2pPPb26. Relationship between distortion and working memory for digital noise-reduction processing in hearing aids

Kathryn Arehart*, Pamela Souza, Thomas Lunner, Michael Syskind Pedersen and James Kates

*Corresponding author's address: Speech, Language, Hearing Sciences, University of Colorado Boulder, 409 UCB, Boulder, Colorado 80309, kathryn.arehart@colorado.edu

Several recent studies have shown a relationship between working memory and the ability of older adults to benefit from specific advanced signal processing algorithms in hearing aids. In this study, we quantify tradeoffs between benefit due to noise reduction and the perceptual costs associated with distortion caused by the noise reduction algorithm. We also investigate the relationship between these tradeoffs and working memory abilities. Speech intelligibility, speech quality and perceived listening effort were measured in a cohort of elderly adults with hearing loss. Test materials were low-context sentences presented in fluctuating noise conditions at several signal-to-noise ratios. Speech stimuli were processed with a binary mask noise-reduction strategy. The amount of distortion produced by the noise reduction algorithm was parametrically varied by manipulating two binary mask parameters, error rate and attenuation rate. Working memory was assessed with a reading span test. Results will be discussed in terms of the extent to which intelligibility, quality and effort ratings are explained by the amount of distortion and/or noise and by working memory ability.[Funded by NIH, Oticon and GN ReSound]

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Noise reduction algorithms aim to reduce adverse effects of noise on communication by listeners with hearing loss. Such algorithms are intended to improve speech intelligibility, but may also affect speech quality and ease of listening. For example, Sarampalias et al. (2009) showed that a noise reduction algorithm (Ephraim and Malah, 1984) decreased listening effort as evidenced by improved performance in a memory task for words in listeners with normal hearing. Some of our previous work has demonstrated a relationship between listener characteristics—specifically, working memory—and signal distortion (Arehart et al., in press). This work, together with other recent research, suggests that poor working memory may result in a general susceptibility to signal distortion. With regard to noise reduction, recent reports (Ng et al., 2010; Rudner et al., 2012) suggest a relationship between working memory and benefit from noise reduction for listeners with hearing loss. That implementation of noise reduction utilized a binary time-frequency masking technique (Wang et al., 2009) with one specific implementation of processing parameters. To expand upon that work, we hypothesize that the effects of working memory on speech intelligibility, speech quality, and ease of listening may be related to tradeoffs in signal-to-noise ratio (SNR) improvement vs. the amount of distortion produced by the signal processing. That is, a comparison is needed in which the aggression of signal processing is evaluated in terms of the amount of SNR improvement and the costs associated with distortions in processing. As a first step in evaluating these tradeoffs, this paper evaluates the cumulative effects of processing using several objective measures, including spectrograms, estimates of SNR improvements, a coherence-based metric of speech intelligibility and an envelope-based metric of speech quality.

METHODS

Stimuli

The stimuli consisted of two pairs of concatenated sentences from the Hearing-in-Noise Test (HINT) (Nilsson et al., 1994). The first set included two sentences spoken by a male talker (“The boy got into trouble. The yellow pears taste good.”). The second set included two sentences spoken by a female talker (“Two cats played with yarn. She needs an umbrella.”). The sentences were digitized at a 44.1 kHz sampling rate and downsampled to 20 kHz for the binary mask processing. Speech was mixed with one of two types of background noise. The first noise type was four-talker babble consisting of two male and two female American English talkers. The level of each talker was adjusted so that all talkers were weighted equally. The second noise was stationary speech-shaped noise from the HINT. The SNR for both the babble and steady-state noises ranged from +10 dB to -30 dB in steps of 1 dB.

Binary Mask

The mixed speech plus noise stimuli were processed with a binary mask noise-reduction strategy (Kjems et al., 2009). The target speech signal, the masker signal and the target+masker mixture were separately converted into the frequency domain by analysis filterbanks consisting of 128 Gammatone filters with center frequencies equally distributed on the ERB scale in the frequency range between 50 and 8000 Hz. The processing was done in time frames each having a duration of 20 ms with an overlap of 10 ms resulting in frame shifts every 10 ms.

In each time-frequency unit, the local SNR was determined by comparing the intensities of the separate target and noise signals. The local SNR was then compared to a local criterion (LC) of 0 dB, resulting in an ideal binary mask decision equal to 1 if the local SNR was above LC, and 0 otherwise. The data of Kjems et al. (2009) indicate that a LC of 0 dB is most effective for SNRs in the range of approximately +5 to -10 dB. Similar to the procedure in Li and Loizou (2008), errors were introduced into the ideal binary mask by randomly flipping a certain percentage (0%, 5%, 10%, 20% or 30%) of the time-frequency units either from 0 to 1 or from 1 to 0. The binary patterns were converted into gain values, where values of 1 were converted into 0 dB gain and values of 0 were converted into an attenuation of either 10 dB or 100 dB. The noisy speech signal was then multiplied by the binary gain values to give the processed signal in the frequency domain. The processed signal was synthesized back into the time domain by use of a time-reversed Gammatone filterbank, thereby ensuring a constant group delay.

The effect of the binary mask processing is illustrated in the spectrograms in Figure 1. Figure 1a shows the spectrogram for sentences spoken by the female talker in quiet. The spectrogram for the same pair of sentences, but now with additive four-talker babble at an SNR of -5 dB, is presented in Figure 1b. The babble fills in the gap between the sentences, blurs the onsets of the speech sounds, and hides many of the low-intensity speech components. However, many of the harmonics in the formant areas are still visible. The result of applying the ideal
binary mask (that is, no errors in the gain pattern) with the mask attenuation set to 100 dB is shown in Figure 1c. Those regions of the speech spectrogram where the noise is more intense than the speech have been attenuated, restoring some of the speech envelope structure and resulting in a spectrogram that is a closer match to that shown in Figure 1a for the clean speech. Note, however, that some of the low-intensity speech components, particularly at high frequencies, have been removed by the binary mask as well.

**Estimated SNR**

The output SNR was computed using the magnitude-squared coherence (MSC) (Kates, 1992). The reference signals, originally sampled at 44.1 kHz, were downsampling to 22.05 kHz. Because output of the binary mask processing was at 20 kHz, the processed signals were upsampling to 22.05 kHz for compatibility with the references. The cross-correlation of the reference and processed signals was computed over the duration of the utterances. The mean-squared energy of each signal was also computed. The MSC is given by the squared magnitude of the cross-correlation divided by the product of the signal energies. The MSC was converted to the signal-to-distortion ratio (SDR) using the relationship $\text{SDR} = \frac{\text{MSC}}{1 - \text{MSC}}$ (Kates, 1992), and the SDR was then converted to dB.

**CSII Intelligibility**

The coherence SII (CSII) (Kates and Arehart, 2005) is an intelligibility index based on a modification of the Speech Intelligibility Index (SII) (ANSI 1997). It estimates the fraction of sentences understood correctly for noisy and distorted speech. The speech was divided into 16-ms segments having a 50% overlap. Each segment was multiplied by a raised-cosine Hamming window. The power in each segment was computed, and the segments were assigned to one of three levels: low-level (-30 to -10 dB re: RMS), mid-level (-10 to 0 dB re: RMS), and high-level (greater than 0 dB re: RMS) where RMS is the RMS level of the entire utterance. The short-time FFT was also computed for each segment. The MSC (Kates, 1992) was then computed over all of the segments in the low-, mid-, and high-level groups to give three MSC values. In the case of the signal being entirely attenuated by the processing, the resultant MSC is zero. Each MSC was converted to the SDR, and the SDR was then converted to dB. The SII was computed for each intensity region, giving a set of CSII values. The intelligibility index $I_3$ for normal hearing was then given by a weighted combination of the CSII values, followed by a logistic function transformation:

$$c = -3.47 + 1.84 \text{CSII}_{\text{Low}} + 9.99 \text{CSII}_{\text{Mid}} + 0.0 \text{CSII}_{\text{High}}$$

$$I_3 = \frac{1}{1 + e^{-c}}$$

The frequency band spacing and segment size used in the CSII calculation was applied independently of the frequency analysis and segmentation used for the binary mask.

Some recent studies have found that the CSII is a poor predictor of intelligibility for binary-masked speech (Christiansen et al., 2010; Taal et al., 2011). The processing used in those papers matched the binary mask LC to the SNR of the noisy signal, while the experiment in this paper used a fixed local criterion of 0 dB independent of the SNR. In the limit of a -60 dB SNR, the processing used in the earlier papers produced a binary-modulated noise that still approximated the envelope of the clean speech and yielded high intelligibility. However, the CSII, which is based on the signal coherence, gives a near-zero value for the -60 dB SNR condition since the temporal fine structure of the clean and modulated noisy signals is uncorrelated. On the other hand, in the present experiment, the local criterion was kept constant at 0 dB independent of the SNR, so the number of 1s in the binary gain pattern decreases as the SNR decreases. In the case of -60 dB SNR and 100-dB attenuation, the processing implemented in this paper would attenuate the entire signal since all cells would have negative SNRs, giving an inaudible output having no intelligibility. This result is consistent with a CSII of zero. Because of these experimental differences, the CSII would be expected to be an accurate predictor of intelligibility for the processing used in this paper.

**HASQI**

The Hearing Aid Sound Quality Index (HASQI) (Kates and Arehart, 2010) was used to quantify the total amount of signal alteration caused by the binary mask and additive noise. HASQI estimates the speech quality for sentence materials, and normal hearing was assumed. HASQI measures signal envelope and spectral fidelity in comparison
with an undistorted reference signal, with the nonlinear term measuring the envelope fidelity. It returns a value between 0 and 1, with 1 representing perfect fidelity and 0 indicating very low fidelity. Additive noise, for example, will fill in the valleys of the speech signal. The noise changes the envelope peak-to-valley ratio and thus reduces the envelope correlation between the noisy signal and a clean reference. A binary mask will reduce the gain of individual time-frequency tiles: a tile containing speech that receives a reduced gain will reduce the envelope correlation, while reducing the gain of a tile that is predominantly noise may improve the envelope correlation. Compared to the HASQI value for noisy speech without further processing, binary-masked noisy speech may show an improved envelope correlation. The frequency band spacing and segment size used in the HASQI calculation was applied independently of the frequency analysis and segmentation used for the binary mask.

RESULTS AND DISCUSSION

The results of the binary mask are illustrated in Figure 2 for speech with additive 4-talker babble. Results for speech with additive speech-shaped stationary noise were similar. The panels in the left column (2a, 2b, 2c) are for the binary mask with 100-dB attenuation, and the panels in the right column (2d, 2e, 2f) are for the binary mask with 10-dB attenuation. The estimated SNR is plotted in Figure 2a as a function of the actual SNR for an attenuation of 100 dB. The estimated SNR for no processing agrees with the actual SNR until the -25 dB limit used in the calculation is reached. The estimated SNR for the ideal binary mask (no errors) shows a substantial improvement in SNR, with speech input at an SNR of -30 dB emerging from the ideal processing with an estimated SNR of just -7 dB. Adding errors to the binary pattern produces estimated SNR values intermediate between the ideal and unprocessed stimuli. The somewhat irregular nature of the error curves is the result of the random error assignments. A 5 percent error rate has little effect at -10 dB and above, but has a much stronger effect as the SNR decreases below -10 dB and is equivalent to no processing at the experimental limit of -30 dB SNR. Even with an error rate of 30 percent, the binary mask improves the estimated SNR, even though the effect is small. The estimated SNR is plotted in Figure 2d as a function of the actual SNR for an attenuation of 10 dB. The estimated SNR benefit of the binary mask is smaller for the 10-dB attenuation than for the 100-dB attenuation. In particular, the ideal binary mask no longer provides the substantial improvement at very low SNRs that was found for the 100-dB attenuation.

The speech intelligibility for normal hearing is estimated using the CSII. The estimated CSII intelligibility is plotted in Figure 2b as a function of the actual SNR for an attenuation of 100 dB. The predicted intelligibility for no processing shows no speech understanding would be expected for SNRs of -10 dB or lower, and perfect intelligibility would be achieved at an SNR of 10 dB. The ideal binary mask provides an intelligibility index of about 0.98 at an SNR of -5 dB, and still gives some intelligibility at an SNR of -30 dB. The intercept for 70 percent intelligibility has shifted from about 2 dB for no processing to about -17 dB for the ideal binary mask, a projected improvement of about 19 dB. The introduction of errors into the binary mask pattern results in predicted performance levels that are between the ideal mask and no processing. Adding a 5 percent error rate causes a substantial reduction in processing benefit at SNRs below -5 dB, but even with 30 percent error rate there is a predicted intelligibility benefit for the binary mask. The estimated CSII is plotted in Figure 2e as a function of the actual SNR for an attenuation of 10 dB. As for the estimated SNR, the CSII benefit of the binary mask is smaller for the 10-dB attenuation than for the 100-dB attenuation.

The speech quality for normal hearing listeners is estimated using HASQI. The estimated HASQI nonlinear quality term is plotted in Figure 2c as a function of the actual SNR for an attenuation of 100 dB. The predicted quality is lower than the predicted intelligibility, with the unprocessed sentences at a 10-dB SNR giving a quality index of about 0.3 even though the intelligibility index was nearly 1. The ideal binary mask increases the predicted quality over the entire range of SNRs, giving a small improvement of about 0.07 at -30 dB and increasing with SNR to an improvement of 0.22 at 10 dB. The HASQI improvement indicates that the time-frequency envelope of the processed speech is closer to that of the reference signal than is the envelope of the noisy speech. Introducing errors into the binary mask pattern again results in performance intermediate between the ideal mask and no processing. At SNRs below -10 dB the fraction of errors appears to make little difference, suggesting that the specific errors may be more important than the average number of errors. At SNRs above -10 dB, increasing the error rate more consistently leads to reduced quality. The HASQI nonlinear term is plotted in Figure 2f as a function of the actual SNR for an attenuation of 10 dB. The benefit of the ideal binary mask is lower for 10-dB attenuation than for 100-dB attenuation, but the benefit when errors are introduced into the processing is quite similar.

The HASQI quality index is plotted in Figure 3 as a function of the estimated SNR for the binary mask with 100-dB attenuation. This plot relates an envelope-based measurement (HASQI) to a coherence-based measurement (estimated SNR) for the binary mask processing. The solid line shows the relationship between HASQI and the estimated SNR for the unprocessed condition; the unprocessed curve in Figure 2a shows that estimating the SNR
using coherence is accurate over a wide range of noise levels for the unprocessed stimuli. The points for the binary mask processing are consistent with the curve for the unprocessed data, which indicates that the estimated SNR accurately summarizes the interaction of the noise reduction and the noise intensity for the processing used in this study. That is, the same estimated SNR always corresponds to the same HASQI prediction of speech quality independent of the combination of noise intensity and processing parameters used. The predictive value of the estimated SNR indicates that coherence can accurately be used to measure the processing effects, and that the computed CSII values are also valid for this study. Similar results were observed for the binary mask using 10-dB attenuation.

CONCLUSIONS

This paper has presented an acoustic analysis of noisy speech processed with a binary-mask noise suppression algorithm. The binary mask processing included the ideal binary mask and binary masks with random errors in the mask pattern. The analysis showed that the estimated SNR, estimated intelligibility, and estimated speech quality all improved when the binary mask processing was applied to the noisy signal. The metrics thus predict that signal fidelity for the noisy speech signals will be better for all binary mask conditions, even those that include significant errors. These predictions, however, are based on models developed for a group of normal hearing listeners, and do not consider individual differences that may occur due to hearing thresholds, suprathreshold processing, and working memory abilities. We are now extending our analysis to consider these individual differences and to further explore the interaction of signal-processing benefit and the nonlinear distortion introduced by the processing.

ACKNOWLEDGEMENTS

Work supported by NIH R01 DC012289, by a research grant to University of Colorado from GN ReSound and by Oticon.

REFERENCES


FIGURE 1. Spectrograms for a pair of sentences spoken by a female talker with (a) no noise or binary mask applied, (b) with additive four-talker babble at -5 dB SNR but with no binary mask applied, and (c) with additive stationary speech-shaped noise at -5 dB SNR with ideal binary mask applied with no masking errors.
FIGURE 2. Performance metrics for speech with additive 4-talker babble. The left column is for binary mask processing with 100-dB signal attenuation, and the right column is for binary mask processing with 10-dB signal attenuation. The legend gives the binary mask pattern error in percent cells randomly changed, and No Proc refers to the unprocessed noisy speech.
FIGURE 3. The HASQI nonlinear term plotted as a function of the estimated SNR for the unprocessed noisy sentences (solid line) and for the binary mask processing. The legend gives the binary mask pattern error in percent cells randomly changed.