# Measuring and predicting quality ratings of fast-acting single-microphone noise reduction

William S. Woods, Cheryl Eiler, Brent Edwards

SHRC, Berkeley, CA (william_woods@starkey.com)

## Motivation

Four main questions motivate this work:

- What is the best fast-acting single microphone noise reduction (SMNR) design from a quality standpoint?
- Does the best design depend on hearing loss?
- Does the best design depend on signal conditions?
- Is the best design predictable from its effects on speech in noise?

These questions are addressed by assessing the quality of a large number of SMNR designs under a variety of signal conditions and reducing the number through multiple stages of paired comparisons.

## General Methods

Normal-hearing subjects and subjects with high-frequency sloping hearing losses evaluated SMNR algorithms in **6 noise conditions: either car interior or cafeteria noise, at 0, 5, and 15 dB SNRs**.

Performed four stages of listening to find the best SMNR setting:

**Stage 1**: Independently in each noise condition one listener with normal hearing listened to **216 different SMNR settings of a simplified algorithm (SA)**, eliminating obviously poor versions. This yielded 18 settings for each noise condition.

**Stage 2**: Five listeners with normal hearing did A/B comparisons of 18 settings against unprocessed only. Top 5 of 18 move to next stage.

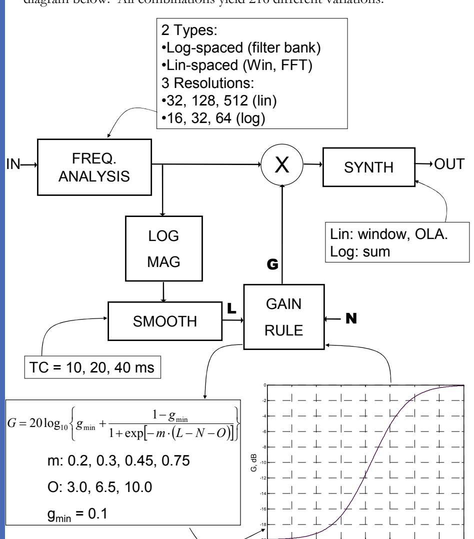
**Stage 3**: Five listeners from each of normal, mild, moderate, and moderate-severe loss groups, in a full pairwise comparison, rated top 5 SA versions against unprocessed and Ephraim, Malah (1985; EM) a la Martin et al. (2004). Performed separately for each noise condition.

**Stage 4**: Five top-rated algorithms across noise conditions from Stage 3 were rated against each other in all noise conditions.

Sentences used were "A boy fell from the window" from the HINT corpus (male), and "A pot of tea helps to pass the evening" (female).

For all algorithms used here a noise spectrum estimate **N** is required. This spectrum was pre-computed with no speech present.

The **SA algorithm and its stage 1 parameter ranges** are shown in the diagram below. All combinations yield 216 different variations.




## Stage 1

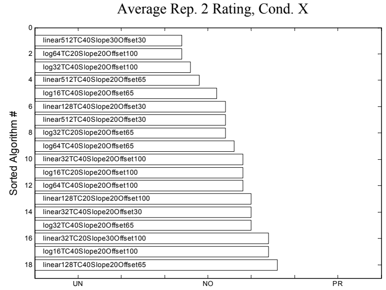
All 216 parameter variations of the SA algorithm were evaluated in all 6 noise conditions. Final 18 settings in each condition determined by eliminating lowest-quality variations. This was performed by listening to signals one at a time, deciding whether to eliminate them or keep them, and cycling through the remaining signals until only 18 were left.

## Stage 2

- Final 18 settings from Stage 1, in A/B comparison to unprocessed alone, were given a rating of one of the following: "A MUCH BETTER", "A BETTER", "A SLIGHTLY BETTER", "NO DIFFERENCE", "B SLIGHTLY BETTER", "B BETTER", "B MUCH BETTER".
- Performed twice by each of 5 listeners with normal hearing.
- A and B were randomly chosen as processed or unprocessed on each trial.
- Subjects could repeat the trial as often as desired before providing preference rating.
- 75 dB SPL overall level.

## Sample Result

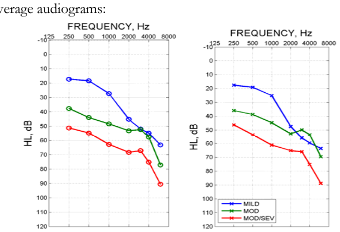
Average preference ratings across subjects from second repetition were used to rank the 18 settings. Slope and offset values are multiplied by 100 and 10, respectively, for plotting clarity. TCs are in milliseconds.




## Stage 3

Five listeners from each of normal, mild, moderate, and moderate-severe loss groups, in a full pairwise comparison, rated top 5 SA versions from stage 2 against unprocessed and Ephraim, Malah (1985). Performed separately for each noise condition.

- Average age of loss listeners was 74, 78, and 72 yrs.
- NAL R linear gain applied for losses.
- Subjects with loss perform each condition twice.
- Average audiograms:




## Stage 3 Results

Two mild-loss listeners focused on speech distortions. Their data not analyzed. For remaining listeners, count # of times an algorithm is preferred (a "win") in A/B comparison, collapsed across listeners in group. Rank the algorithms by number of wins.

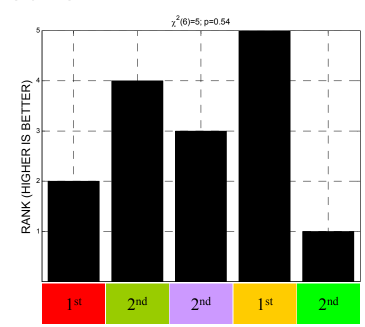
| | | | SUBJECT GROUP | | | |
|---|---|---|---|---|---|---|
| NOISE | SNR | STAGE 3 RANK | NORM | MILD | MOD | MOD/SEV |
| Babble | 0 | 1st | 2nd | EM | 2nd | EM |
| | | 2nd | 2nd" | 2nd" | EM | 2nd" |
| | 5 | 1st | 1st | 1st | 1st | 1st |
| | | 2nd | 2nd | 3rd | EM | EM |
| | 15 | 1st | 2nd | 2nd | 2nd | 2nd |
| | | 2nd | 1st | 1st | 3rd | 3rd |
| Car | 0 | 1st | 2nd | 2nd | EM | 2nd |
| | | 2nd | 2nd' | EM | 2nd | 2nd' |
| | 5 | 1st | 2nd | 2nd | 2nd | 2nd |
| | | 2nd | 3rd | EM | EM | EM |
| | 15 | 1st | 1st | 2nd | 1st | 3rd |
| | | 2nd | 2nd | 1st | 2nd | 2nd |

Table entries are either rank of algorithm from Stage 2, or "EM" for the Ephraim/Malah algorithm. Single quote indicates a tie from Stage 2, double quote an algorithm that differs only in frequency resolution. Each color corresponds to a single algorithm setting.

## Stage 4

- Same subjects as in Stage 3, except no moderate-severe.
- Choose 5 SA variations from top 6 in Stage 3 (2 of top 6 differed only in slope parameter).
- Full pairwise comparison of 5 SA variations in each signal-in-noise condition, with added female talker.
- Fit Bradley-Terry-Luce model to preference data, collapsed across subjects and conditions, using OptiPt (Wickelmaier & Schmid, 2004).
  - Provides rank ordering of settings based on probabilistic model of preference judgment.
  - Provides $\chi^2$ test of null hypothesis: BTL model well-fits input data.

## Results

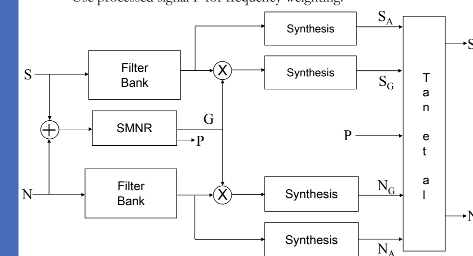



- $\chi^2$ test indicates null hypothesis cannot be discarded.
  - Null hypothesis is that the BTL model, and thus its rank order, is valid for all subjects and conditions.
  - Implies that the "best" SMNR setting is likely 1st and does not differ across subjects and conditions.

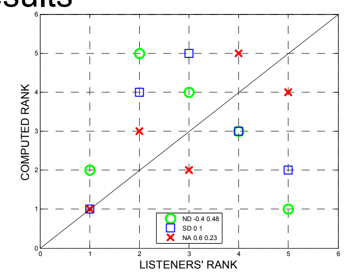
## Predicting Rank Order

Several metrics were used to model the stage-4 rankings:

- Noise attenuation (NA) - use leading and trailing noise-alone portions to compute amount of noise attenuation.
- Speech and noise distortion (SD and ND) - apply simplified version of Tan et al.'s (2004) method to gained speech alone $S_G$ and gained noise alone $N_G$.
  - Use $S_A$ and $N_A$ as time-aligned reference signals.
  - Use processed signal P for frequency weighting.




## Results




- Legend shows metric name, Kendall's τ, and the p-value for τ.
- Noise attenuation NA most highly correlated with subject ranking.
- Speech and noise distortions (SD, ND) zero or negatively correlated with subject ranking. (NA typically positively correlated with SD and ND.)
- Might be due to subjects' focus on NA due to high presentation level.
  - Tolerate some distortion for more noise reduction.
  - Did not analyze data from two subjects who focused on distortion.

## Conclusions

- Top settings of simplified SMNR algorithm are preferred over unprocessed (Stage 3).
  - Mod-severe may show weaker preference.
- Top settings appear to be similar across loss groups and conditions studied (Stage 4).
- Ranking under our conditions was most highly correlated with amount of noise attenuation.
  - Current measure of speech and noise distortions negatively correlated with algorithm rank.

## References

- Ephraim, Y. and Malah, D. (1985). "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,"*IEEE Trans. ASSP* 33 (2), pp. 443–445.
- Kendall, M. (1938). "A new measure of rank correlation", *Biometrika* 30 (1/2), 81-93.
- Martin, R., Malah, D., Cox, R., Accardi, A. (2004). "A Noise Reduction Preprocessor for Mobile Voice Communication", *EURASIP J. App. Sig. Proc.* 8, 1046–1058.
- Tan, C., Moore, B., Zacharov, N., Mattila, V. (2004). "Predicting the Perceived Quality of Nonlinearly Distorted Music and Speech Signals", *J. Audio Eng. Soc.* 52 (7/8), 699-711.
- Wickelmaier F., and Schmid C. (2004). "A Matlab function to estimate choice model parameters from paired-comparison data", *Beh. Res. Meth., Instr., & Comp.* 36 (1), 29-40.