SNR is not enough: Noise modulation and speech quality

James M. Kates and Kathryn H. Arehart
James.Kates@colorado.edu, Kathryn.Arehart@colorado.edu

Department of Speech, Language, and Hearing Sciences
University of Colorado, Boulder, Colorado, USA

ABSTRACT

Speech quality metrics are used to predict quality under a variety of noise and distortion conditions. A simple metric for noisy speech is the signal-to-noise ratio (SNR), which gives the ratio of the long-term average speech power to the long-term average noise power. However, the average noise level fails to include the impact of the noise amplitude modulation on speech quality judgments. This paper reviews how the noise characteristics affect quality ratings even when the SNR is held constant. In one study, ratings for speech combined with stationary Gaussian noise are compared to ratings for noise having the same envelope modulation as the speech. In a second study, ratings for speech combined with continuous noise or multi-talker babble are compared to speech where the interference is gated off during the silent intervals in the stimulus. An accurate quality metric must take into account the modulation as well as the intensity of the noise, and the implications for speech quality metrics are discussed.

Index Terms— Auditory Modeling and Hearing Aids (1.6), Hearing and Psychoacoustics (13.2.2), Distortion Measures (13.3.2)

1. INTRODUCTION

Background noise can reduce both speech intelligibility and speech quality. The noise intensity is often characterized in terms of the signal-to-noise ratio (SNR), which measures the long-term noise power relative to that of the speech signal. However, even at the same SNR, quality ratings vary depending on the noise characteristics [1] [2]. Leman et al. [1], for example, found that listeners gave significantly higher quality ratings to speech accompanied by non-stationary noises such as restaurant babble and city street noises than for stationary noises at the same SNR. Arehart et al. [2] found higher quality ratings for speech in a background of multi-talker babble than for speech in a background of stationary speech-shaped noise.

The impact of noise on speech is most accurately measured using subjective listening tests. But listening tests are expensive and time-consuming, which has led to the development of objective quality measures [3] [4] [5]. Measures based on the SNR, in particular the segmental SNR [3], are often used to indicate the effectiveness of speech-processing algorithms [6] [7]. The segmental SNR is computed by dividing the noisy speech into segments, computing the SNR for each segment, and averaging the SNR values over the utterance. Comparative evaluations of quality indices, however, have shown that the segmental SNR is generally the least accurate of the quality metrics studied. In contrast, approaches based on other signal characteristics, such as changes in the envelope modulation, are much more accurate in predicting quality ratings [8] [9] [10] [11].

One factor that contributes to the differences in noise perception is the noise envelope modulation. Jin et al [12] measured sentence quality ratings for noisy speech where the noise varied from having no superimposed envelope modulation (i.e. stationary noise) to having envelope modulation that duplicated that of speech. In all cases the long-term average noise level (and hence the SNR) was held constant. They found that the rated quality increased as the amount of noise modulation increased even though there was no change in the SNR. The Jin et al [12] data will be reviewed in this paper, and it will be shown that the Hearing Aid Speech Quality Index (HASQI) [5], which is based on measuring changes in the envelope modulation, is very accurate in predicting the quality ratings.

A second factor that may contribute to the differences in noise perception is whether the noise is audible during silent intervals in the speech. Data from a new experiment is presented in which noise was added to a pair of concatenated sentences, with the noise either present continuously before and after the sentences and throughout the gap between the sentences, or gated off before the stimuli, after, and in the gap. The impact of noise during speech silences is considered for quality ratings and for procedures, such as HASQI, that are used to compute quality indices.

2. MODULATED NOISE

Jin et al [12] tested ten subjects with normal hearing. The test material comprised a concatenated pair of sentences

spoken by a male talker extracted from the hearing-in-noise test (HINT) [13]. The stimuli included one pair of sentences without any noise and eleven noisy sentences. The sampling rate was 22.05 kHz. Subjects were tested monaurally using headphones, with the stimuli presented to the right ear.

The stimuli were processed through a 16-band FIR filter bank covering the frequencies from 80 Hz to 10 kHz. The speech envelope was extracted in each frequency band using the Hilbert transform and low-pass filtered using a cutoff frequency of 30 Hz. Stationary Gaussian noise was passed through the same filter bank as the speech. Fully-modulated noise was formed by multiplying the noise in each frequency by the speech envelope lowpass filtered at 30 Hz, while unmodulated noise bypassed the envelope multiplication. Intermediate amounts of noise modulation were produced by blending the stationary noise with the fully-modulated noise while holding the overall long-term noise level constant. The amount of noise modulation varied from 0 to 100 percent in steps of 10 percent, and all frequency bands for a given stimulus were assigned the same percent modulation. Overall SNR values of 0, 10, and 20 dB were used along with speech not having any added noise.

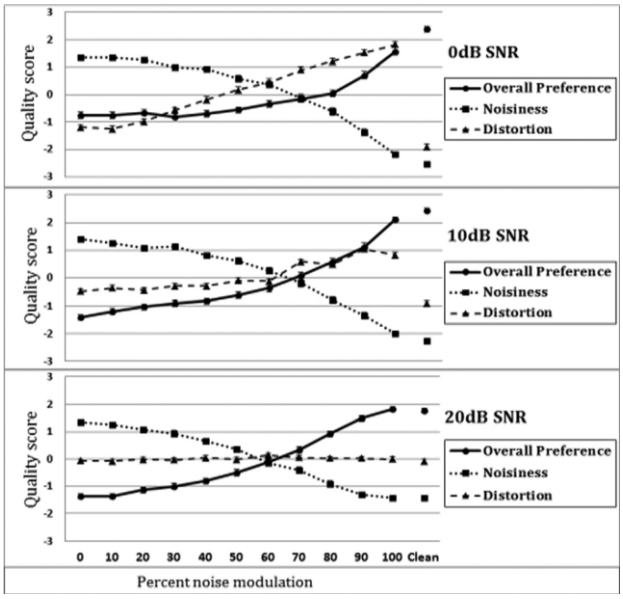

Figure 1. Average quality ratings for clean speech and eleven noisy speech stimuli. Three rating scales are shown: overall preference, noisiness, and distortion (from [12]).

Listeners rated the stimuli on three scales: overall preference, distortion, and noisiness. A six-point paired-comparison forced-choice method was used in which the listeners indicated which sentence of the A-B pair being presented they preferred and by how much. The rating scale ranged from +3 to -3, with +3 corresponding to "A is much preferred" and -3 corresponding to "B is much preferred". For example, if the listener selected "A is moderately preferred", then sentence A was assigned a score of +2 and sentence B was assigned a score of -2. The stimuli were presented in random order, and all stimuli were presented both in position A and in position B. The subjects rated the different SNRs in different test blocks, so stimuli at one SNR were never compared to stimuli at a different SNR.

The average subject ratings are presented in Fig 1. A repeated measures ANOVA showed that for each of the three SNRs, the ratings for overall preference and for noisiness were significantly different for the amounts of noise modulation. The ratings for distortion were significantly different for the 0- and 10-dB SNR but not for the 20-dB SNR. At all three SNRs, the overall preference increases as the amount of modulation increases even though the noise intensity averaged over the duration of the stimulus remains constant at each SNR. Thus modulated noise is preferred over stationary noise at any given SNR.

The preference for modulated noise cannot be predicted by the SNR since amount of modulation was varied while the SNR was held constant. The preference, however, is predicted by HASQI [5]. HASQI measures the signal envelope and spectral fidelity in comparison with an undistorted reference signal. It returns a value between 0 and 1, with 1 representing perfect fidelity and 0 indicating very low fidelity. HASQI was designed for use with both normal-hearing and hearing-impaired listeners and comprises a model of the auditory periphery followed by the extraction of signal features. The nonlinear term of HASQI correlates the time-frequency envelope modulation of the processed signal with that of the clean reference signal. The linear term measures the difference in the long-term spectra of the output and reference signals. The HASQI value is the product of the nonlinear and linear terms. In this experiment the long-term output spectrum was adjusted to match that of the reference signal, so the HASQI value is dominated by the envelope-modulation nonlinear term.

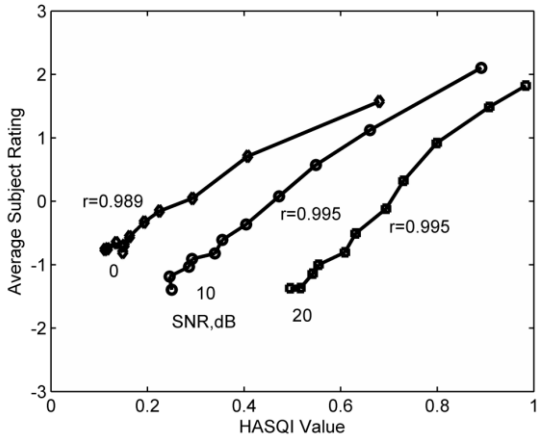

Figure 2. Correlation between HASQI predictions and average subject quality ratings for different amounts of noise envelope modulation.

The HASQI value was computed for each combination of SNR and amount of noise modulation used in the experiment. The subject quality ratings are plotted in Fig 2 as a function of the HASQI values. The ratings at each SNR are plotted as separate curves since the ratings were produced in separate test blocks, with each block referenced to its own perceptual anchors of best and worst quality. The Pearson correlation coefficient between the HASQI prediction and the subject ratings is extremely high, ranging from 0.989 for the 0-dB SNR to 0.995 for the 10- and 20-dB SNRs. Thus measuring the changes in envelope modulation provides quality information that the SNR misses and results in a much more accurate quality model.

3. NOISE DURING SPEECH SILENCES

In this experiment, the impact of noise audible during silences in the speech was measured for 16 subjects with normal hearing. The test material comprised a concatenated pair of sentences spoken by a male talker extracted from the hearing-in-noise test (HINT) [13] and a pair of sentences spoken by a female talker. There was a silent period of 400 msec before and after the pair of sentences and the gap between the two sentences was 250 msec. The sampling rate was 22.05 kHz. Subjects were tested monaurally using headphones, and the level of presentation was 65 dB SPL.

The stimuli were presented in quiet and at SNRs of 20, 15, 10, and 5 dB. The noise was either speech-shaped stationary noise or multi-talker babble. For each SNR and type of noise, the noise was either present continuously before and after the sentence pair and throughout the gap between the two sentences, or was turned off before, after, and in the gap between the sentence pair but was still present during the speech. Subjects rated the speech quality on a 10-point scale, with 1 representing poor quality and 10 representing excellent quality [14]. The subjects were asked to rate the overall quality, the pleasantness, and the noisiness of the speech. The stimuli were presented in random order. The data for the overall quality, which corresponds to the HASQI calculation, are presented here.

The subject ratings, averaged over the male and female talkers and then averaged over the listeners, are presented in Fig 3 for the stationary noise and in Fig 4 for the multi-talker babble. The "gap" condition refers to noise being removed before, after, and in the gap between the sentences in the pair. A repeated measures ANOVA showed that the differences in quality at each SNR were statistically significant. The type of noise (stationary or babble) and gap were not significant. There was, however, a significant two-way interaction between the type of noise and gap, and a significant three-way interaction between the type of noise, gap, and SNR. As can be seen in Fig 3, there is little difference between the overall quality ratings for the stationary HINT noise presented continuously or only when the speech signal is present, but there is a small but consistent preference in babble for the gap condition over the continuous interference.

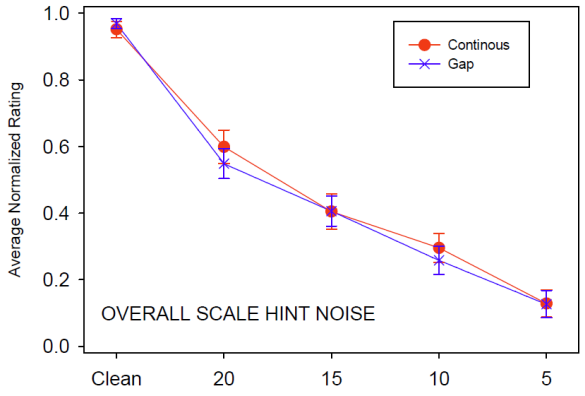

Figure 3. Overall quality ratings for continuous stationary background noise and noise that is eliminated during the silences and gaps in the speech. The ratings are averaged over talker and listener, and have been normalized to give a maximum value of 1 for each listener before averaging.

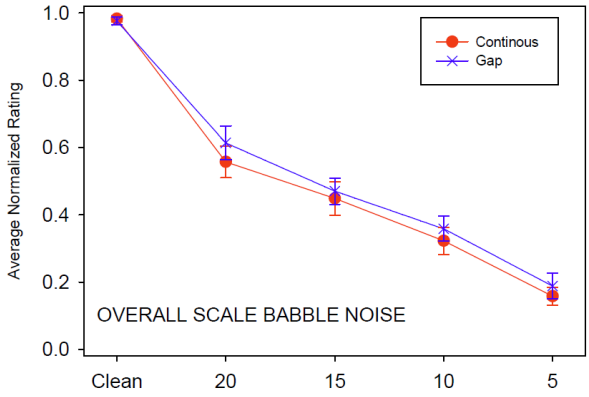

Figure 4. Same as Fig 3, except that the interference is multi-talker babble.

The HASQI calculation removes the silences in the clean speech and the corresponding samples of the noisy speech before computing the quality prediction. Because the silent intervals are ignored, HASQI predicts the same quality for noise present during the silent intervals as for noise gated off during those intervals. The speech ratings [2] used to calibrate HASQI had the noise gated off during the speech silences, so the HASQI prediction corresponds to the gap data in Figs 3 and 4. The data indicate that the HASQI model will also be accurate for speech in continuous stationary noise, but may overestimate the quality of speech in the presence of continuous babble.

4. DISCUSSION AND CONCLUSIONS

The results from both experiments show that noise modulation affects speech quality judgments. In the first

experiment, the overall speech quality increased as the envelope modulation of the noise became more similar to that of the speech. The distortion rating also increased as the noise modulation matched that of the speech, while the noisiness decreased with increasing envelope correlation. These results suggest that signal degradations that track the speech envelope are perceived as distortion, while degradations that are independent are perceived as noise. Measuring just the SNR is insufficient in determining these distinctions in noise perception and their impact on quality.

In the second experiment, the impact of noise during the speech silences depended on the type of noise. There was no significant difference between the speech quality for continuous versus gated stationary noise, while there was a small but significant difference for the multi-talker babble. The babble differs from the stationary speech-shaped noise both in its envelope modulation and in the signal temporal fine structure, so again the perceptual distinctions depend on noise characteristics that cannot be determined from the SNR alone.

The quality results also have consequences for metrics used to predict speech quality. The HASQI index, for example, removes the silent portions of the speech signal before computing the index. The justification is that the gap duration will change the sentence envelope modulation spectrum at low modulation frequencies. Including the silent portions of the speech also affects average values computed over the utterance [15]. However, the results presented here indicate that improved accuracy could be obtained by testing for the presence of noise during silences in the speech signal and then measuring the noise characteristics within these intervals.

5. RELATION TO PRIOR WORK

This paper presented new experimental results that illustrate the effect of noise modulation on speech quality judgments. Systematic studies of the effects of noise envelope modulation on quality and the effects of noise during silences in the speech materials have not been reported previously in the literature. The results show that the SNR is a poor predictor of speech quality [8] [9] [11], and indicates that metrics that measure changes in the speech envelope modulation [4] [5] can be much more accurate.

6. REFERENCES

[1] A. Leman, J. Faure, and E. Parizet, "Influence of informational content of background noise on speech quality evaluation for VoIP application," J. Acoust. Soc. Am. 123: 3066-3066, 2008.

[2] K.H. Arehart, J.M. Kates, and M.C. Anderson, "Effects of noise, nonlinear processing, and linear filtering on perceived speech quality," Ear and Hearing 31: 420-436, 2010.

[3] J.H.L. Hansen and B. L. Pellom, "An effective quality evaluation protocol," Proc. Int. Conf. Spoken Lang. Proc., Sydney: 2819-2822, 1998.

[4] R. Huber and B. Kollmeier, "PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception," IEEE Trans. Audio Speech and Lang. Proc. 14: 1902-1911, 2006.

[5] J.M. Kates, and K.H. Arehart, "The hearing aid speech quality index (HASQI)," J. Audio Eng. Soc. 58: 363-381, 2010.

[6] A. Das and J.H.L. Hansen, "Phoneme selective speech enhancement using the generalized parametric spectral subtraction estimator," Proc. Int. Conf. Acoust. Speech and Sig. Proc., Prague: 4648-4651, 2011.

[7] M.A. Haque, T. Islam, and Md. K. Hasan, "Robust speech dereverberation based on blind adaptive estimation of acoustic channels," IEEE Trans. Audio Speech and Lang. Proc. 19: 775-787, 2011.

[8] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," Speech Comm. 49: 588–601, 2007.

[9] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Trans. Acoust. Speech Signal Proc. 16: 229–238, 2008.

[10] K. Kokkinakis and P. C. Loizou, "Evaluation of objective measures for quality assessment of reverberant speech," Proc. Int. Conf. Acoust. Speech and Sig. Proc., Prague, 2420-2424, 22-27 May 2011.

[11] A.A. Kressner, D.V. Anderson, and C.J. Rozell, "Robustness of the hearing aid speech quality index (HASQI)," Proc. IEEE Workshop Appl. Sig. Proc. Audio Acoust., New Paltz, N.Y.: 209-212, 2011.

[12] I-K. Jin, J.M. Kates, and K.H. Arehart, "The effect of noise envelope modulation on quality judgments of noisy speech," J. Acoust. Soc. Am. 132: EL277-283, 2012.

[13] M. Nilsson, S. Soli, and J. Sullivan, "Development of the hearing in noise test for the measurement of speech reception threshold in quiet and in noise," J. Acoust. Soc. Am. 95: 1085-1099.

[14] International Telecommunication Union. 2003. ITU-R: BS. 1284–1, *General methods for the subjective assessment of sound quality*. Geneva: ITU.

[15] J.G. Beerends, A.P. Hekstra, A.W. Rix, and M.P. Hollier, "Perceptual Evaluation of Speech Quality (PESQ) the new ITU standard for end-to-end speech quality assessment Part II - Psychoacoustic model," J. Audio Eng. Soc. 50: 765-778, 2002.