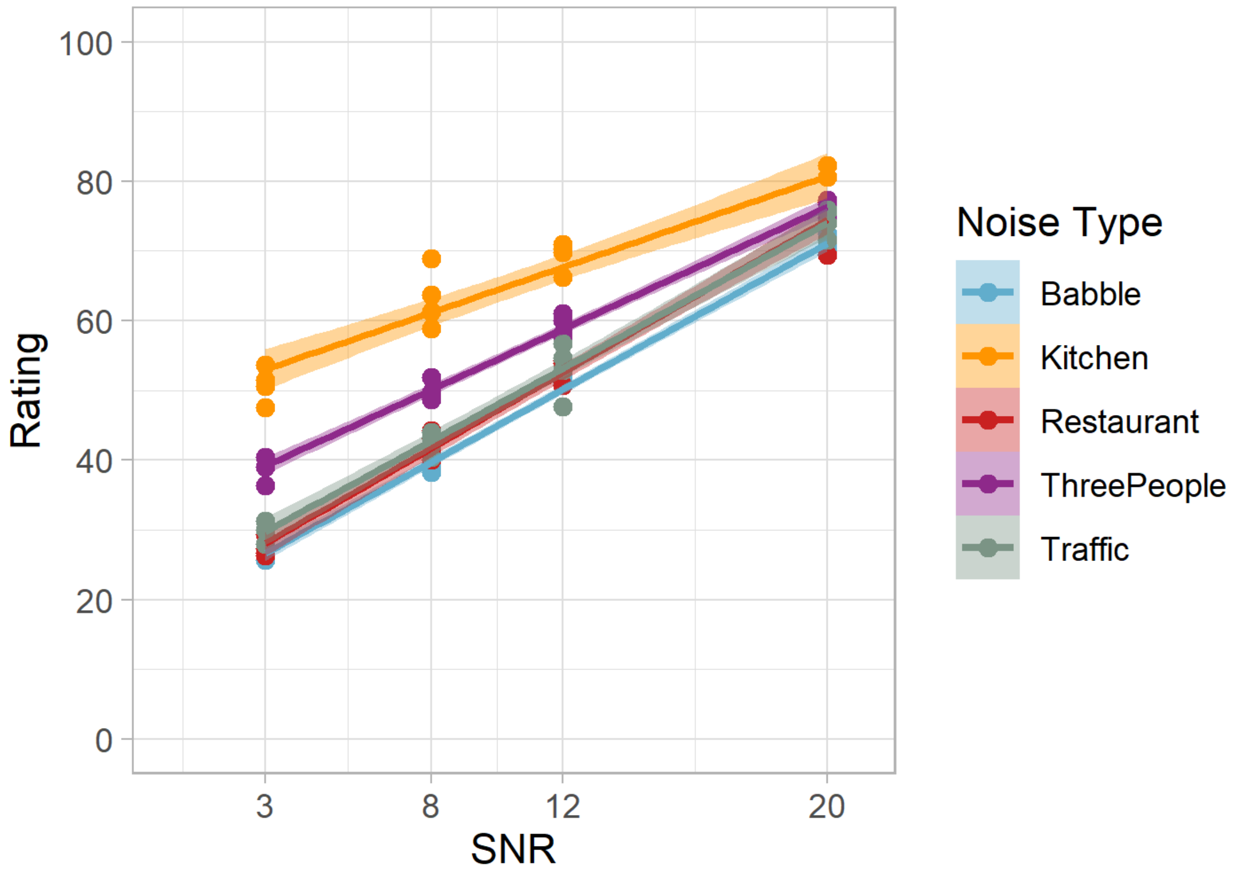
(Lundberg et al, 2020)

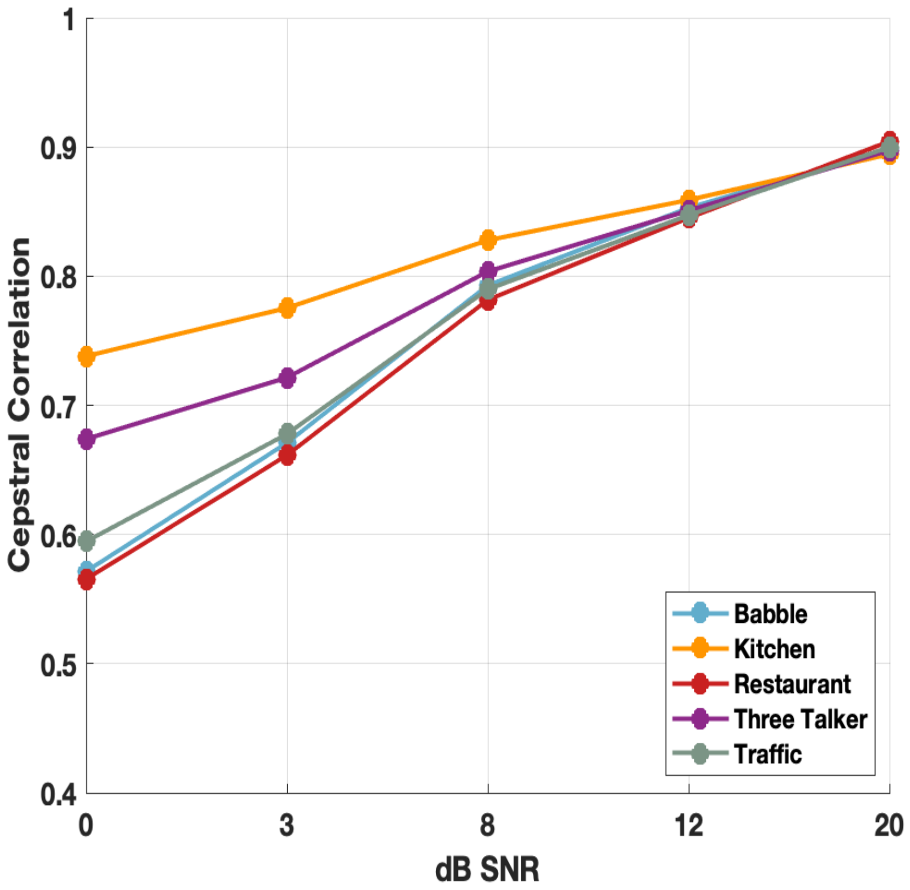
University of Colorado Boulder

Noise Waveforms



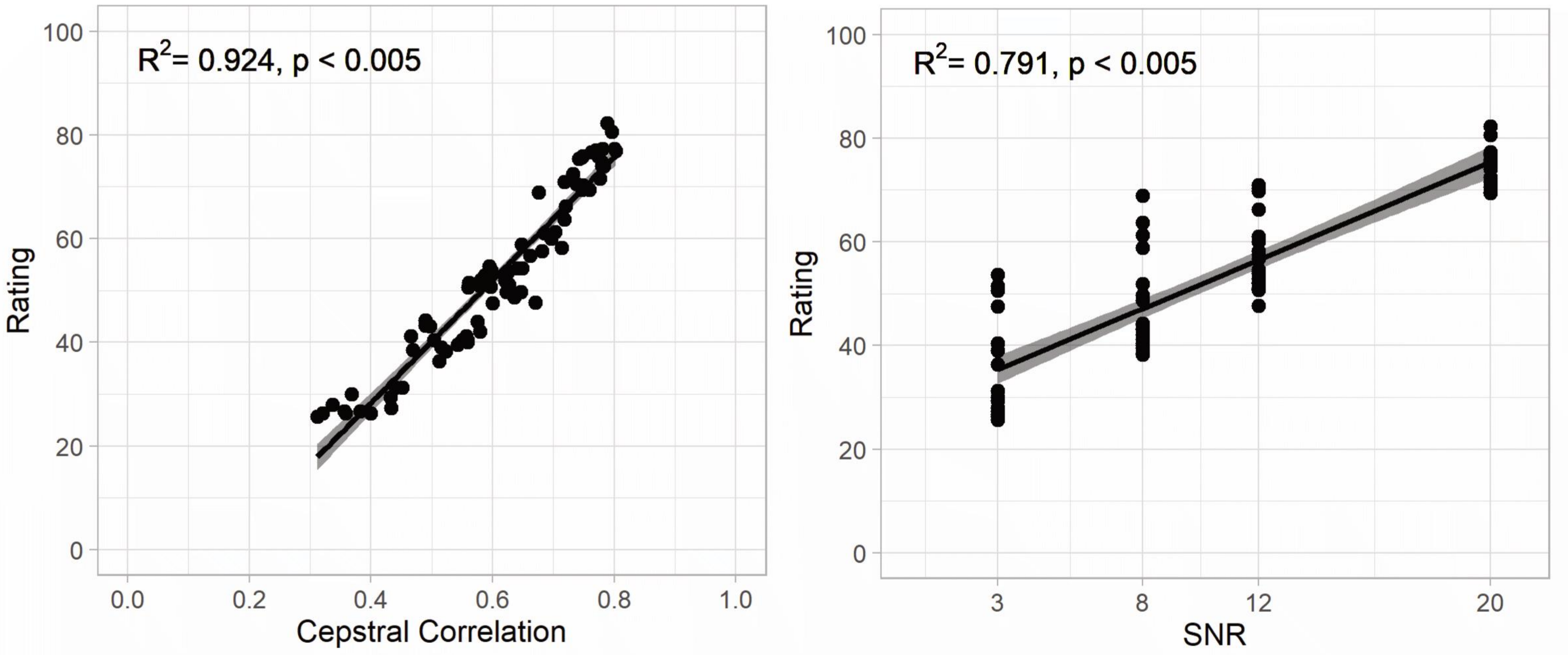
Averaged Quality Ratings





University of Colorado Boulder

Accuracy of Envelope Modulation Model



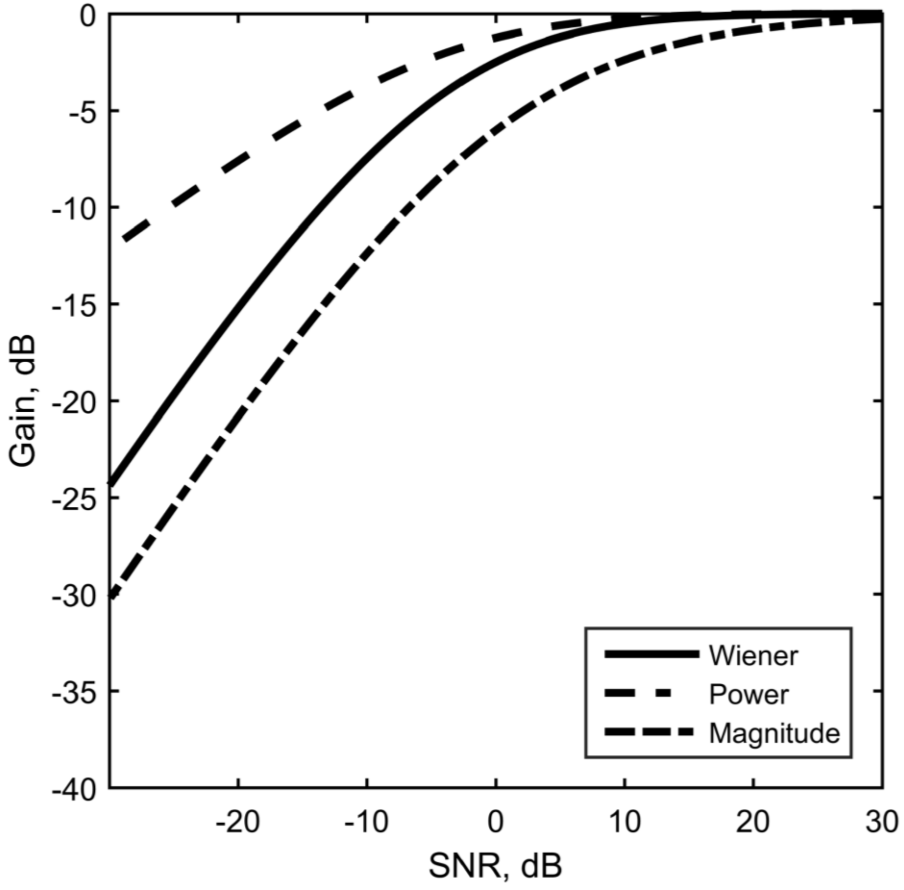
2. Single-Microphone Noise Suppression

- Use metrics to compare single-microphone algorithms
  - Spectral subtraction, 18 frequency bands
  - Spectral subtraction with upward spread of masking
  - Ideal binary mask (IBM)
  - Restore envelope of noisy speech to match that of clean speech
- Compare noise estimation procedures
  - Ideal knowledge of separate speech and noise RMS level fluctuations
  - Gives exact SNR in each time-frequency cell, 16-ms raised cos window
  - Or replace exact noise values with average over time in each band
- NH, N3 audiogram (moderate flat loss) with gain compensation
- Average over 20 IEEE sentences in 6-talker babble

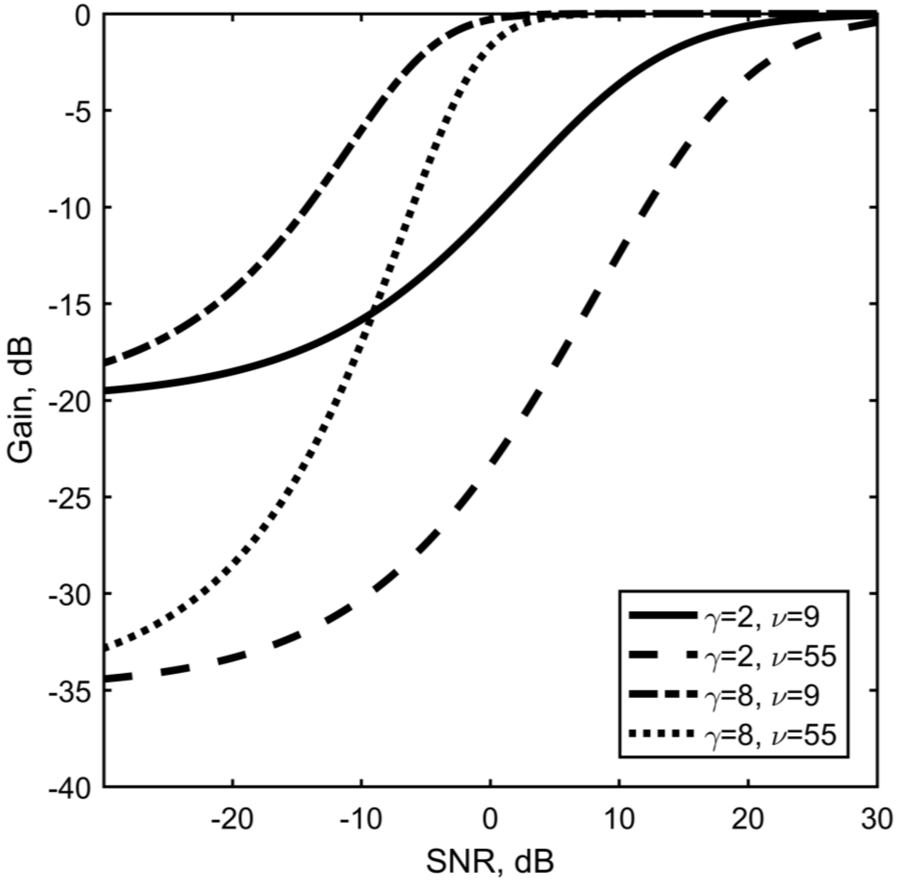
(Kates, 2017)

Gain vs SNR Rules

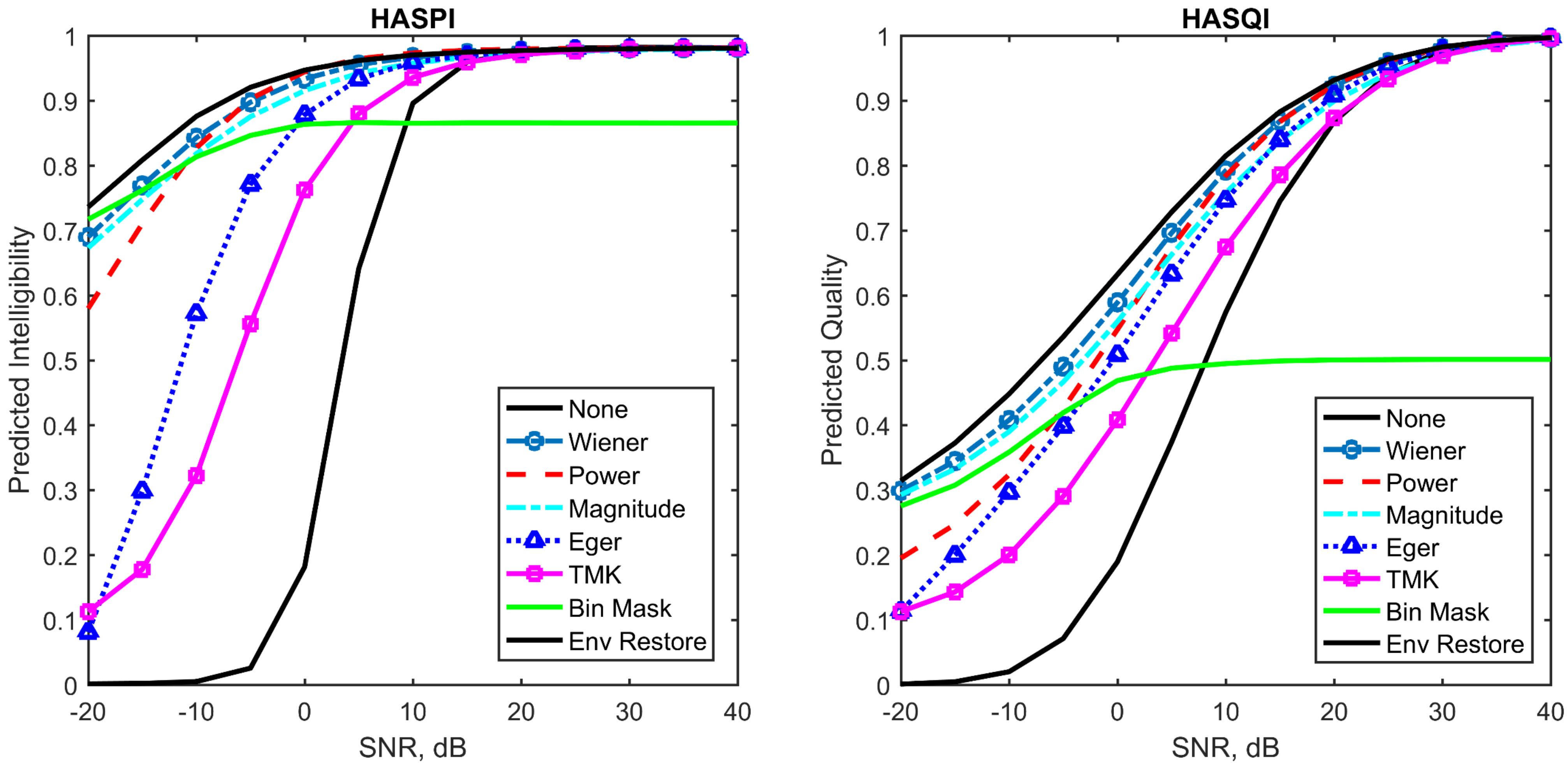
Spectral Subtraction



Eger

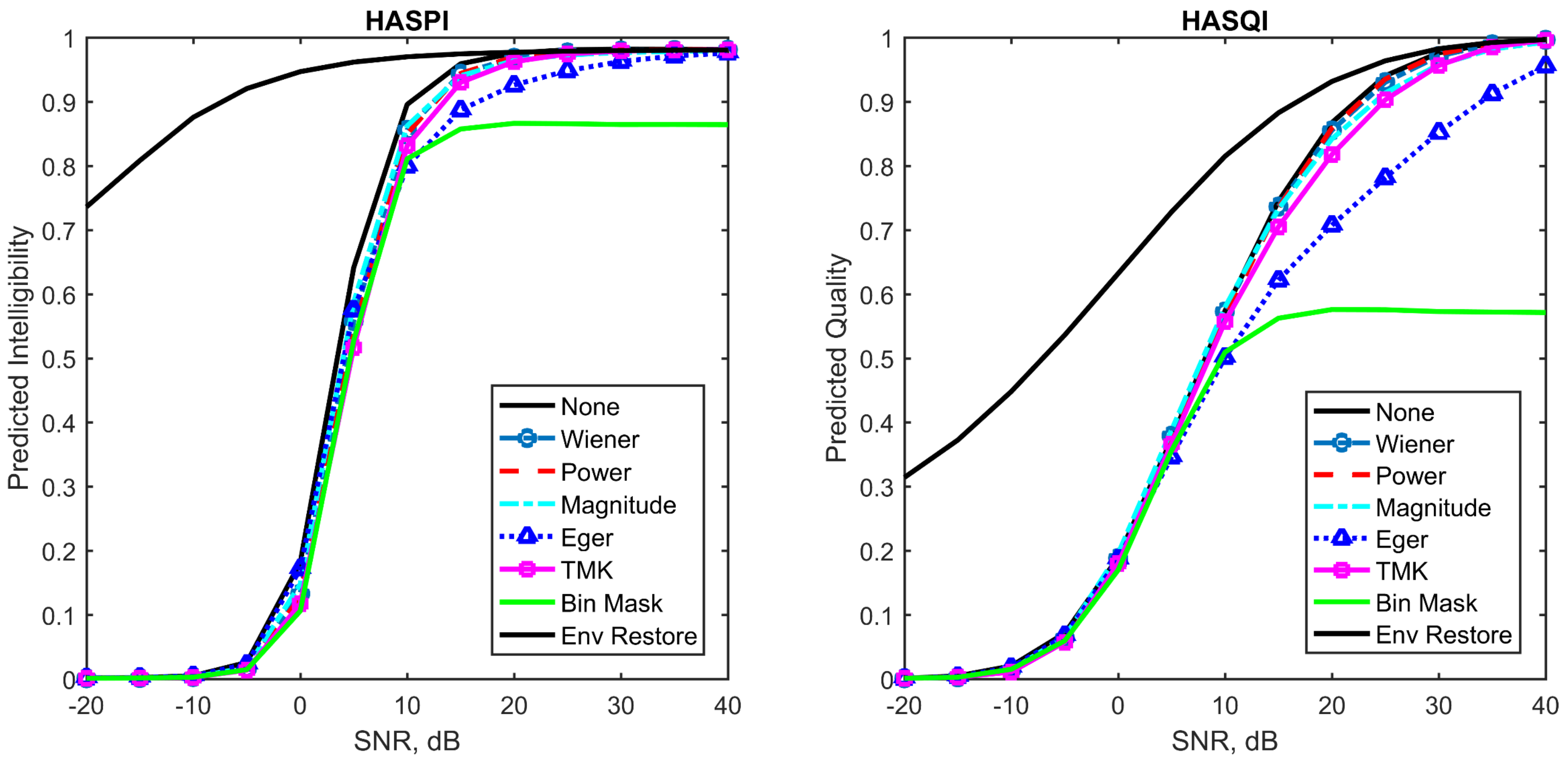


Ideal Processing, N3



University of Colorado Boulder

Average Noise Power, N3

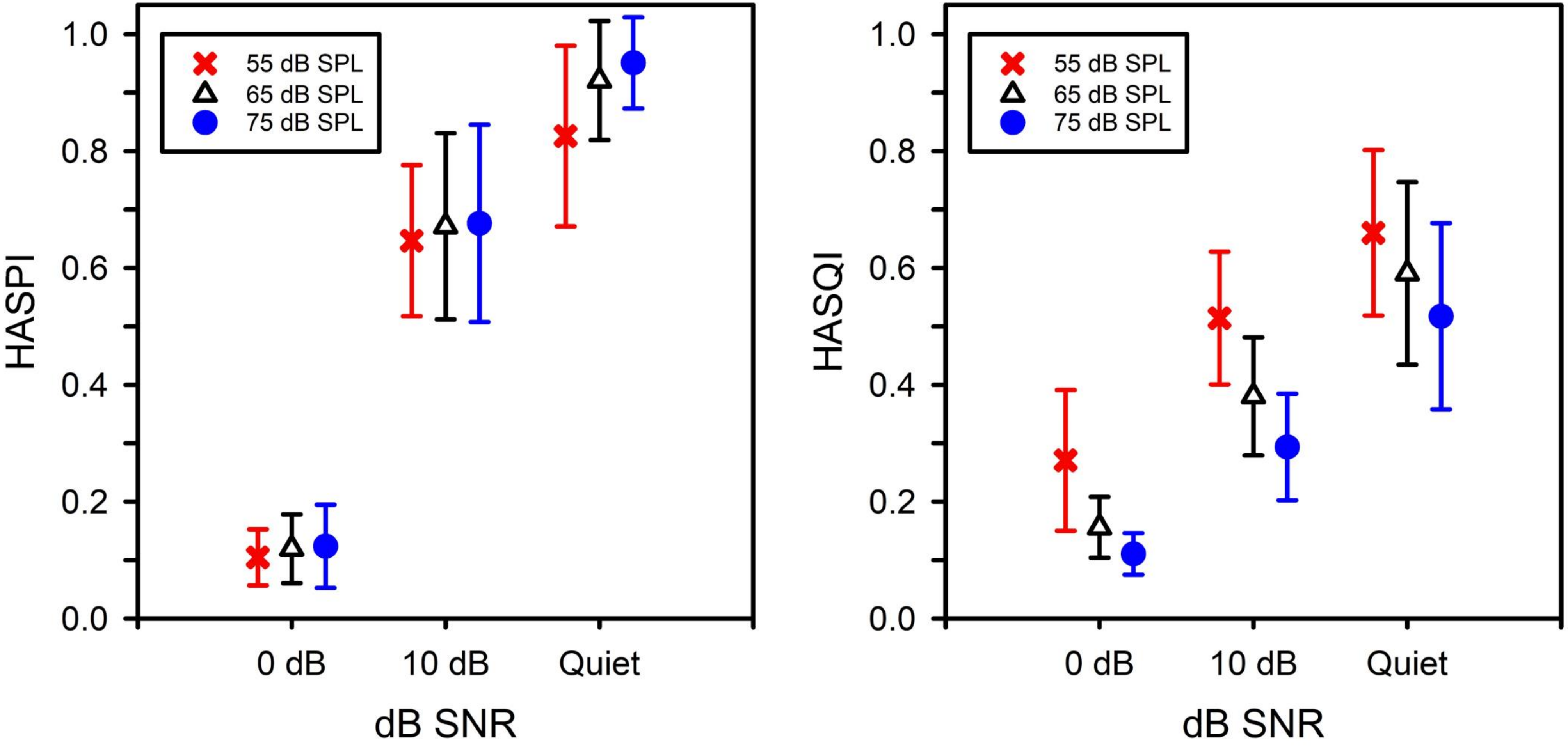


3. Commercial Hearing Aid Measurements

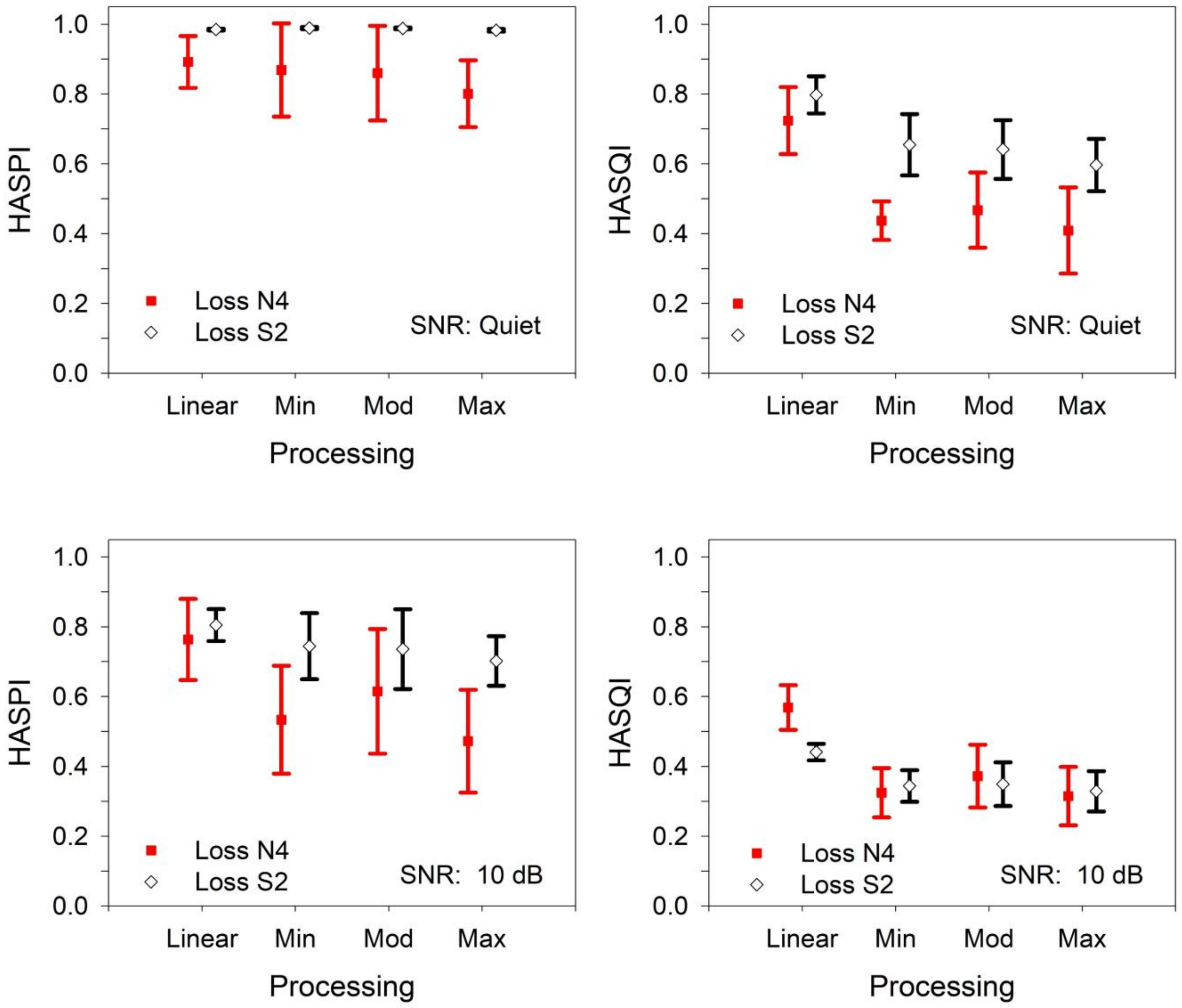
- Use metrics to distinguish between commercial devices
- Speech is HINT sentence pair
- Hearing aids
  – Three major manufacturers
  – Basic and premium model from each
  – WDRC, Noise suppression (NS), Frequency lowering (FL)
- Processing: NAL-R, Mild (no NS or FL), Moderate, Maximum
- Two audiograms: S2 (mild sloping), N4 (mod-severe flat)
- Vary SNR (6-talker babble), level of presentation
- Measurements use acoustic test box in sound booth

(Kates et al, 2018)

SNR, Multi-talker Babble



Processing Setting



Conclusions

- HASPI and HASQI accurate in predicting listener responses
  - Same peripheral model for both normal and impaired hearing
  - Measure nonlinear distortion, noise, linear response modifications
  - Envelope fidelity is important
  - Measures complete system, including processing interactions
  - Trade-off between audibility and nonlinear distortion
- Limitations
  - Derived for monaural headphone listening
  - Based on sentence test materials
  - Not validated for tonal languages
- MATLAB code free for the asking: *James.Kates@colorado.edu*

References

- J.M. Kates (2013), “An auditory model for intelligibility and quality predictions,” Proc. Mtgs. Acoust. (POMA) 19, 050184: Acoust. Soc. Am. 165th Meeting, Montreal, June 2-7, 2013.
- J.M. Kates and K.H. Arehart (2014a), “The hearing aid speech quality index (HASQI) version 2,” J. Audio Eng. Soc. 62, 99-117.
- J.M. Kates and K.H. Arehart (2014b), “The hearing aid speech perception index (HASPI),” Speech Comm. 65, 75-93.
- J.M. Kates and K.H. Arehart (2015), “Comparing the information conveyed by envelope modulation for speech intelligibility, speech quality, and music quality,” J. Acoust. Soc. Am. 138, 2470-2482.
- J.M. Kates (2017), “Modeling the effects of single-microphone noise-suppression,” Speech Comm. 90, 15-25.
- J.M. Kates, K.H. Arehart, M.C. Anderson, R. Kumar Muralimanohar, and L.O. Harvey, Jr. (2018), “Using objective metrics to measure hearing-aid performance,” Ear & Hearing 39, 1165–1175
- V. Rallapalli, M. Anderson, J. Kates, L. Balmert, L. Sirow, K. Arehart, and P. Souza (2020), “Quantifying the range of signal modification in clinically-fit hearing aids,” Ear & Hearing 41, 433-441.
- E.M.H. Lundberg, S.-H. Chon, J.M. Kates, M.C. Anderson, and K.H. Arehart (2020), “The type of noise influences quality ratings for noisy speech in hearing aid users,” J. Speech Lang. Hear. Res. 63, 4300-4313.
- J.M. Kates and K.H. Arehart (2021), “The hearing-aid speech perception index (HASPI) version 2,” Speech Comm. 131, 35-46.