Determination of the Potential Benefit of Time-Frequency Gain Manipulation

Michael C. Anzalone, Lauren Calandruccio, Karen A. Doherty, and Laurel H. Carney

Objective: **The purpose of this study was to determine the maximum benefit provided by a time-frequency gain-manipulation algorithm for noise-reduction (NR) based on an ideal detector of speech energy. The amount of detected energy necessary to show benefit using this type of NR algorithm was examined, as well as the necessary speed and frequency resolution of the gain manipulation.**

Design: **NR was performed using time-frequency gain manipulation, wherein the gains of individual frequency bands depended on the absence or presence of speech energy within each band. Three different experiments were performed: (1) NR using ideal detectors, (2) NR with nonideal detectors, and (3) NR with ideal detectors and different processing speeds and frequency resolutions. All experiments were performed using the Hearing-in-Noise test (HINT). A total of 6 listeners with normal hearing and 14 listeners with hearing loss were tested.**

Results: **HINT thresholds improved for all listeners with NR based on the ideal detectors used in Experiment I. The nonideal detectors of Experiment II required detection of at least 90% of the speech energy before an improvement was seen in HINT thresholds. The results of Experiment III demonstrated that relatively high temporal resolution (<100 msec) was required by the NR algorithm to improve HINT thresholds.**

Conclusions: **The results indicated that a single-microphone NR system based on time-frequency gain manipulation improved the HINT thresholds of listeners. However, to obtain benefit in speech intelligibility, the detectors used in such a strategy were required to detect an unrealistically high percentage of the speech energy and to perform the gain manipulations on a fast temporal basis.**

(Ear & Hearing 2006;27;480–492)

One of the greatest problems facing listeners with hearing loss is understanding speech in the presence of background noise. Although current hearing aid technology improves the audibility and comfort of speech in noisy backgrounds, improvement in the intelligibility of speech is limited. One general approach to noise-reduction (NR) is a strategy known as time-frequency gain reduction. Studies using this strategy have shown mixed results in intelligibility, with some showing benefits (Rankovic, Freyman & Zurek, 1992; Stein & Dempesy–Hart, 1984), and others showing none (Fabry & Van Tasell, 1990; Klein, 1989). It is difficult to evaluate the general strategy from these studies, because different speech detectors were used in each study. In the current study an ideal detector was used; that is, the detector was based on complete knowledge of the speech signal to be detected. This is not a realistic detector, because the signal to be detected is typically unknown. However, use of an ideal detector allows determination of the upper limit of noise reduction based on a particular time-frequency gain manipulation strategy.

Quantitatively, a listener's understanding of speech in noise can be measured using a speech reception threshold (SRT). The SRT is a measure of the signal-to-noise ratio (SNR) that is required to achieve a preset level of intelligibility, generally 50 or 100 percent (Moore, 2003). The SRTs of listeners with hearing loss are increased relative to listeners with normal hearing. For speech-spectrum shaped noise, the increase is 2 to 5 dB (Plomp, 1994), whereas the SRT increases 7 to 15 dB when the noise amplitude fluctuates (Eisenberg, Dirks & Bell, 1995; Takahashi & Bacon, 1992) or is a competing speaker (Carhart & Tillman, 1970).

There has been extensive research in the development of noise-reduction algorithms. The general goal of all noise reduction algorithms is to restore the SRT of listeners with hearing loss to that of listeners with normal hearing. This goal is two-fold: to restore lost intelligibility and to improve the quality of noisy speech (Schum, 2003). Single-microphone NR systems perform the second goal of improving quality, without increasing intelligibility (Chabries & Bray, 2002; Levitt, 2001; Schum, 2003). Directional microphones are the only NR strategy that meets the first goal (Chabries & Bray, 2002; Levitt, 2001; Schum, 2003). Although directional microphones can improve intelligibility when examined in a laboratory setting, the presence of reverberation in real-world listening environments limits the SRT improvement to a few decibels (Hawkins & Yacullo, 1984; Ricketts, 2000; Ricketts & Hornsby, 2003). In addition, the directional microphone requires that the noise and speech be spatially separated to maximize the benefit.

Department of Biomedical and Chemical Engineering (M.C.A., L.H.C.), Department of Speech and Communication Disorders (L.C., K.A.D.), and Department of Computer Science (L.H.C.), Institute for Sensory Research, Syracuse University, Syracuse, New York.

0196/0202/06/2705-0480/0 • Ear & Hearing • Copyright © 2006 by Lippincott Williams & Wilkins • Printed in the U.S.A.

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

The single-microphone NR system attempts to increase intelligibility by increasing the SNR of the speech. Several strategies that have been used to accomplish this include Wiener filtering (Wiener, 1949), spectral subtraction (Boll, 1979), adaptive filtering (Graupe, Grosspietsch & Basseas, 1987), and speech synthesis (Kates, 1994; McAulay & Quatieri, 1986). Wiener filtering involves estimating the characteristics of the signal and the noise, and creating a filter that optimizes the SNR at the output of the filter. In spectral subtraction, the spectrum of the noise is estimated and subtracted from the noisy signal, leaving only the spectrum of the speech (Boll, 1979). The adaptive filtering strategy is similar to Wiener filtering, but uses a time-varying filter that is varied based on the difference of the filter output and a noise-estimate. Speech synthesis is a NR strategy that replaces the speech detected in the original signal with speech that has been synthesized. A common form of this, called sine wave modeling, replaces the detected components of speech with sine waves matched to the amplitude and frequency of the detected components (Kates, 1994; McAulay & Quatieri, 1986).

All of these single-microphone NR strategies depend on having an accurate model of the noise and/or speech. Such models are difficult to create or measure, given that noise varies greatly across everyday listening environments. Also, the speech signal varies across speakers. Although these strategies have been shown to improve the SNR of the noisy signal, none have significantly improved intelligibility, primarily because the processing adds distortions to the signal (Boll, 1979; Kates, 1994; Levitt, 2001; McAulay & Quatieri, 1986; Schum, 2003).

Another frequently used single-microphone NR strategy is time-frequency gain manipulation, which has been reported to increase SNR and intelligibility in some, but not all cases (Rankovic et al., 1992; van Dijkhuizen, Festen & Plomp, 1989, 1990). In this strategy, the gain of each frequency band is time-varying; when the SNR within the band is high (e.g., favorable), the gain is high, when the SNR is low, the gain for that frequency band is reduced. This strategy has been shown to be effective when the noise is limited to one frequency band (Rankovic et al., 1992), but the results conflict when the noise is spread over a wide band of frequencies (Fabry & Van Tasell, 1990; Klein, 1989; Ono, Kanzaki & Mizoi, 1983). There have been several implementations of this strategy, all of which have used their own, unique speech-component detectors to determine when to change the gain within each frequency band. The differences in detection strategy across different studies complicate the evaluation of the general strategy of time-frequency gain manipulation. The use of an ideal detector in this study avoided this issue.

The ideal detector was used for each frequency channel, and time-frequency gain manipulation was carried out using a binary mask. Within each frequency channel and time window, a binary ("yes" versus "no") decision was made as to whether or not speech was present based on the spectrogram of the non-noisy speech. This decision was determined based on a two-dimensional "mask", which was created by giving each "pixel" that contained speech energy within a specific time window and given frequency channel a high value (for example, "1"), and pixels that did not contain speech energy a low value. The resulting binary mask can be thought of as a two-dimensional representation of the presence of a speech component in a given frequency band at a given time. This binary mask is applied to noisy speech by assigning different gains to the two different values in the mask; for example, a gain of 1 can be applied to a given frequency channel during time windows when the value of the binary mask is high, and attenuation can be applied when the value of the mask is low. Thus, the ideal binary mask effectively separates the speech from the noise by preserving the signal when speech is present, and attenuating it when speech is not present. Binary masks allow for quantitative comparison of time-frequency gain patterns produced by different systems. Binary masks have been shown to increase the performance of automatic speech recognition in noise (Cooke, Green, Ljubomir, et al., 2001; Srinivasan, Roman & Wang, 2004), as well as to separate acoustical sources (Roman, Wang & Brown, 2003; Yilmaz & Rickard, 2004). Although performance of these systems improves because of the increased overall SNR produced by the binary mask, the binary masks produce several other modifications to the stimuli that a human listener could use to improve intelligibility. These modifications include, in addition to the improvement in the SNR, a decrease in the spread of masking and an increase in the envelope cues available.

The overall speech SNR should be improved by application of the ideal binary mask because the noise is attenuated during periods when the speech is not present. The frequency bands that contain speech components during a given period, however, will be preserved, along with any noise present in that frequency band at that time. The articulation index (AI) and speech intelligibility index (SII) (ANSI, 1969; ANSI, 1997; French & Steinberg, 1947; Mueller & Killion, 1990; Pavlovic, 1988) suggest that the SNR of each frequency band contributes to the overall intelligibility of speech, provided that the band is audible; if there is no change in the bands' SNRs, there should not be an increase in intelligibility. The effect of time-frequency gain manipulation on the SNR in individual frequency bands

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

depends on the frequency resolution of the binary mask. If the frequency resolution of the binary mask is high, the SNRs of the individual frequency bands may be improved, leading to an increase in intelligibility. If the frequency resolution of the binary mask is low, the SNRs of the individual frequency bands are unaffected, as the binary mask preserves both the speech and noise within a band when speech is present, suggesting that no improvement in intelligibility will result.

Regardless of the frequency resolution of the binary mask, the application of the mask will result in attenuation of frequency bands that do not contain speech components. Reducing the amplitude of these frequency bands should decrease the spread of masking in adjacent bands that may contain speech components. Listeners with hearing loss have been shown to be susceptible to upward spread of masking (Klein, Mills & Adkins, 1990; Trees & Turner, 1986). By reducing the amount of masking caused by adjacent frequency bands, the intelligibility of the speech may be increased.

Application of the binary mask may also result in enhancement of certain speech cues (e.g., onsets and offsets of speech). The application of a speech signal's envelope to a noise carrier has been shown to provide the necessary cues for comprehension of simple speech in quiet for listeners with normal hearing when there are greater than four frequency bands representing the range of speech (Shannon, Zeng, Kamath, et al., 1995). For adverse SNRs, the ideal binary mask acts to impose a crude version of the speech components' envelopes onto the noise for each of the frequency bands. The envelopes imposed by the binary mask are crude because of the binary nature of the mask; the gains are allowed to only take on the "on" and "off" values, instead of the full values of the envelope.

The purpose of this study was to apply several versions of the binary mask, a single-microphone NR strategy, to noisy speech and examine the effect on speech intelligibility. It was hypothesized that speech intelligibility, as indirectly measured using the Hearing-in-Noise test (HINT) (Nilsson, Soli & Sullivan, 1994), would improve for the ideal binary mask compared with unprocessed stimuli. The HINT provides an estimate of the SRT, which is the SNR needed for the listener to correctly identify 50% of the words in a given sentence. Three experiments were performed: Experiment I involved the application of the ideal binary mask to the noisy speech to determine the maximum intelligibility achieved, Experiment II used degraded versions of the ideal binary mask to determine the amount of energy detected needed to see an improvement, and Experiment III varied the frequency resolution and processing speed of the ideal binary mask.

METHODS

Signal Processing

The processing that was performed falls into the general category of time-frequency gain manipulation. Stimuli were filtered into separate frequency bands. The gain of each band was dynamically changed depending on the presence or absence of speech energy within that band as a function of time. The bands were then recombined to form the final, processed output. The processing consisted of three main stages: the separation of the input stimuli into frequency bands, the detection of the speech components within each band, and the application of the gain changes to each band. Each of these stages will be discussed in more detail. All of the processing was performed offline using MATLAB (MathWorks, Natick, MA).

To separate the incoming stimuli into separate frequency bands, the NR algorithm used a Gammatone filter bank (Hohmann, 2002). The output of this filter bank, which is based on the tuning characteristics of the auditory system, resulted in separate frequency bands that were equally spaced on an equivalent rectangular bandwidth (ERB) scale (Glasberg & Moore, 1990). The ERB scale is based on the frequency-dependant bandwidths of auditory filters for human listeners. For Experiments I and II, filters with a bandwidth of 1 ERB were spaced at one-half-ERB intervals over the range of 70 to 7000 Hz. For Experiment III filters with a bandwidth of 1, ERB were spaced at either one-half-ERB or 1-ERB intervals over the same frequency range as for the first two experiments.

The second stage of processing required the detection of speech components within each of the separate frequency bands produced by the Gammatone filter bank. Rather than use the noisy stimuli, the detectors were "perfect" in that they were based on the original, non-noisy samples of the stimuli. The ensemble of detection patterns for a given stimulus for all frequency bands and times was the binary mask for that stimulus: The mask had two states, either present or not present. To generate a binary mask, a spectrogram was computed by filtering the non-noisy speech with a second filter bank that consisted of 4th-order Gammatone filters with one-half-ERB bandwidths. The spacing of the spectrogram's frequency bands matched that of the Gammatone filter bank used in the first stage. An example of a spectrogram computed in this way is shown in Figure 1A for the sentence "Her shoes were very dirty." The darkened regions in the figure represent the energy present in the non-noisy speech for each frequency

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

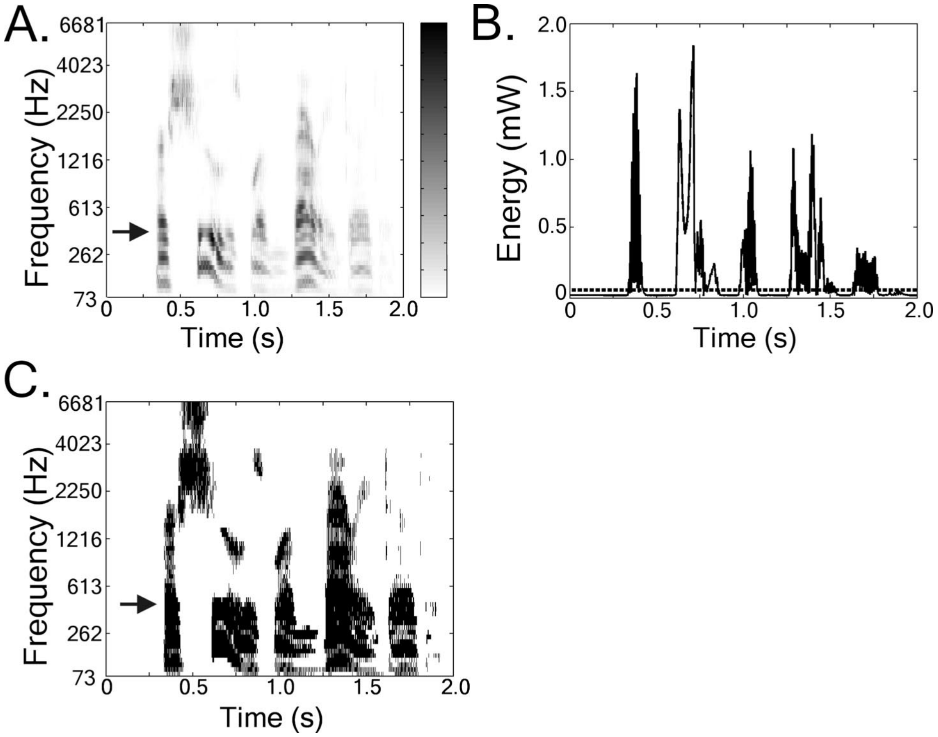

Fig. 1. Ideal binary mask generation. (*A*), Spectrogram of the speech in quiet ("Her shoes were very dirty") produced by filtering the speech with a filter bank with center frequencies matched to the analysis filter bank of the NR algorithm, with one-half-ERB bandwidths to reduce overlap. (*B*), The global criterion used to detect 99% of the speech energy (*dashed line*) is illustrated on a plot of energy versus time for the 414-Hz center frequency band (*arrow* in *A*). The gain in the binary mask (*B*) for this frequency band was set to 1.0 when the energy exceeded the criterion and to 0.2 when the energy was below the criterion. (*C*), The ensemble of gains (the ideal binary mask) shown in a manner similar to the spectrogram. Dark areas represent time periods and frequency bands for which the gain is 1.0. To create the mask, the filter output (*A*) was squared and filtered with a 300-Hz low-pass filter.

band of the spectrogram. An example of the energy present in one frequency band with a CF of 414 Hz (denoted by the arrow in Fig. 1A) is shown in Figure 1B. The energy of each frequency band was computed by squaring the output of each spectrogram filter and low-passed filtering with a 4th-order Butterworth filter that had a cut-off frequency of 300 Hz. To determine if speech was present within this frequency band, a criterion was applied to the energy within that band, as illustrated by the dashed lines in Figure 1B. Speech was considered to be present if the energy in the band exceeded the criterion. The same criterion was applied to all of the frequency bands of the spectrogram to produce the binary mask. The criterion was chosen for each sentence such that a fixed percentage of the total energy contained in the entire stimulus was above the criterion. The percentage was fixed at 99% for Experiments I and III to generate ideal binary masks for each sentence and changed from 75 to 99% to generate nonideal binary masks for Experiment II. Note that each sentence had a unique ideal binary mask; the mask was determined by the specific patterns within each sentence. The ideal binary mask produced by the spectrogram shown in Figure 1A is shown in Figure 1C. The dark regions represent periods when speech components were present, as determined by the application of the criterion to the spectrogram's frequency bands. Comparing the ideal binary mask to the spectrogram (Fig. 1A), one can see that the ideal binary mask simply identified periods of time and frequency when energy was present in the speech signal.

The third stage of processing consisted of the application of the binary mask generated in the second stage. The binary mask was applied by changing the gain within each frequency band with time, depending on the state of the binary mask. The gains used in the processing were allowed to take on two values: 0.2 or 1.0. If the binary mask indicated that a speech component was present, the gain was set to 1.0, otherwise the gain was set to 0.2. Examining Figure 1C, the gains were set to 1.0 for those frequency bands and times denoted by the dark regions; gains were set to 0.2 (i.e., attenuated) for frequency bands and times indicated by white regions. Although a larger SNR gain could have been achieved by setting the gain to zero when the binary mask indicated that no speech components were present, doing so would have decreased the overall quality of the stimulus because of the introduction of "musical" noise. Musical noise refers to random, short tone-like bursts that can be more bothersome than the original noise. Limiting the attenuation to less than 20 dB by using a gain of 0.2 when speech was absent reduced the amount of musical noise introduced by the time-frequency gain manipulation (Berouti, Schwartz & Makhoul, 1979).

To reconstruct the final signal, each frequency band was delayed and scaled such that the peaks of each band's impulse response had a maximum at 4 msec (Hohmann, 2002). All of the frequency bands were then added together to obtain a single waveform. The effect of the processing is shown in Figure 2. A spectrogram is shown in Figure 2A for the sentence "Her shoes were very dirty" at an SNR of 0

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

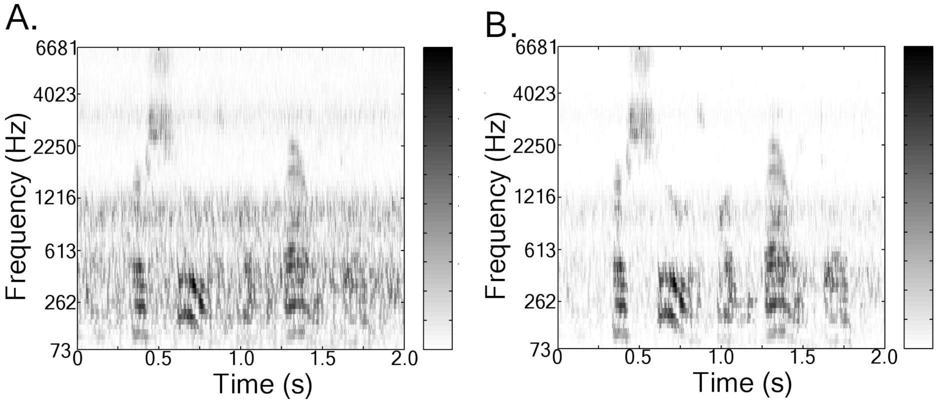

Fig. 2. Application of the ideal binary mask. (*A*), Spectrogram for the sentence of Figure 1 ("Her shoes were very dirty") for speech-spectrum noise added at a SNR of 0 dB. (*B*), Spectrogram of sentence after application of the ideal binary mask shown in Figure 1B. Note that the noise between speech components is attenuated by application of the ideal binary mask. Because the binary mask was applied on a sample-by-sample basis (with 50 μsec sampling time), the reduction occurred both between words as well as within the words themselves.

dB, with long-term speech spectrum noise. The spectrogram of the output of the NR algorithm is shown in Figure 2B, after the application of the ideal binary mask for this particular sentence (Fig. 1C). The result of the processing is the reduction of the noise between the speech components identified by the binary mask.

Experiment I: Ideal Binary Mask

Stimuli were processed for four conditions of noise-reduction: (1) Unprocessed (UNP): the stimuli were simply passed through the Gammatone filter bank without any manipulation of the frequency bands' gains. This condition served as a control, as the analysis-synthesis bank added some minor distortions to the signals and band limited the signal (Hohmann, 2002); (2) Low-frequency (LF) condition: the algorithm was only applied to lower frequencies (from 70 to 1500 Hz), and the remaining frequency bands were passed without modification; (3) High-frequency (HF) condition: the algorithm was only applied to higher frequencies (1.5 to 7 kHz), and the lower frequency bands were passed without modification; (4) All frequency (AF): the noise-reduction algorithm was applied to all frequency bands (70 Hz to 7 kHz).

For all four conditions, the frequency resolution of the binary mask and analysis/synthesis filter bank was set at 2 filters per ERB. The binary mask was applied on a sample-by-sample basis, as determined by the energy contained in the speech signal. The time-constant of the low-pass filter used to determine the energy within each band was always 0.53 msec.

Experiment II: Non-Ideal Binary Masks

In the first experiment, the criterion used to generate the ideal binary mask for each sentence was based on preservation of 99% of the total speech energy. In Experiment II, the criterion was systematically varied such that the binary mask was based on 75 to 95% of the speech energy in steps of 5%, as well as a 99% condition comparable to Experiment I. All other aspects of processing were the same as in Experiment I. An illustration of the effect of changing the criterion on the binary mask is shown in Figure 3, which shows the ideal binary mask for a criterion set at a low level to detect 99% of the speech energy (Fig. 3A), as well as binary masks based on 85% (Fig. 3C) and 75% (Fig. 3D) of the speech energy. All of the binary masks were generated as described in the signal-processing section above, with the only difference being the criterion used. The different criteria are illustrated in Figure 3B, in which the energy of a single frequency band is shown, similar to Figure 1B. When the criterion was set to detect a smaller percentage of the speech energy, the binary mask did not include the low-level speech energy and detection performance was thus decreased. The criterion was determined for each sentence individually; this was done to ensure that the sentences used in each track were processed using binary masks based on an identical percentage of speech energy exceeding the criterion.

The binary mask was then applied to the stimuli as described in Experiment I. Processing was only performed in the LF condition in Experiment II due to the large number of nonideal masks that were tested and the limited number of stimuli available. The LF condition was chosen because of the complicated nature of speech and the performance of a possible real-world detector. The low-frequency components of speech (i.e., the harmonics) have a more defined structure that allows for easier detection as compared with noise-like components at higher fre-

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

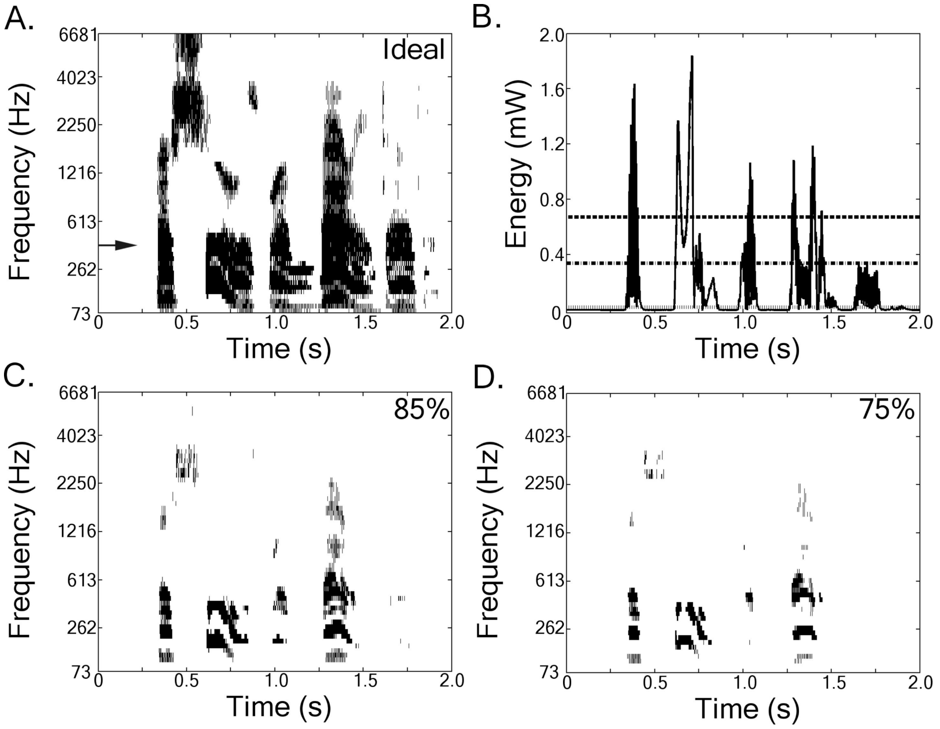

Fig. 3. Degradation of the ideal binary mask. (*A*), Ideal binary mask for the example sentence. B, Energy versus time for the 414-Hz frequency band (*arrow* in *A*) with lines showing the criteria used to detect 99% (*dotted line*), 85% (*dot-dashed line*), and 75% (*dashed line*) of the speech energy. (*C*), Binary mask based on 85% of speech energy. (*D*), Binary mask based on 75% of speech energy.

quencies. The frequency- and time-resolution of the binary mask were the same as in Experiment I.

Experiment III: Binary Mask Frequency Resolution and Temporal Smearing

The methods of Experiment III were similar to those of the LF condition of Experiment I. The primary difference was that in Experiment III the frequency-resolution of the binary mask was lower and a temporal smearing was applied to the ideal binary mask. For the control condition, stimuli were processed as in Experiment I under the UNP condition. Two experimental conditions were tested, one with a reduced frequency resolution (1 filter per ERB), and another with temporally smeared binary masks. The binary masks for each sentence were temporally smeared by varying the way in which the gains were applied to the output of the Gammatone filter bank. In Experiments I and II, the gains were applied on a sample-by-sample basis, for Experiment III whenever the energy-threshold was exceeded for a given band the gain was kept at 1.0 for a set amount of time. The amount of time was set to either 15 msec or 100 msec. This temporal smearing had the effect of removing some fast variations in the temporal patterns of the gain changes; that is, although the gains could rapidly change to 1.0, they returned to the 0.2 state slowly.

Listeners

A total of six listeners with normal hearing and 14 listeners with sensorineural hearing loss participated in this study. Listeners each participated in a single experiment, with no overlap in participants across the three experiments except as noted. All listeners with hearing loss had a mild-to-moderate, sloping sensorineural hearing loss that was bilateral and symmetric (Fig. 4). Listeners with normal hearing had thresholds lower than 15 dB HL from 250 to 6000 Hz (ANSI, 1989). Listeners for Experiment I consisted of 8 subjects, 3 listeners with normal hearing and 5 listeners with sensorineural hearing loss. Listeners for Experiment II consisted of 10 subjects, 3 listeners with normal hearing and 7 listeners with sensorineural

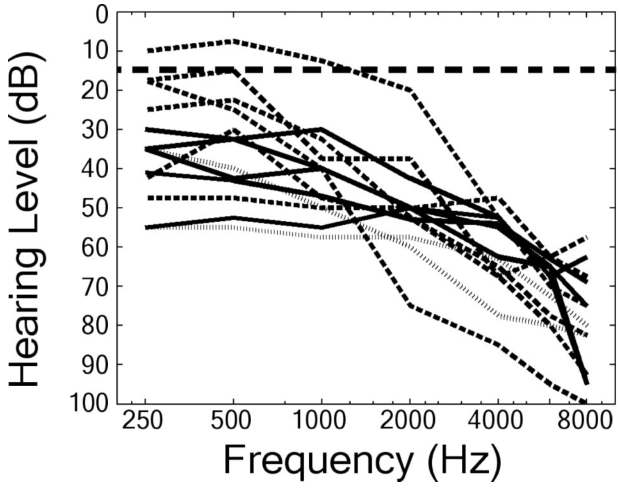

Fig. 4. Audiograms for the listeners with hearing loss in Experiments I, II, and III are shown with *solid, dashed,* and *dotted* lines, respectively. All thresholds of listeners with normal hearing were less than 15 dB HL (*heavy dashed line*).

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

hearing loss. Two of the listeners with hearing loss were used for Experiment III. Listeners with normal hearing were aged 21 to 27 (mean, 23.8); listeners with hearing loss were ages 68 to 88 (mean, 78.5).

Individual audiograms for the listeners with hearing loss for all experiments are shown in Figure 4. There was no greater than 10 dB difference in threshold between ears at all frequencies tested. All but two of the listeners with hearing loss were experienced hearing aid wearers. All testing was performed unaided with no spectral shaping. All listeners were paid for their participation.

Procedure

Listeners were seated in a double-walled, sound-attenuating booth (IAC). Stimuli were presented in the free-field by a speaker located 1 m in front of the listener. Stimuli were presented through a TDT System II 16-bit D/A system and digital attenuator (TDT PA4), and amplified by a Crown D-75A power amplifier. The level of the stimuli was adjusted for each subject to ensure that the stimuli were audible, yet did not exceed the listener's discomfort level. Presentation levels varied from 77 dB (A) to 87 dB (A).

The standard HINT procedure (Nilsson et al., 1994) was used to measure listeners' SRT (50% intelligibility of sentences) of speech in speech-spectrum shaped noise. Briefly, the subjects were presented with a HINT sentence in speech-spectrum shaped noise, and asked to repeat back the sentence to the experimenter. During each track, the SNR had a lower bound of −10 dB. This lower bound was needed because preliminary testing for Experiment I showed that listeners with normal hearing could understand the speech at SNRs down to −30 dB (the lowest SNR used in preliminary testing). At these low SNRs, application of the binary mask resulted in the noise present in each frequency band having an envelope similar to the speech within that band because of the gain changes introduced. Listeners are able to understand speech when noise is separated into frequency bands and amplitude-modulated with the speech envelope, even when there is no speech actually present (Shannon et al., 1995). Because of the ability to understand the speech from the imposed envelopes, even when the speech itself was inaudible, a lower bound on the SNR was necessary.

Listeners were presented with 40 unprocessed HINT sentences to familiarize themselves with the testing procedure and to ensure that they could perform the task. After the initial familiarization, subjects were presented with the processed stimuli. Listeners' HINT thresholds were determined with the standard procedure of taking the average of the SNRs of the final 16 sentences, as well as the SNR at which the 21st sentence would have been presented (Nilsson et al., 1994). For each condition, two HINT thresholds were obtained and the average of the two is shown in the results. Listeners were never presented with the same sentence more than once. The order of presentation for the experimental conditions was randomized for each subject. Listeners were allowed to take short breaks between tracks as necessary.

Results

Experiment I

Results on the HINT test for all four listening conditions (UNP, LF, HF, AF) are shown in Figures 5 and 6 for listeners with hearing loss and normal hearing, respectively. Listeners' pure-tone averages (PTAs) for 500, 1000, and 2000 Hz are included in Figure 5. For the unprocessed condition, the difference in HINT thresholds between the average listener with hearing loss and the average listener with normal hearing was 3.6 dB, which is consistent with previous work for long-term speech-spectrum shaped noise (Plomp, 1994). In general, application of the ideal binary mask decreased the HINT thresholds for all listeners, which indicates an improvement in the listeners' abilities to understand speech in noise. All of the listeners with hearing loss showed the greatest reduction in SNR for the AF condition, followed closely by the LF condition. The amount of reduction for the HF condition was smaller than for the LF or AF conditions and varied more across listeners. Results from a simple linear regression indicated that there was no significant relation between the HINT thresholds and PTAs ($p > 0.15$).

Listeners with normal hearing also showed improved HINT thresholds for all conditions of processing (Fig. 6). For the AF condition, all listeners with normal hearing were operating near the minimum SNR that was used (−10 dB). The arrows in Figure 6 indicate that the listener reached the limit of SNRs used in the processing and their scores would have likely been better if the SNR had not been limited to −10 dB. Unlike the listeners with hearing loss, the listeners with normal hearing showed an improvement in HINT threshold for the HF condition.

In general, the listeners with hearing loss derived more benefit from the LF and AF conditions than did the listeners with normal hearing, whereas the listeners with normal hearing derived more benefit from the HF condition than did the listeners with hearing loss. The larger improvement for the listeners with hearing loss was due to their higher SRTs for unprocessed speech; the improvement of listeners with normal hearing was also limited by the minimum SNR used

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

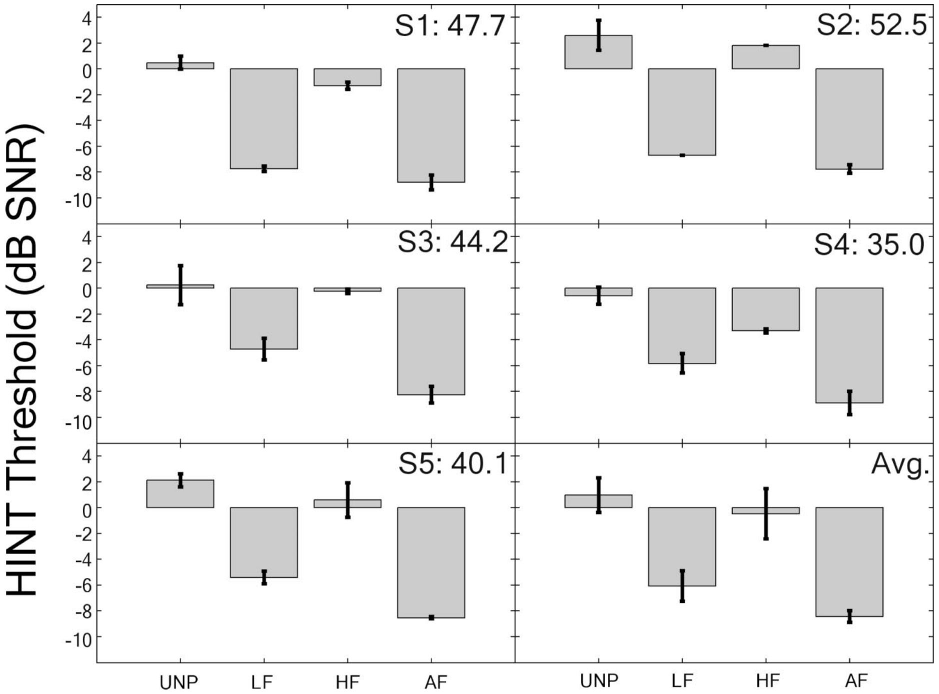

Fig. 5. Individual and average HINT thresholds for each processing condition for listeners with hearing loss. A more negative threshold signifies better performance. Pure-tone average (PTA) thresholds for 500-, 1000-, and 2000-Hz tones are shown in each panel.

(−10 dB). However, the listeners with hearing loss were not affected by the floor imposed on the processing.

Experiment II

The results for Experiment II are summarized in Figures 7 to 9. The SNRs for listeners with hearing loss are plotted as a function of the amount of energy above the threshold of the degraded binary mask. Individual subjects' PTAs are shown in the legend. As the binary mask approached the ideal mask, all subjects with hearing loss showed a decrease in SNR. The SNRs obtained for the ideal mask (99% speech energy above the criterion, the right-most point on the curves) matched the results found in Experiment I, which were obtained with a different set of listeners.

In Figure 8, the results from Figure 7 are replotted to subtract out each listener's unprocessed score. Thus, 0 SNR in Figure 8 indicates that there was no differ-

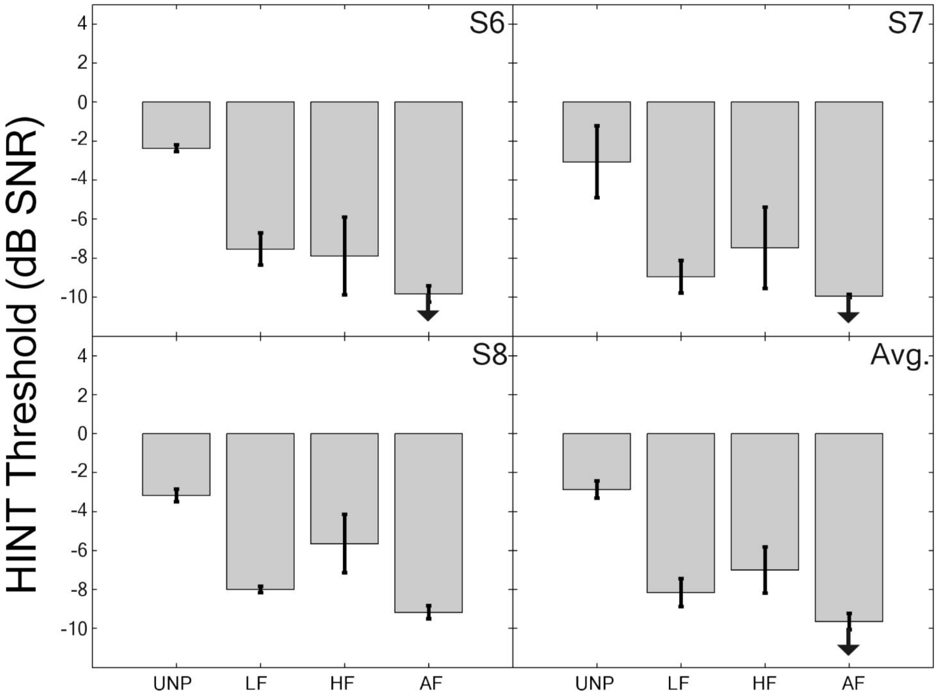

Fig. 6. HINT thresholds for all conditions for listeners with normal hearing. All thresholds for unprocessed speech fell within the norms for the HINT (Nilsson et al., 1994). *Arrows* indicate that the subjects hit the floor of the processed SNRs, for which only the temporal envelope cues were available (see text).

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

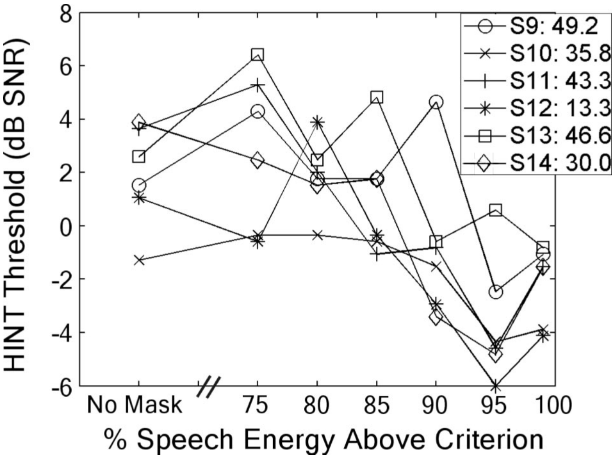

Fig. 7. Effect of degrading the ideal binary mask on listeners with hearing loss. HINT thresholds as a function of the percent of speech energy used to create the binary mask (see Fig. 3). All subjects showed a decreasing trend in HINT threshold as more speech energy was used to create the ideal binary mask. Performance for the ideal mask (the rightmost point of each line) matched the results in Experiment I.

ence in the listeners' perception of processed and unprocessed speech. The change in SNR decreased as the percentage of energy above the criterion increased. The heavy solid line shows the mean change for all subjects with hearing loss, with error bars denoting ±1 SD. Although individual subjects showed improvement in SNR at each percentage, a paired t-test showed no significant difference ($p < 0.05$) until the percentage was equal to or greater than 95% (indicated by the asterisks in Fig. 8). If the outlier from subject 10 was removed (the open circle of Fig. 8), the difference at 90% became significant ($p = 0.02$).

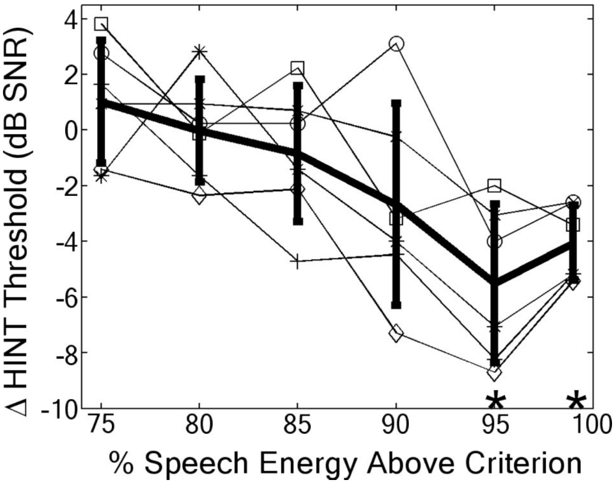

Fig. 8. Changes in HINT threshold for listeners with hearing loss as a function of the percent of speech energy used to create the binary mask. *Solid line* represents mean (±1 SD) change for all listeners with hearing loss. Listeners with hearing loss improved when the binary mask was based on greater than 90% of the speech energy. For binary masks based on less energy, there was either no change from the unprocessed condition, or a slight increase in HINT threshold. Asterisks represent statistically significant differences from the unprocessed condition (*p* < 0.05).

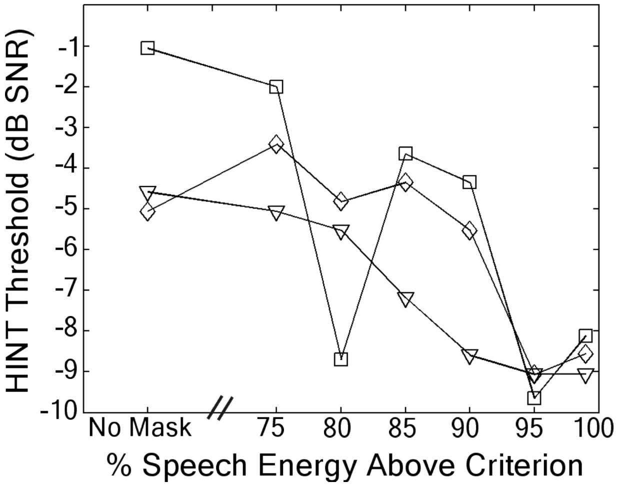

Fig. 9. HINT thresholds for listeners with normal hearing improved as the binary mask approached the ideal binary mask. For binary masks based on higher percentages of speech energy, HINT tracks for listeners with normal hearing reached the minimum SNR used in the study (−10 dB).

Similar to the listeners with hearing loss, HINT thresholds for listeners with normal hearing improved as the binary mask became closer to the ideal binary mask (Fig. 9). Listeners with normal hearing had better HINT thresholds for the entire range of percentages, as expected. The scores are replotted as a change in SNR, along with an average change, in Figure 10. A paired t-test showed that HINT thresholds for listeners with normal hearing for processed speech were significantly better than their unprocessed thresholds ($p < 0.05$) for the ideal binary-mask condition (99% of speech energy above the criterion). Note that only three listeners with normal hearing were tested.

Experiment III

The result of reducing the frequency-resolution and temporally smearing the ideal binary mask is shown in Figure 11. When the frequency resolution was decreased to one filter per ERB, both subjects still showed an improvement in HINT thresholds as compared with the control condition. The size of this improvement matched that found in Experiment I.

The SNR obtained for both listeners for the temporally smeared binary masks are also shown in Figure 11. The two conditions represent gains that were forced to slowly return to their attenuating state, thus removing many of the fast variations in the gain changes as compared with the control

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

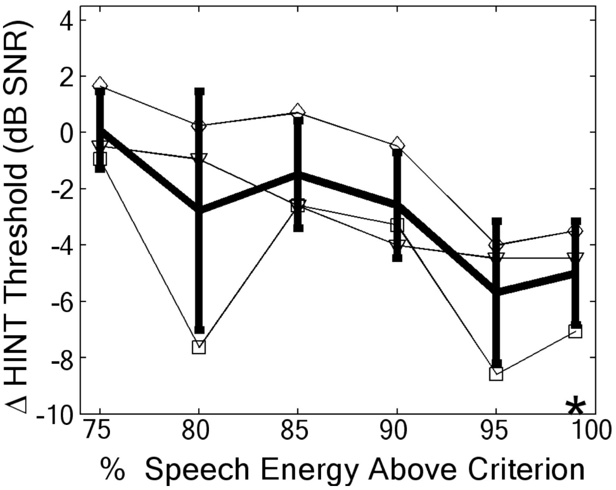

Fig. 10. Change in HINT threshold for listeners with normal hearing as a function of the amount of speech energy used to create the binary mask. *Solid line* represents mean (±1 SD) change for all listeners with normal hearing. Only the 99% condition was statistically different from the unprocessed condition (paired *t*-test, $p < 0.05$, asterisk).

condition, for which the gains changed on a sample-by-sample basis. The temporally smeared masks had a frequency resolution of two filters per ERB. Here, the two listeners showed a similar pattern for the 100-msec condition, with little or no change from the unprocessed condition. For the 15-msec condition, the first listener showed an improvement similar to that obtained in the one-filter-per-ERB condition with no temporal smearing. The change seen for the second listener was difficult to determine, as the subject's variability was high; for the first track, the SNR improved to a level similar to the one-filter-per-ERB condition, whereas the second track resulted in a SNR higher than that of the control.

Discussion

The results of Experiment I demonstrated that a time-frequency gain manipulation strategy could improve the SRT of speech in speech-spectrum noise, provided that the gains were manipulated based on the ideal detector. This result is in contrast to some past studies that have shown no improvement using this technique (Fabry & Van Tasell, 1990; Klein, 1989). Three possible differences between the current study and past studies (Dempsy, 1987; Fabry & Van Tassell, 1990; Klein, 1989; Ono et al., 1983; Rankovic et al., 1992; van Dijkhuizen, Anema & Plomp, 1987; van Dijkhuizen et al., 1989; van Dijkhuizen et al., 1990; van Dijkhuizen, Festen & Plomp, 1991) may account for the different results: (1) reliance on nonideal detectors, (2) the use of fewer frequency bands for gain manipulations, and (3) sluggish manipulation of the gains. The combination of these differences may have led to the large improvements shown in the current study with respect to previous studies. Replicating the current study with the frequency bands used in previous studies (or in current hearing aids) would enable a more direct comparison between studies.

One interesting result from Experiment I indicated that the performance of listeners with normal hearing reached the limits of processing (−10 dB) on the ideal binary mask condition, but listeners with hearing loss did not reach this limit, even when the ideal binary mask was applied to the full frequency range of speech. At the lower levels of SNR, below

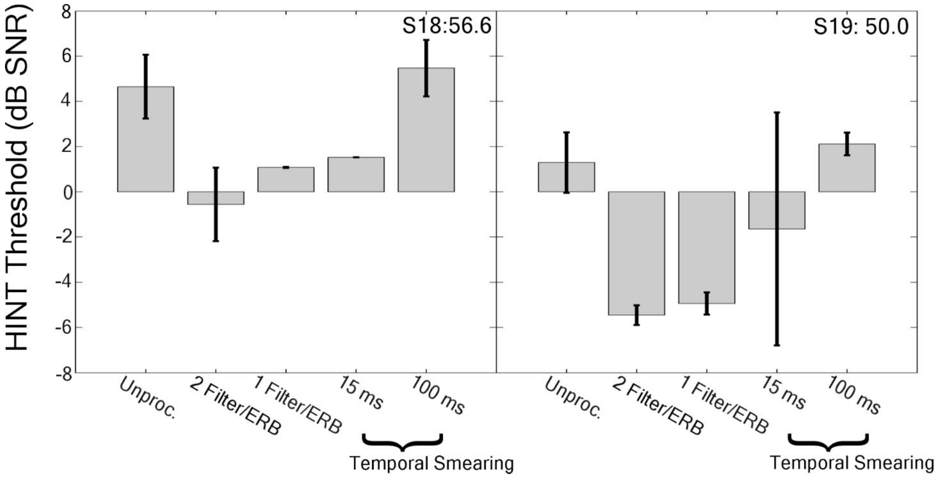

Fig. 11. HINT thresholds for two subjects with hearing loss for changes in the frequency-resolution and temporal smearing of the binary mask. Both subjects had improved thresholds for both frequency resolutions tested. When the binary mask was temporally smeared with a 15-msec rectangular window, one subject showed an improvement similar to that seen in Experiment I, whereas the other showed a variable response. When the binary mask was smeared with a 100-msec rectangular window, both subjects' HINT thresholds matched those for unprocessed speech.

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

about −6 dB, the output was dominated by the noise, which had been given a speech-like envelope by the manipulation of the gains. The inability of the listeners with hearing loss to use this information suggests an impairment in their ability to process envelope cues compared with listeners with normal hearing. Studies have shown that listeners with hearing loss generally perform similarly to listeners with normal hearing in amplitude-modulation detection as long as audibility is taken into account (Bacon & Gleitman, 1992; Moore, Shailer & Schooneveldt, 1992; Moore & Glasberg, 2001; Sek & Moore, 2006). However, it is possible that the wider tuning of the impaired auditory system could have a significant effect on the complex wide-band envelope cues that are contained in speech. Additionally, listeners with hearing loss did not show an improvement when the ideal binary mask was applied to the high-frequency (>1500 Hz) region, unlike the listeners with normal hearing. The improved SNR of higher-frequency components should have a large effect on the subjects' scores. The lack of improvement for a few of the listeners with hearing loss may have resulted because the processed speech was inaudible at some higher frequencies due to the listener's hearing loss (Fig. 4); however, the fact that none of the listeners with hearing loss showed improvement (even those with milder losses) suggests an alternative reason. A more likely reason is the susceptibility of those with hearing loss to upward spread of masking. The fact that all listeners with hearing loss showed additional improvement when the ideal binary mask covered the entire frequency range suggests that the masking effect of the low-frequencies in the HF condition explained the lack of improvement in the HF condition. If the processed speech was unavailable because of the listeners' increased thresholds, the AF and LF conditions should have shown similar improvements. Last, the results could have been influenced by the difference in the age of the two groups of listeners, but it is unlikely because hearing loss has been shown to have a greater effect than age on listeners' recognition of speech in noise (Gordon–Salant & Fitzgibbons, 1995; Halling & Humes, 2000; Humes, 2002).

By degrading the ideal binary mask in Experiment II, it was found that more than 90% of the energy must be detected to improve a listener's SRT. To detect such a high percentage of the speech energy in the presence of noise, however, requires a very good detector that may not be realizable in the computationally limited environment of a hearing aid. Also, everyday listening environments are never "clean" or without some level of background noise in which the speech must be detected. Thus, detection of overall speech energy might not be the best cue. Additional experiments should be done to examine the effect of detecting various other cues in speech, such as onsets and offsets, the transitions of formants, and envelope cues. It is possible that a detector would not have to detect the high percentage of energy shown in Experiment II if it were to detect some of the more salient speech cues mentioned above.

The results of Experiment III, although based on a small number of subjects, help to further define the binary mask needed to improve SRT. The frequency resolution results indicate that a lower frequency resolution can still improve SRT. When the frequency resolution of the NR algorithm was matched to that of the listener (the one filter per ERB condition), the application of the NR algorithm did not improve the instantaneous SNR within each ERB because the noise was unaffected when speech was present in the frequency band. This result suggests that the improvement in the SRT was the result of other factors and not due to an increase in the instantaneous SNR.

The temporal smearing results of Experiment III suggest that any time-frequency gain manipulation must be relatively fast (15 msec or less) for there to be an improvement in the SRT. This result may explain why many of the current time-frequency gain manipulation algorithms used in hearing aids provide little in the way of improved intelligibility. Most current hearing aid algorithms are typically slow-acting (20 msec to a few seconds).

Results from the three experiments reported in this study have practical application for the development of time-frequency gain-manipulation algorithms for hearing aids. Experiment I demonstrated that a time-frequency gain manipulation NR algorithm works and can improve listeners' SRT by processing only the low-frequency components (<1.5 kHz) of speech. This finding allows the use of simpler detectors, because the low-frequency, narrowband components of speech are generally easier to detect than the noise-like, high frequency components of speech. The results of Experiment II provide a set of guidelines for the performance of detectors to be used in a real-world algorithm. To improve SRT, the real-world detector must be able to detect at least 90 to 95% of the speech energy. Last, Experiment III defined the necessary frequency-resolution of the overall system, as well as the temporal speed of the gain changes.

Unfortunately, time-frequency gain manipulations often result in poorer sound quality. This degradation in quality was also observed in the current study; listeners with normal hearing often describing the stimuli as "machine-sounding" or as sounding like artificially generated speech. This "machine-like" quality is the result of the rapid transitions of the gains, as well as the low SNRs of the stimuli. For the low SNRs

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

(−10 to −6 dB) presented to the listeners with normal hearing, the output was dominated by the noise, which was roughly shaped by the gain manipulations to mimic speech; the resulting outputs were similar in quality to cochlear implant demonstrations, in which noise is modulated with a speech envelope (Shannon et al., 1995). Interestingly, the listeners with hearing loss were generally less sensitive to the reduced quality of the stimuli. Their informal comments about the quality of the processed speech were unlike the comments of listeners with normal hearing. This may have been because the SNR was at a level where the actual speech still dominated the output. The degraded quality of speech should not be an issue with a real-world implementation of this class of algorithms, as the detectors would not reach the performance levels (i.e., ideal detection at −10 dB SNR) that were used in the current study. The signal in the SNRs that real-world detectors would typically process would dominate over the speech-envelope imposed on the noise. Informal listening suggests that at these SNRs, the quality of the processed speech is at an acceptable level.

The use of binary masks allows for the evaluation of important cues in understanding speech in noise, and provides a platform for the evaluation of NR algorithms. The ideal binary mask provides an absolute limit of benefit that can be achieved by manipulating the gains of frequency bands. By systematically manipulating the binary mask, one can determine the necessary cues or regions of the stimulus that are needed for intelligibility. Once these cues are determined, quantitative comparisons can be made to the detection patterns that are produced by experimental detectors, allowing for the design and testing of such detectors without having to perform extensive intelligibility testing. By simply comparing the detection patterns of the experimental detector to the binary masks that produce increases in intelligibility, one can predict how that detector would perform. Care must be taken, however, to ensure that the binary masks that produce benefit are consistent across listeners with hearing loss. In addition, sound quality must be assessed for any time-frequency gain-manipulation strategy.

Supported by grant R21-DC006057 from the National Institute on Deafness and Other Communication Disorders.

Address correspondence to: Dr. Laurel H. Carney, Departments of Biomedical and Chemical Engineering and Electrical Engineering and Computer Science, Institute for Sensory Research, Syracuse University, 621 Skytop Road, Syracuse, NY 13244. E-mail: lacarney@syr.edu.

Received September 3, 2005; accepted April 28, 2006.

REFERENCES

ANSI S3.5 (1969). American National Standard Methods for the Calculation of the Articulation Index. ANSI S3.5. New York: ANSI.

ANSI S3.6 (1989). American National Standard Specification for Audiometers. ANSI S3.6-1989. New York: ANSI.

ANSI S3.5 (1997). Methods for Calculation of the Speech Intelligibility Index. New York: ANSI.

Bacon, S. P., & Gleitman, R. M. (1992). Modulation detection in subjects with relatively flat hearing losses, *Journal of Speech and Hearing Research*, *35*, 642–653.

Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, *4*, 208–211.

Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 27*, 113–120.

Carhart, R. C., & Tillman, T. W. (1970). Interaction of competing speech signals with hearing losses. *Archives of Ototlaryngology*, *91*, 273–279.

Chabries, D. M., & Bray, V. (2002). Use of DSP techniques to enhance the performance of hearing aids in noise. In: G. M. Davis, Editor. *Noise Reduction in Speech, Applications*. pp. 379–392). New York: CRC Press.

Cooke, M., Green, P., Ljubomir, J., & Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data, *Speech Communication*, *34*, 267–285.

Dempsey, J. J. (1987). Effect of automatic signal-processing amplification on speech recognition in noise for persons with sensorineural hearing loss, *Annals of Otology, Rhinology, and Laryngology*, *96*, 251–253.

Eisenberg, L. S., Dirks D. D. & Bell, T. S. (1995). Speech recognition in amplitude-modulated noise of listeners with normal and listeners with impaired hearing. *Journal of Speech and Hearing Research*, *38*, 222–233.

Fabry, D. A., & Van Tassel, D. J. (1990). Evaluation of an articulation-index based model for predicting the effects of adaptive frequency response hearing aids. *Journal of Speech and Hearing Research*, *33*, 676–689.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, *19*, 90–119.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*, 103–138.

Gordon-Salant, S., & Fitzgibbons, P. J. (1995). Comparing recognition of distorted speech using and equivalent signal-to-noise ratio index, *Journal of Speech and Hearing Research*, *38*, 706–713.

Graupe, D., Grosspietsch, J. K., Basseas, S. P. (1987). A single-microphone-based self-adaptive filter of noise from speech and its performance evaluation. *Journal of Rehabilitation Research and Development*, *24*, 119–126.

Halling, D., Humes, L. E. (2000). Factors affecting the recognition of reverberant speech by elderly listeners. *Journal of Speech and Hearing Research*, *43*, 414–431.

Hawkins, D. B., Yacullo, W. S. (1984). Signal-to-noise ratio advantage of binaural hearing aids and directional microphones under different levels of reverberation, *Journal of Speech and Hearing Disorders*, *49*, 278–286.

Hohmann, V. (2002). Frequency analysis and synthesis using a Gammatone filterbank. *Acustica Acta United with Acustica*, *88*, 334–347.

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

Humes, L. E. (2002). Factors underlying the speech-recognition performance of elderly hearing-aid wearers. *Journal of the Acoustical Society of America*, *112*, 1112–1132.

Kates, J. M. (1994). Speech enhancement based on a sinusoidal model. *Journal of Speech and Hearing Research*, *37*, 449–464.

Klein, A. J. (1989). Assessing speech recognition in noise for listeners with a signal processor hearing aid *Ear and Hearing*, *10*, 50–57.

Klein, A. J., Mills, J. H., & Adkins, W. Y. (1990). Upward spread of masking, hearing loss, and speech recognition in young and elderly listeners. *Journal of the Acoustical Society of America*, *87*, 1266–1271.

Levitt, H. (2001). Noise reduction in hearing aids: a review. *Journal of Rehabilitation Research and Development*, *38*, 111–121.

McAulay, R. J., Quatieri, T. F. (1986). Speech transformations based on a sinusoidal representation, *IEEE Transactions of Acoust Speech Sig Proc*, *34*, 1449–1464.

Moore, B. C. J. (2003). Speech processing for the hearing-impaired: successes, failures, and implications for the speech mechanisms, *Speech and Communication*, *41*, 81–91.

Moore, B. C. J., & Glasberg, B. R. (2001). Temporal modulation transfer functions obtained using sinusoidal carriers with normally hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, *110*, 1067–1073.

Moore, B. C. J., Shailer, M. J., & Schooneveldt, G. P. (1992). Temporal modulation transfer functions for band-limited noise in subjects with cochlear hearing loss. *British Journal of Audiology*, *26*, 229–237.

Mueller, H. G., & Killion, M. C. (1990). An easy method for calculation the articulation index. *Hearing Journal*, *9*, 14–17.

Nilsson, M, Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, *95*, 1085–1099.

Ono, H, Kanzaki, J., & Mizoi, K. (1983). Clinical results of hearing aid with noise-level-controlled selective amplification, *Audiology*, *22*, 494–515.

Pavlovic, C. V. (1988). Articulation index predictions of speech intelligibility in hearing aid selection, *ASHA*, *8*, 63–65.

Plomp, R. (1994). Noise, amplification, and compression: Considerations of three main issues in hearing aid design, *Ear and Hearing*, *15*, 2–12.

Rankovic, C. M., Freyman, R. L., & Zurek, P. M. (1992). Potential benefits of adaptive frequency-gain characteristics for speech reception in noise. *Journal of the Acoustical Society of America*, *91*, 354–362.

Ricketts, T. (2000). Impact of noise source configuration on directional hearing aid benefit and performance, *Ear and Hearing*, *21*, 194–205.

Ricketts, T. A., & Hornsby, B. W. (2003). Distance and reverberation effects on directional benefit, *Ear and Hearing*, *24*, 472–484.

Roman, N., Wang, D., & Brown, G. J. (2003). Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, *114*, 2236–2252.

Schum, D. J. (2003). Noise-reduction circuitry in hearing aids, II: goals and current strategies. *Hearing Journal*, *56*, 32–41.

Sek, A., & Moore, B. C. J. (2006). Perception of amplitude modulation by hearing-impaired listeners: The audibility of component modulation and detection of phase change in three-component modulators. *Journal of the Acoustical Society of America*, *119*, 507–514.

Shannon, R. V., Zeng, F. G., Kamath, V, Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues, *Science*, *270*, 303–304.

Srinivasan, S., Roman, N., & Wang, D. L. (2004). On binary and ratio time-frequency masks for robust speech recognition. *Proc ICSLP* pp. 2541–2544.

Stein, L. K., & Dempsey-Hart, D. (1984). Listener-assessed intelligibility of a hearing aid self-adaptive noise filter, *Ear and Hearing*, *5*, 199–204.

Takahashi, G. A., Bacon, S. P. (1992). Modulation detection, modulation masking, and speech understanding in noise in the elderly. *Journal of Speech and Hearing Research*, *35*, 1410–1421.

Trees, D. E., & Turner, C. W. (1986). Spread of masking in normal subjects and in subjects with high-frequency hearing loss, *Audiology*, *25*, 70–83.

van Dijkhuizen, J. N., Anema, P. C., & Plomp, R. (1987). The effect of varying the slope of the amplitude-frequency response on the masked speech-reception threshold of sentences. *Journal of the Acoustical Society of America*, *81*, 465–469.

van Dijkhuizen, J. N., Festen, J. M., & Plomp, R. (1989). The effect of varying the amplitude-frequency response on the masked speech-reception threshold of sentences for hearing-impaired listeners. *Journal of the Acoustical Society of America*, *86*, 621–628.

van Dijkhuizen, J. N., Festen, J. M., & Plomp, R. (1990). Speech-reception threshold in noise for hearing-impaired listeners in conditions with varying amplitude-frequency response. *Acta Oto-laryngologica Supplementum*, *469*, 202–206.

van Dijkhuizen, J. N., Festen, J. M., & Plomp, R. (1991). The effect of frequency-selective attenuation on the speech-reception threshold of sentences in conditions of low-frequency noise. *Journal of the Acoustical Society of America*, *90*, 885–894.

Weiner, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series.* New York: Wiley.

Yilmaz, O., Rickard S. (2004). Blind separation of speech mixtures via time-frequency masking, *IEEE Transaction on Sig Process*, *52*, 1830–1847.

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.