Speech-cue transmission by an algorithm to increase consonant recognition in noise for hearing-impaired listeners

Eric W. Healy[a),b)] and Sarah E. Yoho[b)]
Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210

Yuxuan Wang
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210

Frédéric Apoux
Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210

DeLiang Wang[b)]
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210

(Received 12 June 2014; revised 19 September 2014; accepted 3 November 2014)

Consonant recognition was assessed following extraction of speech from noise using a more efficient version of the speech-segregation algorithm described in Healy, Yoho, Wang, and Wang [(2013) J. Acoust. Soc. Am. **134**, 3029–3038]. Substantial increases in recognition were observed following algorithm processing, which were significantly larger for hearing-impaired (HI) than for normal-hearing (NH) listeners in both speech-shaped noise and babble backgrounds. As observed previously for sentence recognition, older HI listeners having access to the algorithm performed as well or better than young NH listeners in conditions of identical noise. It was also found that the binary masks estimated by the algorithm transmitted speech features to listeners in a fashion highly similar to that of the ideal binary mask (IBM), suggesting that the algorithm is estimating the IBM with substantial accuracy. Further, the speech features associated with voicing, manner of articulation, and place of articulation were all transmitted with relative uniformity and at relatively high levels, indicating that the algorithm and the IBM transmit speech cues without obvious deficiency. Because the current implementation of the algorithm is much more efficient, it should be more amenable to real-time implementation in devices such as hearing aids and cochlear implants.
 [http://dx.doi.org/10.1121/1.4901712]

PACS number(s): 43.71.Ky, 43.71.Es, 43.66.Ts, 43.72.Dv [FJG] Pages: 3325–3336

I. INTRODUCTION

It is well known that a primary limitation for hearing-impaired (HI) listeners involves poor speech intelligibility in background noise (see Moore, 2007). Whereas modern hearing aids generally allow users to perform quite well in quiet, performance and satisfaction in noise remain low, despite currently implemented noise-suppression methods. An ultimate goal for remedying this limitation may involve a monaural algorithm to improve intelligibility by extracting speech from background noise. Because the algorithm, not the listener, performs the extraction of intelligible speech, the limitations of the HI listener would be substantially minimized. Numerous such "speech enhancement" or "segregation" algorithms have been proposed, but substantial increases in intelligibility have remained elusive despite decades of effort (see Dillon, 2012, p. 232).

Healy *et al.* (2013) reported substantial gains in intelligibility by HI listeners using a monaural binary-masking algorithm. Listeners having sensorineural hearing loss were tested using the Hearing In Noise Test (HINT) sentences mixed with speech-shaped noise (SSN) and with babble. Intelligibility was generally greater than 80% following algorithm processing, despite scores in unprocessed speech in noise that ranged considerably across listeners and were often lower than 20%. The benefit from algorithm processing was larger for HI listeners than for normal-hearing (NH) control subjects, due primarily to the lower unprocessed speech-in-noise scores of the former group. Benefit was also larger for less-favorable signal-to-noise ratios (SNRs), the nonstationary background, and listeners with the most elevated audiometric thresholds. It was concluded that the algorithm operates to restore speech perception in noise for listeners who need it most (those with the greatest losses) and under conditions that they find most troublesome (modulated backgrounds and unfavorable SNRs).

In binary masking, the extraction of speech from noise is viewed as a classification task. It operates by classifying each time-frequency (T-F) unit of a speech-plus-noise mixture according to the SNR within the unit (Hu and Wang, 2001; Wang, 2005). Units having an SNR greater than a predefined local criterion (LC) are retained, whereas units having an SNR less than this value are simply discarded. The binary-masked speech signal corresponds to the mixture within retained T-F units only, and represents speech extracted from noise. In the Ideal Binary Mask (IBM), the premixed speech and noise signals are known, and the mask corresponds to

[a)]Author to whom correspondence should be addressed. Electronic mail: healy.66@osu.edu
[b)]Also at Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210

ideal classification. IBM processing produces remarkable speech-intelligibility improvements in noise for both HI and NH listeners, even at extremely low SNRs (Anzalone *et al.*, 2006; Brungart *et al.*, 2006; Li and Loizou, 2008; Wang *et al.*, 2008, 2009; Cao *et al.*, 2011; Sinex, 2013).

In contrast, the algorithm employed by Healy *et al.* (2013) *estimates* the IBM directly from the speech-noise mixture. This ability to estimate based only on the mixture is obviously critical for eventual applicability of such an algorithm. The estimation was done by training deep neural networks (DNNs) to classify T-F units as either dominated by the target speech (SNR $>$ LC) or dominated by the background noise (SNR $\leq$ LC). The DNNs received as input acoustic features extracted from sentences mixed with noise, and the sentences used for training were not used for testing human subjects.

Key to the success of the algorithm is its ability to accurately estimate the IBM. This accuracy cannot be simply inferred from high sentence intelligibility, because it is known that sentences contain a multitude of redundant cues, and that they can retain high intelligibility when many of these cues are distorted or missing (e.g., Shannon *et al.*, 1995; Warren *et al.*, 1995; Stickney and Assmann, 2001). One technique that has been used to assess classification accuracy of an estimated mask involves an acoustic analysis of hit minus false-alarm rate (HIT−FA), where HIT denotes the percentage of speech-dominant T-F units correctly classified, and FA denotes the percentage of noise-dominant units incorrectly classified (Kim *et al.*, 2009). This analysis therefore involves the acoustic similarity between the estimated and ideal masks. An analysis of the estimated IBMs in Healy *et al.* (2013) indicated a HIT−FA rate of approximately 80%, which was substantially greater than that produced by the classification algorithm of Kim *et al.* (2009) based on Gaussian mixture models.

In the current study, we employ an alternative technique to assess the accuracy with which the binary mask is estimated. Here, we assess mask accuracy and speech-segregation effectiveness using a perceptual analysis of speech cues transmitted to listeners. The rationale is as follows: If a mask estimated by the algorithm captures and delivers speech cues in a fashion similar to the IBM, then the pattern of speech information transmitted to listeners should be similar. Thus, the current measure is one of "effective" or perceptual accuracy. This analysis has the additional benefit of revealing the extent to which various speech features are preserved during binary masking. Although it is known that ideal binary masking substantially improves the intelligibility of speech in noise, little is known about the specific cues transmitted to HI and to NH listeners following such binary masking.

In the current study, consonant recognition in noise by HI and NH listeners was assessed using a more efficient version of the algorithm described by Healy *et al.* (2013; see also Wang and Wang, 2013). Recognition was assessed for speech extracted from noise using the algorithm and the IBM. An information-transmission analysis (Miller and Nicely, 1955; Wang and Bilger, 1973) was then employed to determine the features of speech transmitted to listeners. The goals of the current study are (1) to determine, for both HI and NH listeners, the ability of the more efficient algorithm and the IBM to improve recognition of speech that lacks the substantial redundancy of cues that characterizes sentences and therefore requires additional accuracy of bottom-up acoustic cues; (2) to determine the accuracy of the estimated binary mask relative to the IBM in terms of similarity in speech features transmitted to listeners; and (3) to determine the features of speech that are conveyed to HI and NH listeners by the estimated and ideal binary masks.

II. METHOD

A. Subjects

A group of ten listeners having bilateral sensorineural hearing loss of cochlear origin participated. They were bilateral hearing-aid wearers recruited from The Ohio State University Speech-Language-Hearing Clinic to represent typical patients. These listeners ranged in age from 25 to 73 yr (mean = 58.4) and seven were female. Prior diagnoses were confirmed on day of test using otoscopy, tympanometry [American National Standards Institute (ANSI), 1987], and pure-tone audiometry (ANSI, 2004, 2010). The hearing losses ranged from mild to severe and were moderate on average. Pure-tone average audiometric thresholds (PTAs, average of thresholds at 500, 1000, and 2000 Hz, averaged across ears) ranged from 35 to 71 dB hearing level (HL) with an average of 48 dB HL. The configurations of hearing loss ranged from flat to sloping. Audiograms obtained on day of test are presented in Fig. 1, along with subject number, age, and gender.

Also recruited were ten younger listeners having NH, as defined by audiometric thresholds of 20 dB HL or below at octave frequencies from 250 to 8000 Hz (ANSI, 2004, 2010). These listeners were recruited from undergraduate courses at The Ohio State University. They ranged in age from 20 to 22 yr (mean = 20.8) and eight were female. All subjects received a monetary incentive or course credit for participating. As in Healy *et al.* (2013), age matching between HI and NH subjects was not performed because the goal was to assess the abilities of typical (older) HI listeners relative to the gold-standard performance of young NH listeners.

B. Stimuli

The stimuli were 16 consonants (/p, t, k, b, d, g, f, v, s, ʃ, θ, ð, z, ʒ, m, n/) in /a/-consonant-/a/ format. The 44.1-kHz, 16-bit digital files were produced using two male and two female talkers for a total of 64 vowel-consonant-vowels (VCVs, recordings from Shannon *et al.*, 1999). The backgrounds included speech-shaped noise (SSN) and multi-talker babble. The SSN was created by shaping a 10-s white noise to match the long-term average amplitude spectrum of all 64 VCV utterances. This shaping was performed using a 1000-order arbitrary-response finite impulse-response filter (fir2 in MATLAB) having frequency-magnitude characteristics derived from the 64 000 point, Hanning-windowed, fast-Fourier transform of the concatenated utterances. Signal-to-

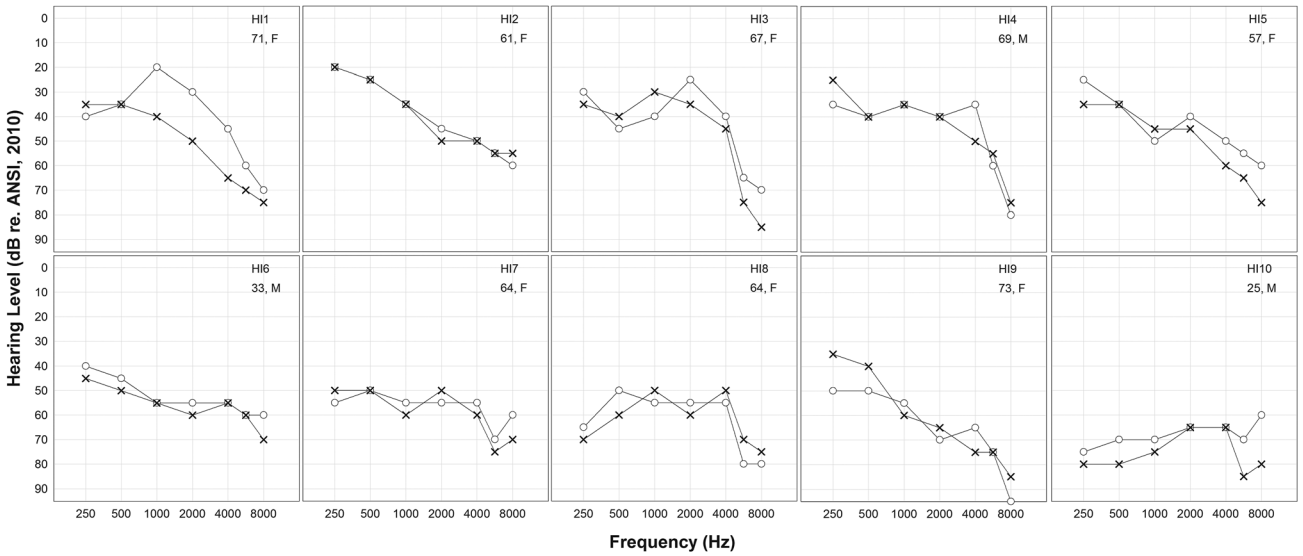

FIG. 1. Pure-tone air-conduction audiometric thresholds for the listeners with sensorineural hearing impairment. Right ears are represented by circles and left ears are represented by ×'s. Also displayed is subject number, listener age in years, and gender. Listeners are numbered and arranged according to increasing pure-tone average audiometric threshold.

noise ratios in SSN were −5 and −8 dB for the HI listeners and −8 and −12 dB for the NH listeners. The multi-talker babble was that employed in Healy *et al.* (2013) and was created by mixing at equal amplitudes sentences from the TIMIT database (Garofolo *et al.*, 1993; four male and four female talkers, two sentences each). Signal-to-noise ratios in babble were −3 and −5 dB for the HI listeners and −5 and −8 dB for the NH listeners. All stimuli were downsampled to 16 kHz prior to processing.

These stimuli were presented under four conditions: Unprocessed speech in noise, algorithm processed, IBM processed, and unprocessed speech in quiet. The first condition consisted of speech mixed with SSN or babble. Each individual utterance was mixed with the background having a random start point within the looped noise or babble sample. The noise began approximately 50 ms before the speech and ended approximately 100 ms after. It was anticipated that increases in masking that can occur as a result of signal placement near the onset of a masker (termed temporal effects of masking or overshoot, see Bacon and Healy, 2000) should be largely absent because the target consonant was preceded by an initial vowel in every utterance and therefore distanced from masker onset. The algorithm- and IBM-processed conditions were based on these same speech-noise mixtures and are described in the next two sections. Because the algorithm is based on IBM estimation, the IBM is described first.

1. IBM Processing

Ideal binary-mask processing followed closely that of Wang *et al.* (2009). Specifically, the signals were divided using a 64-channel gammatone filterbank having center frequencies ranging from 50 to 8000 Hz, equally spaced in equivalent rectangular bandwidths (Glasberg and Moore, 1990). The signals were then divided into 20-ms frames having 10-ms overlap. Based on this cochleagram representation (Wang and Brown, 2006), the IBM was derived by calculating the local SNR (computed from premixed signals) within each T-F unit. If the local SNR was greater than the LC, the mask value was set to 1 and the T-F unit was designated as target dominant. Otherwise, the mask value was set to 0 and the T-F unit was designated as noise dominant. That is,

$$\mathrm{IBM}(t,f) = \begin{cases} 1, & \text{if SNR}\,(t,f) > \mathrm{LC} \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where SNR(t, f) denotes the local SNR within the T-F unit centered at time t and frequency f. The IBM-processed speech was synthesized by gating the speech-noise mixture using the IBM (Wang and Brown, 2006). The selection of LC value is important for speech recognition. To preserve a sufficient amount of speech information, LC was set as follows: [−8, −10, −12, and −16 dB] for input SNRs of [−3, −5, −8, and −12 dB], respectively. These LC choices were made to produce roughly the same difference between LC and input SNR, which is referred to as Relative Criterion (RC). The IBM is determined by the RC, and it has been shown that the same RC values yield comparable intelligibility scores over a wide range of input SNRs (Kjems *et al.*, 2009).

2. Algorithm processing

The algorithm processing followed generally that of Healy *et al.* (2013). In that system, the speech-noise mixture was divided into T-F units using the same 64 frequency channels and 20-ms overlapping frames as described for IBM processing. A set of acoustic features (see below) was extracted from the mixture in each T-F unit and fed to an array of DNNs, one for each frequency channel. These features were used by the DNNs to classify each T-F unit as target dominant or noise dominant, thus estimating the IBM. The DNNs were then trained to minimize the difference between the estimated binary mask and the true IBM, for

each speech utterance, using backpropagation. Once the DNNs were trained using one set of sentences mixed with noise having random start points, they were used to estimate the IBMs for a different set of sentences mixed with noise having random start points, and these latter sentences were used to test listeners.

In the system of Healy *et al.* (2013), a trained DNN for each frequency channel (a subband DNN) output the probability of target dominance for the T-F units in that channel. A second subband DNN was then trained on a window of these posterior probabilities to incorporate spectro-temporal context. In the current study, a similar but more efficient system was employed. Instead of training two DNNs for each of 64 frequency channels (128 in total), only a single DNN was employed to estimate the IBM. This single DNN output a vector of estimated mask values (zeros or ones) for all 64 gammatone frequency channels in each 20-ms time frame.

The same set of complementary features (Wang *et al.*, 2013) employed previously was input to the DNN during system training. This feature set consisted of (1) the amplitude modulation spectrogram (AMS), (2) relative spectral transform and perceptual linear prediction (RASTA-PLP), and (3) the mel-frequency cepstral coefficient (MFCC). The interested reader is directed to Wang *et al.* (2013) for descriptions of these features, which are commonly used in speech processing. In the current implementation, the features were extracted from each frame of the speech-noise mixture (i.e., across all frequencies simultaneously), rather than from each individual T-F unit. Additionally, the features were mean-variance normalized and then post-processed by an autoregressive and moving average (ARMA) filter (Chen and Bilmes, 2007), which has been shown to improve speech-segregation performance in low SNR conditions (Chen *et al.*, 2014). Temporal context was incorporated in the current system using a five-frame spliced window of the combined features as input to the DNN.

The DNN had three hidden layers, each having 1024 rectified linear hidden units. Supervised training of the DNN was performed using the backpropagation algorithm coupled with a dropout regularizer (Hinton *et al.*, 2012). No unsupervised pretraining was used. The adaptive stochastic gradient descent (Duchi *et al.*, 2011), with momentum as the optimizer, was employed to minimize the difference between the estimated IBM and the IBM. The current training set for each SNR condition consisted of 513 utterances not used during subject testing. This corresponded to an average of eight utterances of each VCV by each of the four talkers.

C. Procedure

There were a total of 13 conditions heard by each listener (2 noise types $\times$ 2 SNRs $\times$ 3 processing conditions, plus speech in quiet). Each listener heard all 64 utterances (16 VCVs $\times$ 4 talkers) in random order in each condition. Conditions were blocked and randomized, such that noise type and SNR were heard in random order for each subject, but the three processing conditions within a particular noise type and SNR were heard contiguously in random order. Speech in quiet was presented as the last condition.

Each utterance in each condition was scaled to play back at the same RMS level. The presentation level for NH listeners was 65 dBA. Levels for the HI listeners were set to 65 dBA plus frequency-specific insert gains as determined by the NAL-R hearing-aid fitting formula (Byrne and Dillon, 1986). These individualized gains maximized audibility across frequencies for the HI subjects, who had primarily sloping losses. This was important because the various speech cues examined here differ in their spectral distribution of energy, and it was desired to reduce the influence of differential audibility on the transmission of these cues. Insert gain (IG) at each frequency f was determined by

$$\mathrm{IG}_f = 0.15\mathrm{PTA} + 0.31H_f + k_f, \quad (2)$$

where H_f is the audiometric threshold at frequency f in dB HL, and k_f takes the following values [−17, −8, 1, −1, −2, and −2] at the following frequencies [250, 500, 1000, 2000, 4000, and 6000 Hz]. Gain applied at 6000 Hz was also applied at 8000 Hz, and gain applied at 250 Hz was also applied at 125 Hz (see Table I). A 500-order fir2 filter was employed to perform this subject-specific shaping of stimuli. Hearing-impaired listeners were tested with hearing aids removed, and all levels were calibrated using a sound-level meter and flat-plate headphone coupler (Larson Davis models 824 and AEC 101; Depew, NY).

Signals were delivered using a personal computer running custom software written in MATLAB. Subjects responded by selecting with a computer mouse from a matrix of alternatives displayed on the computer screen and labeled using everyday labels (e.g., "aSHa") and example words (e.g., "Ship"). The signals were transformed to analog form using Echo Digital Audio (Santa Barbara, CA) Gina 3G digital-to-analog converters and presented diotically over Sennheiser HD280 headphones (Wedemark, Germany).

Listeners were tested while seated in a double-walled audiometric booth. Testing began with audiometric evaluation followed by a brief familiarization. During this familiarization, the experimenter presented each consonant while pointing out its location on the response matrix. The subject then heard and responded to all 16 consonants produced by one male and one female talker not used for testing. These

TABLE I. Gains (in dB) for each hearing-impaired subject at each frequency (in Hz) as prescribed by the NAL-R.

	125	250	500	1000	2000	4000	6000	8000
HI1	0	0	8	16	17	20	23	23
HI2	−5	−5	5	17	19	19	20	20
HI3	−2	−2	11	17	14	17	25	25
HI4	−2	−2	10	18	17	17	22	22
HI5	−1	−1	9	22	18	21	23	23
HI6	4	4	15	26	25	23	25	25
HI7	7	7	16	27	23	24	28	28
HI8	12	12	17	26	25	23	30	30
HI9	5	5	14	27	28	28	30	30
HI10	18	18	26	34	30	29	33	33

stimuli were unprocessed and in quiet, and feedback was provided during familiarization but not during testing. This phase was repeated if desired by the subject or experimenter up to a maximum of three times. During familiarization, HI listeners were asked if the signals were adequate in level and comfortable, and "if they would turn it up or down if they could." Six subjects indicated that the stimuli were adequate and comfortable, and four indicated that they would turn the stimuli down if they could. For these listeners, the overall levels were reduced by 5 dB, additional stimuli were presented, and the question was repeated. All found the level adequate and comfortable after this 5-dB reduction. Overall presentation levels for the HI listeners were at or below 92 dBA.

III. RESULTS

A. Consonant recognition

Figure 2 shows recognition for individual HI (filled symbols) and NH listeners (open symbols). The upper two rows of panels display data for SSN and the lower two rows of panels display data for multi-talker babble. Panels displaying performance following algorithm processing are immediately above those for IBM processing. Unprocessed speech-in-noise scores are represented by circles, algorithm-processed scores are represented by triangles, IBM-processed scores are represented by squares, and scores in quiet are represented by horizontal dashes. The benefit from processing is therefore represented by the height of the bar connecting symbols.

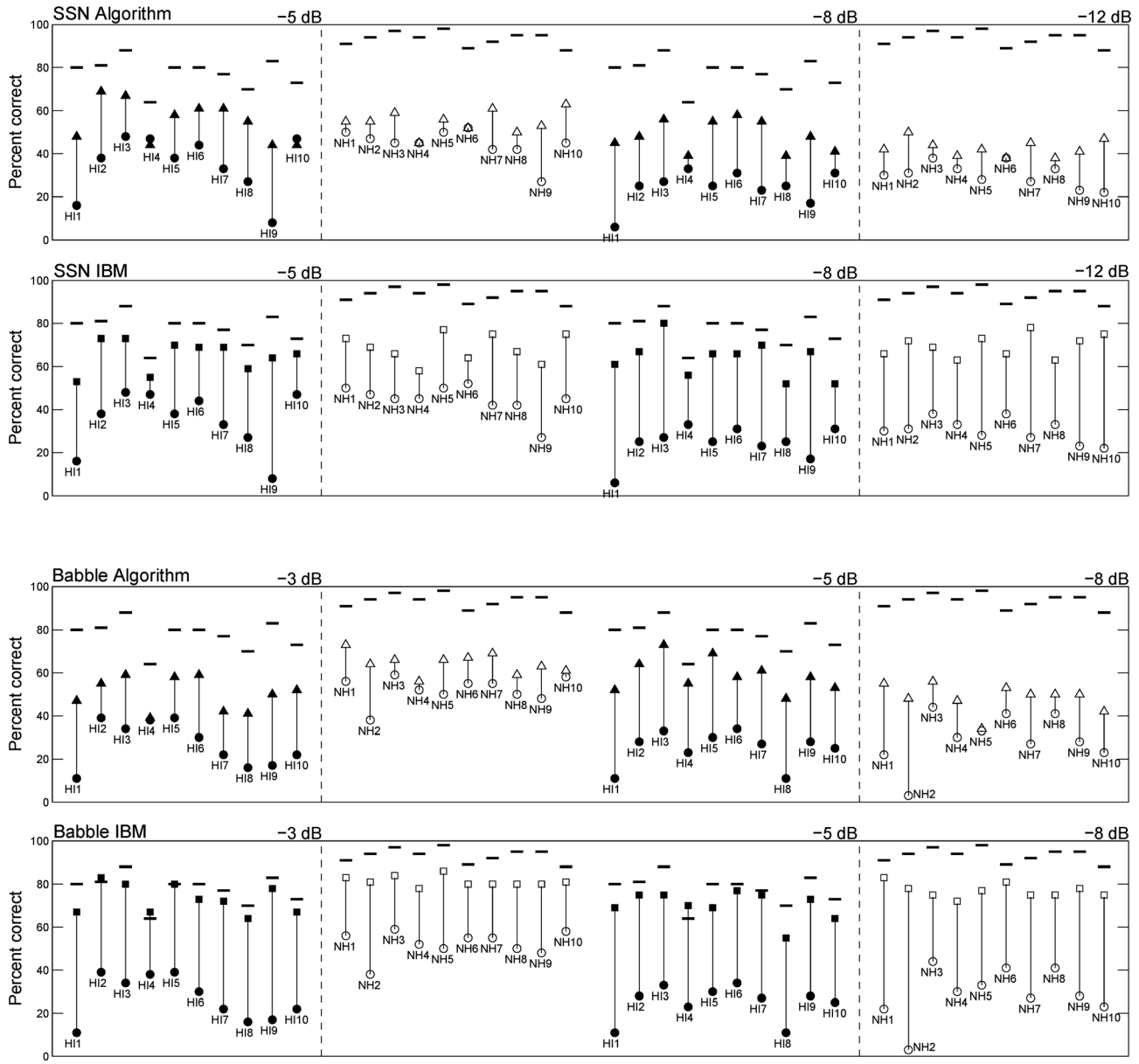

FIG. 2. Mean VCV phoneme recognition for each listener. Hearing-impaired listeners are represented by filled symbols and NH listeners are represented by open symbols. Unprocessed speech-in-noise conditions are represented by circles, algorithm-processed conditions are represented by triangles, IBM-processed conditions are represented by squares, and speech-in-quiet conditions are represented by horizontal lines. The top two rows represent recognition in speech-shaped noise at the three SNRs indicated, and the bottom two rows represent recognition in multi-talker babble at three SNRs. Panels displaying scores following algorithm processing are immediately above those displaying scores following IBM processing. Hearing-impaired listeners are numbered and plotted in order of increasing pure-tone average (as in Fig. 1).

All of the HI listeners received some benefit from algorithm processing for at least one of the SSN SNRs and both babble SNRs. Nine of the ten NH listeners received some benefit from algorithm processing for at least one of the SSN SNRs, and all received some benefit at both babble SNRs. With regard to IBM processing, all listeners received some benefit in all conditions. Generally speaking, scores for algorithm- or IBM-processed speech did not reach scores for speech in quiet.

Figure 3 displays group-mean recognition scores for HI and NH listeners. Data for SSN are displayed in the top row and data for babble are displayed in the bottom row. The dashed lines represent group-mean performance on speech in quiet. Apparent from the figure are increased group-mean recognition scores for both algorithm and IBM processing, relative to corresponding unprocessed speech in noise. Further, group-mean scores in the IBM conditions are greater than those in the corresponding algorithm conditions. This pattern held for both listener groups, both noise types and all SNRs. These observations were confirmed using a series of paired *t* tests and the procedure of Benjamini and Hochberg (1995) to limit the false-discovery rate to 0.05 or lower. It was found that algorithm and IBM processing increased scores significantly (alpha equivalent to 0.005 or lower) relative to corresponding unprocessed speech in noise. It was also found that IBM scores were significantly greater than corresponding algorithm scores (0.005 or lower). These significant effects held for all listener, noise type, and SNR conditions. Finally, algorithm and IBM scores were all significantly lower than corresponding scores for speech in quiet (0.05 or lower).

1. Algorithm processing

In accord with the results observed previously for everyday sentences (Healy *et al.*, 2013), the average benefit resulting from algorithm processing was greater for HI than for NH listeners. This advantage for HI listeners is reflected as (a) larger increases in scores from unprocessed speech in noise to algorithm processed and (b) smaller differences in scores between algorithm processed and speech in quiet. This HI-listener advantage is apparent within each center panel of Fig. 3, where SNR was equated across listeners. It is also observed when the rightmost panel is compared to the leftmost panel, for each noise type. This latter comparison allows an evaluation across similar baseline (unprocessed speech-in-noise) scores. Averaged across SNRs for SSN, the

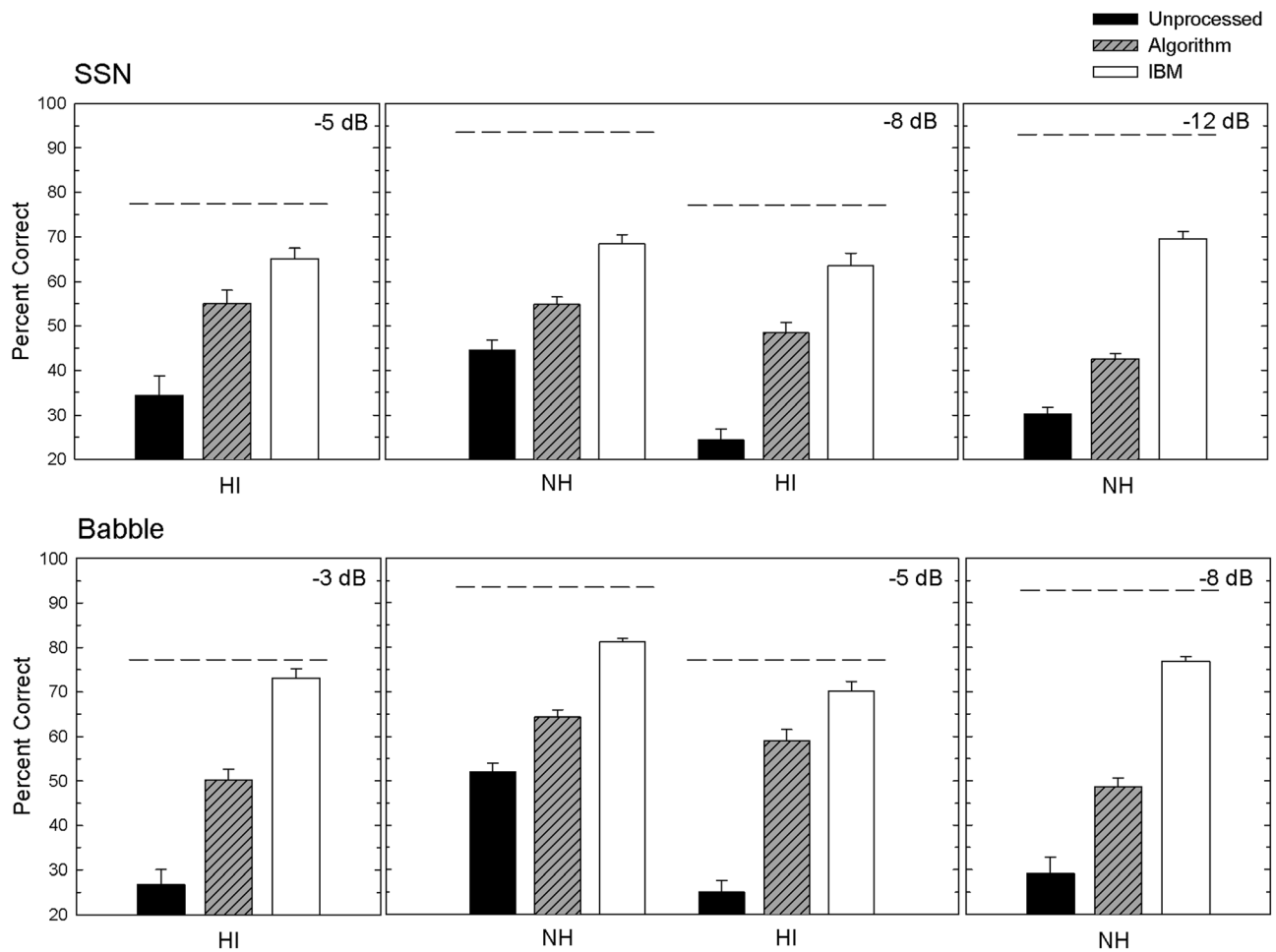

FIG. 3. Group mean recognition scores and standard errors for VCVs presented in SSN (top panels) and babble (bottom panels), at the SNRs indicated. Scores are presented separately for HI and NH listeners. The three columns reflect scores in unprocessed speech in noise, speech extracted by algorithm processing, and speech extracted by IBM processing. The dashed lines represent recognition of unprocessed speech in quiet for HI (SE = 2.12) and NH listeners (SE = 1.09).

algorithm benefit relative to unprocessed speech in noise was 22.3 percentage points for HI listeners vs 11.3 for NH. The distance to scores in quiet averaged 25.8 percentage points for HI listeners vs 44.6 for NH. Averaged across SNRs for babble, the algorithm benefit was 28.7 percentage points for HI listeners vs 15.9 for NH. The distance to scores in quiet averaged 22.9 points for HI listeners vs 36.8 for NH.

Benefit resulting from algorithm processing was calculated as a proportion to capture both increase from unprocessed speech in noise and distance to scores in quiet. This proportion increase (P) was calculated as

$$P = \frac{\text{Proc} - \text{UNP}}{Q - \text{UNP}}, \tag{3}$$

where Proc is the score following algorithm processing, UNP is the corresponding score in unprocessed noise, and Q is the corresponding score in quiet. The HI-listener advantage was confirmed using this metric and planned comparisons (uncorrected unpaired t tests) to assess benefit averaged across SNRs. Algorithm benefit was significantly greater for HI than for NH listeners in SSN [0.42 vs 0.20, $t(18) = 2.94$, $p < 0.01$] and in babble [0.55 vs 0.29, $t(18) = 5.94$, $p < 0.001$].

Also in accord with prior results involving sentences, the benefit resulting from algorithm processing was generally larger in babble than in SSN. This babble advantage is somewhat smaller than the HI-listener advantage just described, but it appears for both benefit relative to unprocessed speech in noise and for distance to scores in quiet. This advantage holds for all but one comparison across corresponding top versus bottom panels in Fig. 3. Averaged across SNRs, the algorithm benefit relative to unprocessed speech in noise was 28.7 percentage points for babble vs 22.3 for SSN (HI listeners) and 15.9 percentage points for babble vs 11.3 for SSN (NH listeners). The distance to scores in quiet averaged 22.9 percentage points for babble vs 25.8 for SSN (HI listeners) and 36.8 percentage points for babble vs 44.6 for SSN (NH listeners). The babble advantage was also assessed using planned comparisons (uncorrected paired t tests) on the proportion metric averaged across SNRs. Algorithm benefit in terms of proportion increase was numerically larger for babble than for SSN, but statistically equivalent, for HI listeners [0.55 vs 0.42, $t(9) = 2.00$, $p = 0.08$, power = 0.33] and for NH listeners [0.29 vs 0.20, $t(9) = 1.88$, $p = 0.09$, power = 0.29].

Another planned comparison of interest involves HI-listener performance following algorithm processing versus NH-listener performance prior to algorithm processing (unprocessed speech in noise), in conditions of identical noise type and SNR (Fig. 3, center panels). It was found that mean HI-listener recognition scores were significantly greater than those for their NH counterparts in babble [59.1% vs 52.0%, $t(18) = 2.21$, $p < 0.05$], and numerically greater than but statistically equivalent in SSN [48.4% vs 44.5%, $t(18) = 1.22$, $p = 0.24$, power = 0.10].

2. IBM processing

With regard to IBM processing, the benefit was again generally larger for HI than for NH listeners. The increases relative to unprocessed speech in noise were substantially larger for HI listeners in conditions of common noise type and SNR (Fig. 3, center panels). Although this HI-listener advantage is less apparent when conditions are compared across similar baseline scores (rightmost panel compared to leftmost panel, for each noise type), it appears again when IBM scores are compared to scores in quiet: Whereas large differences between IBM and speech-in-quiet scores remain for NH listeners, these differences are substantially reduced for HI listeners.

Averaged across SNRs for SSN, the IBM benefit relative to unprocessed speech in noise was 35.0 percentage points for HI listeners vs 31.6 for NH. The distance to scores in quiet averaged 13.1 percentage points for HI listeners vs 24.3 for NH. Averaged across SNRs for babble, the IBM benefit relative to unprocessed speech in noise was 45.8 percentage points for HI listeners vs 38.5 for NH. The distance to scores in quiet averaged 5.9 points for HI listeners vs 14.2 for NH. The HI-listener advantage was found to be significant for SSN [0.71 vs 0.56, $t(18) = 3.42$, $p < 0.005$] and for babble [0.90 vs 0.72, $t(18) = 4.56$, $p < 0.001$] using planned comparisons of the proportion-benefit metric [Eq. (3)], again averaged across SNRs.

Also in accord with what was observed for algorithm processing, the benefit resulting from IBM processing was greater in babble than in SSN. This babble advantage appears for both benefit relative to unprocessed speech in noise and for distance to scores in quiet, across each corresponding top versus bottom panel in Fig. 3. Averaged across SNRs, the IBM benefit relative to unprocessed speech in noise was 45.8 percentage points for babble vs 35.0 for SSN (HI listeners) and 38.5 percentage points for babble vs 31.6 for SSN (NH listeners). The distance to scores in quiet averaged 5.9 percentage points for babble vs 13.1 for SSN (HI listeners) and 14.2 percentage points for babble vs 24.3 for SSN (NH listeners). Planned comparisons of the proportion metric averaged across SNRs indicated that the IBM benefit was significantly greater for babble than for SSN, for HI listeners [0.90 vs 0.71, $t(9) = 4.41$, $p < 0.005$] and for NH listeners [0.72 vs 0.56, $t(9) = 4.82$, $p < 0.001$].

B. Speech-cue transmission

An information-transmission analysis (Miller and Nicely, 1955) employing the sequential information analysis (SINFA) of Wang and Bilger (1973) was conducted on the averaged consonant-confusion matrices to determine the percentage of speech cues transmitted to listeners in each condition. See http://web.cse.ohio-state.edu/~dwang/ for these consonant-confusion matrices. The cues examined were those associated with voicing (voiced, unvoiced), manner of articulation (fricative, plosive, nasal, sibilant), and place of articulation (front, mid, back). Figure 4 displays the results of this analysis. Displayed in separate panels are results for HI and NH listeners and the two noise types. Displayed within each panel is information transmitted in each processing condition as well as values averaged across the two SNRs in each panel. Finally, information transmitted in the quiet condition is displayed at the right of each panel.

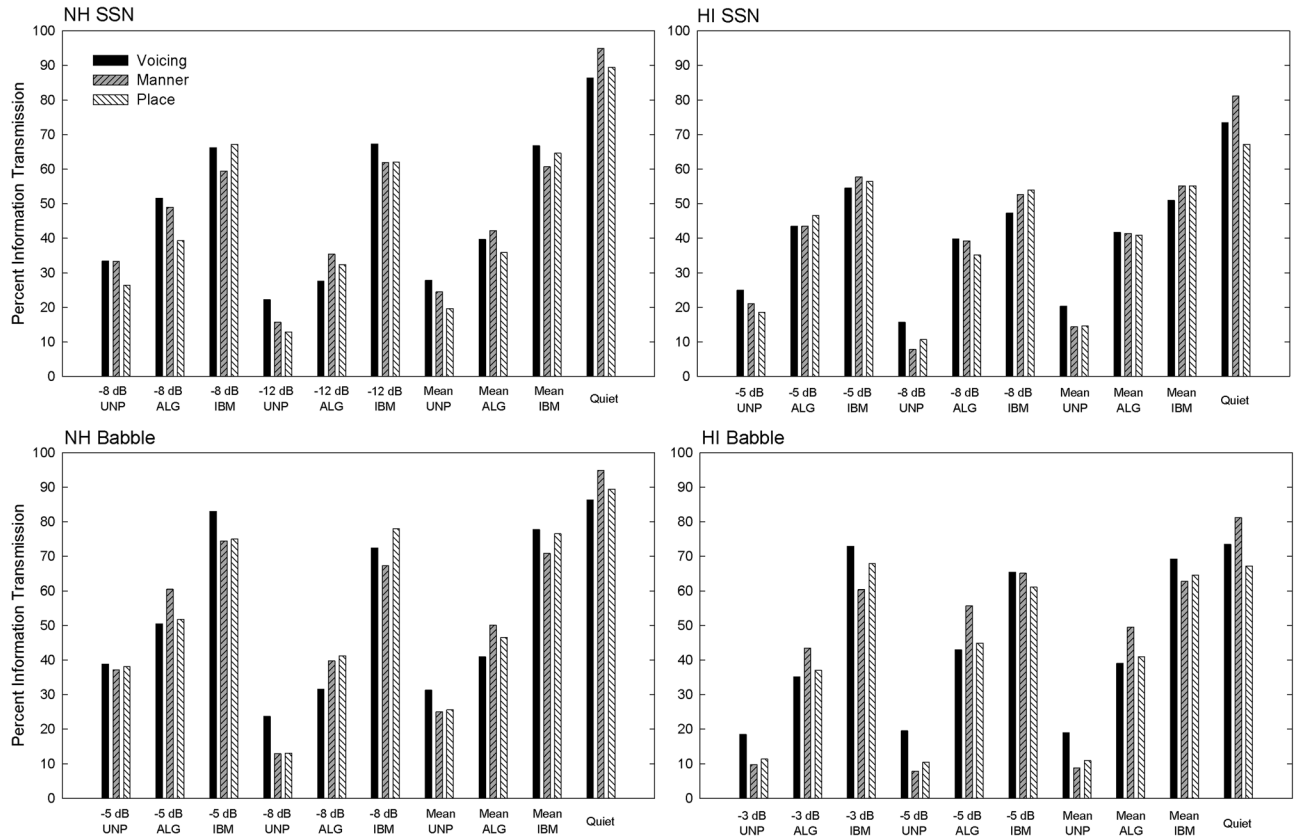

FIG. 4. Percent information transmitted for the speech features associated with voicing, manner of articulation, and place of articulation. The four panels display information for HI and NH listeners in SSN and in babble. Each panel contains information transmission for unprocessed speech in noise (UNP), algorithm processing (ALG), and IBM processing, at the SNRs indicated, as well as averaged across the two SNRs. Also displayed is information transmission for speech in quiet.

Overall, all three classes of speech cues were transmitted at similar levels within each condition. For unprocessed speech in noise, the transmission of voicing information was greatest, and transmission of the other two cues was similar and slightly reduced. For speech in quiet, transmission of manner information was greatest and transmission of the remaining cues was similar and slightly reduced. These effects held for both listener and noise types. The speech cues transmitted by the algorithm followed generally the pattern of cues transmitted by speech in quiet. For most IBM conditions, the transmission of voicing and place was slightly better than the transmission of manner.

Figure 5 displays information transmission as ratios of the values obtained in the corresponding unprocessed speech-in-noise conditions, then averaged across the two SNRs. Thus, these values reflect increases in information transmission resulting from algorithm or IBM processing. Most notable is the similarity in patterns between corresponding algorithm versus IBM conditions, for each noise and listener type (compare the rightmost pair of column trios to one another, and the leftmost pair of column trios to one another, in each panel). Increases in speech-cue transmission were generally greatest for manner of articulation in HI listeners and for place of articulation in NH listeners. This pattern generally held for both noise types. The ratio increases in speech-cue transmission resulting from algorithm processing averaged 3.4 for the HI listeners and 1.9 for the NH listeners. Ratio increases for the IBM averaged 4.7 for the HI listeners and 3.2 for the NH listeners.

IV. DISCUSSION

A. Consonant recognition

1. Algorithm and IBM processing

The first main goal of the current study was to assess consonant-in-noise recognition after processing by the algorithm and by the IBM. The current results indicate that a more efficient version of the algorithm described by Healy *et al.* (2013) is capable of improving recognition of isolated consonants. This is encouraging for at least two reasons. First, isolated phonemes lack the semantic context and multitude of redundant cues that characterize sentences, which limits listeners' ability to use top-down processing mechanisms. Accordingly, phoneme recognition typically requires additional acoustic information to achieve similar levels of recognition, and it is typically presumed to involve a greater reliance on bottom-up processing of these cues. Second, the introduced efficiency, including a reduction from 128 DNNs to 1, did not prevent the algorithm from substantially improving speech recognition for NH and HI listeners. As they did with sentences (Healy *et al.*, 2013), the older HI listeners tested here performed as well or better than their young NH counterparts

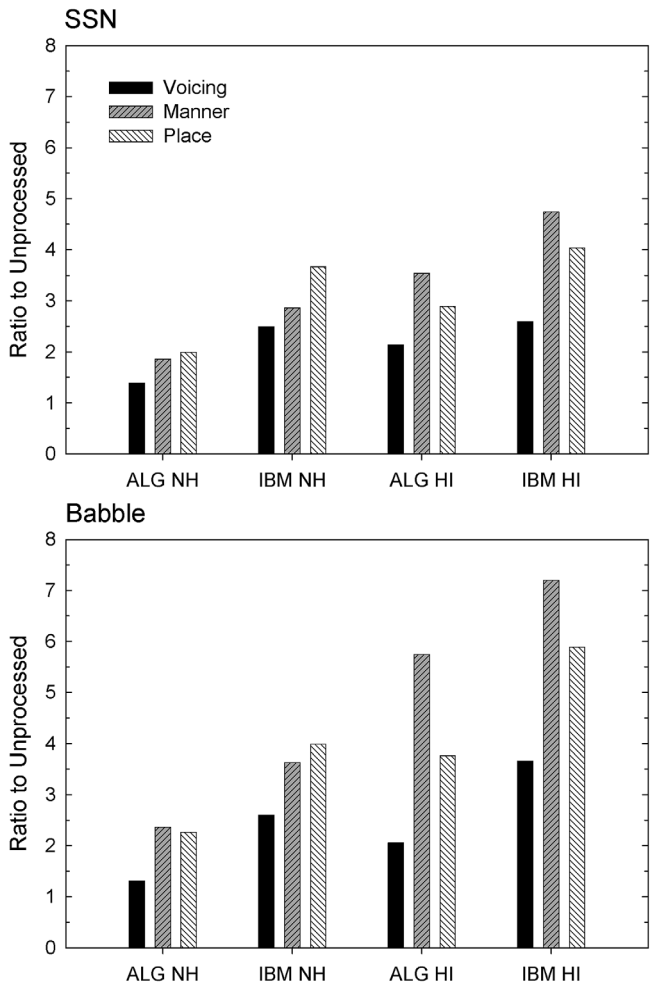

FIG. 5. Increases in information transmission for the speech features associated with voicing, manner of articulation, and place of articulation. Values are displayed separately for algorithm (ALG) and IBM processing, and for HI and NH listeners. They are displayed as ratios of the values obtained in the corresponding unprocessed speech-in-noise conditions, then averaged across the two SNRs employed. Values in SSN are displayed in the top panel and values in babble are displayed in the bottom panel.

in identical conditions of noise, once the HI listeners had access to the current algorithm (compare HI algorithm to NH unprocessed). This effectiveness despite increased processing efficiency bodes well for the eventual goal of real-time implementation in hearing technology, including hearing aids and cochlear implants.

One issue that must be addressed before an algorithm such as this can be implemented into wearable devices involves processing speed. A major advantage of the current classification-based framework is that much of the processing load is transferred to an earlier stage involving DNN training. The operational stage primarily involves the extraction of features from the sound mixture and binary labeling of each T-F unit by the trained DNN classifiers. The time required to perform these tasks on a 3-s. speech-plus-noise mixture, using an Intel Xeon X5650 2.67 GHz CPU, was 143 ms per frequency channel (times 64 channels) using the 2013 implementation of the algorithm. The current efficient implementation requires 215 ms to perform these same tasks, which represents a 43-fold increase in processing speed.

The magnitude of benefit observed currently for isolated consonants is smaller than that observed by Healy *et al.* (2013) for sentences, at least in terms of percentage-point benefit. This may be understood in terms of the psychometric transfer functions relating recognition to information content, which are far shallower for isolated phonemes than for sentences (e.g., ANSI, 1969, Fig. 15). Although transfer functions for the specific recordings employed here are not available, the functions that do exist support the view that speech-information benefits resulting from algorithm processing are similar across the current study and the former study involving sentences, despite the different percent-point gains.

Surprisingly, group-mean algorithm scores were lower at −3 dB SNR babble than at −5 dB, for nine of the ten HI listeners. This apparent reversal in scores was not observed for the IBM. However, the corresponding unprocessed speech-in-noise scores were similar in these conditions (group means within 2 percentage points, and similarly reversed in direction for four of the ten HI listeners). We do not have a good explanation for the algorithm's poorer performance at this input SNR.

As with the algorithm, it was found that the IBM improved recognition of isolated phonemes, thus extending results observed previously for sentences (Anzalone *et al.*, 2006; Brungart *et al.*, 2006; Li and Loizou, 2008; Wang *et al.*, 2008, 2009; Cao *et al.*, 2011; Sinex, 2013). This result should not be surprising, given the effectiveness of algorithm processing observed here and the fact that the algorithm aimed to estimate the IBM. The fact that IBM scores were uniformly superior to algorithm scores indicates that the binary masks estimated by the algorithm were not as effective as the ideal masks, which were constructed from premixed speech and noise signals. This topic of mask-estimation accuracy is discussed in Sec. IV B.

2. HI-listener and babble advantages

In accord with previous work involving sentences and either algorithm processing (Healy *et al.*, 2013) or IBM processing (Anzalone *et al.*, 2006; Wang *et al.*, 2009), gains were greater for HI than for NH listeners and for modulated than for stationary backgrounds. The HI-listener advantage is promising because it indicates that the algorithm and the IBM are effective despite the multitude of limitations imposed by the impaired auditory system. Speech extracted from noise using the estimated or ideal binary masks contains a sufficient concentration of cues to be successfully processed despite these limitations, even for isolated phonemes. The babble advantage (which was statistically significant for sentences, but only marginally so here for algorithm processing using a low-power test) is promising because it indicates that the algorithm and the IBM are most effective in backgrounds that represent real-world interferers. Finally, algorithm improvements in these nonstationary backgrounds represent a far larger technical challenge than do improvements in stationary noise (Wang and Brown, 2006).

Relationships between various HI-subject variables and performance were examined in an attempt to identify subject characteristics associated with maximum benefit. No

significant correlations were observed between subject variables including age or PTA, and performance including unprocessed speech-in-noise scores, algorithm or IBM scores, or algorithm or IBM benefit. Rather, most every HI subject received some benefit from algorithm and from IBM processing, despite a range of audiometric profiles and a wide range of ages, from 25 to 73 yr.

3. Assessing benefit

Benefit can be assessed in two ways: Increases in scores from unprocessed speech in noise and distance between processed scores and scores in quiet. The proportion metric on which planned comparisons were made reflects both of these components. With regard to recognition of unprocessed consonants in noise, scores obtained here match or are lower than those obtained in other studies employing these speech materials, NH listeners, similar SNRs, and different simplified SSNs (e.g., Fu *et al.*, 1998; Apoux and Healy, 2011). The somewhat lower scores obtained currently may potentially be attributed to the greater masking by the current SSN, which closely matched the long-term average amplitude spectrum of the speech. The current unprocessed scores are also similar to or slightly greater than those for corresponding sentence-in-noise conditions in Healy *et al.* (2013). But it should be noted that comparisons across speech materials can be difficult to interpret. As suggested earlier, isolated phonemes require additional acoustic information content to yield recognition scores similar to sentences. Some of this advantage for sentences likely reflects the benefit of semantic context. But in contrast, the open-set response paradigm typically employed for sentences can potentially reduce scores relative to the closed-set paradigm typically employed for consonants. The comparison between the current unprocessed consonant-recognition scores and those involving sentences likely reflects this tradeoff.

Consonants allow a different level of analysis relative to sentences. Sentences better reflect real-world recognition because everyday communication involves words in sentences that follow grammatical rules and a contextual theme. But because of their redundancy and ability to be understood with only a small subset of cues, sentence scores quickly reach ceiling values. Consonants' relative robustness to ceiling effects can be advantageous. In the current study, speech in quiet was always heard last. This provided a maximum amount of practice prior to this condition and tended to maximize scores in this condition. The procedure likely downplayed the benefit from algorithm and IBM processing. But despite this procedure, speech-in-quiet scores were slightly below ceiling levels for NH listeners and far below ceiling for HI listeners. These lower scores for HI listeners should be anticipated, because only audibility was corrected, and no attempt was made to rectify suprathreshold deficits associated with hearing impairment.

One consequence of the current reduction in ceiling effects is the finding that IBM scores were significantly lower than speech-in-quiet scores. This difference is generally obscured when using sentence materials, as IBM scores tend to reach ceiling values (e.g., Brungart *et al.*, 2006; Li and Loizou, 2008; Wang *et al.*, 2009; Sinex, 2013). Another consequence is that the increased effectiveness of IBM processing for HI listeners is reflected with clarity when assessed as the distance between IBM scores and scores in quiet (see Fig. 3). For NH listeners, a substantial difference in scores remains between these conditions. In contrast, this difference for HI listeners is reduced to as low as 4 percentage points. This result suggests that IBM processing is effective enough to improve HI-listener scores, but not NH-listener scores, to values near those for speech in quiet, even in the absence of ceiling effects.

The possibility exists that these IBM scores could be increased further through the manipulation of LC values. The choices made here, with LC about 5 dB smaller than input SNR, were motivated by values shown to be effective for noisy sentences (Brungart *et al.*, 2006; Li and Loizou, 2008; Wang *et al.*, 2009; Kjems *et al.*, 2009). It is possible that LC values that are most effective for consonant materials will differ from those for sentence materials, perhaps due to the increased requirements for acoustic speech information and increased reliance on bottom-up processing.

B. Accuracy of the estimated binary mask

A second main goal of the current study was to assess the effective accuracy of the binary mask estimated by the current algorithm. This accuracy was assessed by comparing speech features transmitted to listeners by the algorithm versus by the IBM. Apparent from Fig. 5 is the marked similarity in how the algorithm and the IBM increase cue transmission for a given noise and listener type. This indicates that the masks estimated by the algorithm are transmitting cues in a fashion highly similar to that of the IBM, which suggests that the masks are being estimated with perceptually relevant accuracy. This further indicates that the lower performance in algorithm-processed conditions relative to corresponding IBM-processed conditions is not due to a deficiency in any particular aspect of mask estimation or a deficiency in the transmission of any particular speech feature. Instead, these results suggest that the algorithm is estimating the IBM with general accuracy, but simply with imperfect fidelity.

C. Speech features transmitted by algorithm- and IBM-processed speech

A third main goal of the current study was to determine the features of speech that are conveyed to HI and NH listeners by the estimated and ideal binary masks. Figure 4 displays the results of this information-transmission analysis. The primary conclusion that can be drawn from these results is that both the algorithm and the IBM transmit speech cues with substantial uniformity. The speech features examined here were transmitted to listeners at generally similar levels, and no class of cues failed to be transmitted. The possibility existed that the high sentence recognition obtained previously using the algorithm and/or IBM was based on transmission of only a subset of speech features. However, the current results indicate that this is not the case.

Figure 5 shows that the increases in cue transmission were highly similar for the algorithm versus the IBM when compared within the HI group and when compared within the NH group. The patterns are also highly similar across the two noise types. However, a somewhat different pattern of speech-cue transmission increase is observed across HI versus NH listeners. Whereas HI listeners experienced greatest gains in manner of articulation, NH listeners generally experienced greatest gains in place of articulation. This modest difference in the pattern of speech-cue transmission increases across listener groups should not be surprising, as differences have been observed previously (e.g., Bilger and Wang, 1976; Wang *et al.*, 1978; but see Gordon-Salant, 1987). The difference between listener groups observed here is likely attributable to a reduction in the transmission of place-of-articulation cues to HI listeners, due to their broad auditory tuning. Place of articulation is known to be a cue encoded with substantial spectral detail (e.g., Shannon *et al.*, 1995) and one requiring fine frequency resolution for accurate perception.

V. CONCLUSIONS

(1) The current segregation algorithm is capable of improving recognition of isolated consonants, for which semantic context is absent and increased reliance on bottom-up acoustic speech cues is typically required. In accord with results from sentences (Healy *et al.*, 2013) these improvements were greatest for HI listeners. Older HI listeners having access to the algorithm performed as well or better than young NH listeners in conditions of identical noise.
(2) The IBM also substantially increased recognition of isolated consonants. Not surprisingly, these increases exceeded those resulting from the algorithm and were greatest for HI listeners and for a babble background. These scores generally did not reach scores in quiet.
(3) The binary masks estimated by the segregation algorithm transmitted speech features to listeners in a fashion similar to that of the IBM, suggesting that the algorithm is estimating the IBM with considerable effective or perceptual accuracy. Further, both the algorithm and the IBM appear to be without any specific deficiency in the transmission of speech cues.
(4) The current algorithm has remained highly effective, while being made substantially more efficient. The 43-fold increase in segregation speed makes the current algorithm more amenable to real-time implementation in devices such as hearing aids and cochlear implants.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Institute on Deafness and other Communication Disorders (R01 DC008594 to E.W.H. and R01 DC012048 to D.L.W.) and AFOSR (FA9550-12-1-0130 and an STTR subcontract from Kuzer to D.L.W.). Portions were presented at the 167th meeting of the Acoustical Society of America. We gratefully acknowledge the manuscript-preparation contributions of Carla Youngdahl, Brittney Carter, and Jordan Vasko, and computing resources from the Ohio Supercomputer Center.

American National Standards Institute (**1969**). S3.5 (R1986), *Methods for the Calculation of the Articulation Index* (Acoustical Society of America, New York).
American National Standards Institute (**1987**). S3.39 (R2012), *Specifications for Instruments to Measure Aural Acoustic Impedance and Admittance (Aural Acoustic Immittance)* (Acoustical Society of America, New York).
American National Standards Institute (**2004**). S3.21 (R2009), *Methods for Manual Pure-Tone Threshold Audiometry* (Acoustical Society of America, New York).
American National Standards Institute (**2010**). S3.6, *Specification for Audiometers* (Acoustical Society of America, New York).
Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.
Apoux, F., and Healy, E. W. (**2011**). "Relative contribution of target and masker temporal fine structure to the unmasking of consonants in noise," J. Acoust. Soc. Am. **130**, 4044–4052.
Bacon, S. P., and Healy, E. W. (**2000**). "Effects of ipsilateral and contralateral precursors on the temporal effect in simultaneous masking with pure tones," J. Acoust. Soc. Am. **107**, 1589–1597.
Benjamini, Y., and Hochberg, Y. (**1995**). "Controlling the false discovery rate: A practical and powerful approach to multiple testing," J. R. Stat. Soc. Ser. B (Methodol.) **57**, 289–300.
Bilger, R. C., and Wang, M. D. (**1976**). "Consonant confusions in patients with sensorineural hearing loss," J. Speech Hear. Res. **19**, 718–748.
Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.
Byrne, D., and Dillon, H. (**1986**). "The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," Ear Hear. **7**, 257–265.
Cao, S., Li, L., and Wu, X. (**2011**). "Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise," J. Acoust. Soc. Am. **129**, 2227–2236.
Chen, C.-P., and Bilmes, J. A. (**2007**). "MVA processing of speech features," IEEE Trans. Audio. Speech Lang. Process. **15**, 257–270.
Chen, J., Wang, Y., and Wang, D. L. (**2014**). "A feature study for classification-based speech separation at very low signal-to-noise ratio," *Proc. ICASSP*, pp. 7089–7093.
Dillon, H. (**2012**). *Hearing Aids*, 2nd ed. (Boomerang, Turramurra, Australia), p. 232.
Duchi, J., Hazan, E., and Singer, Y. (**2011**). "Adaptive subgradient methods for online learning and stochastic optimization," J. Machine Learn. Res. **12**, 2121–2159.
Fu, Q.-J., Shannon, R. V., and Wang, X. (**1998**). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," J. Acoust. Soc. Am. **104**, 3586–3596.
Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1993**). "DARPA TIMIT acoustic-phonetic continuous speech corpus," Technical Report No. NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD.
Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.
Gordon-Salant, S. (**1987**). "Consonant recognition and confusion patterns among elderly hearing-impaired subjects," Ear Hear. **8**, 270–276.
Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. L. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**, 3029–3038.
Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (**2012**). "Improving neural networks by preventing co-adaptation of feature detectors," arXiv:1207.0580.
Hu, G., and Wang, D. L. (**2001**). "Speech segregation based on pitch tracking and amplitude modulation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (IEEE, New Paltz, NY), pp. 79–82.
Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494.
Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. L. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.

Li, N., and Loizou, P. C. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Moore, B. C. J. (**2007**). *Cochlear Hearing Loss*, 2nd ed. (Wiley, Chichester, UK), pp. 201–232.
Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., and Wang, X. (**1999**). "Consonant recordings for speech testing," J. Acoust. Soc. Am. **106**, L71–74.
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Sinex, D. G. (**2013**). "Recognition of speech in noise after application of time-frequency masks: Dependence on frequency and threshold parameters," J. Acoust. Soc. Am. **133**, 2390–2396.
Stickney, G. S., and Assmann, P. F. (**2001**). "Acoustic and linguistic factors in the perception of bandpass-filtered speech," J. Acoust. Soc. Am. **109**, 1157–1165.
Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer, Norwell, MA), pp. 181–197.
Wang, D. L., and Brown, G. J., Eds. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley, Hoboken, NJ), pp. 1–44.
Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2008**). "Speech perception of noise with binary gains," J. Acoust. Soc. Am. **124**, 2303–2307.
Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**, 2336–2347.
Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.
Wang, M. D., Reed, C. M., and Bilger, R. C. (**1978**). "A comparison of the effects of filtering and sensorineural hearing loss on patterns of consonant confusions," J. Speech Hear. Res. **21**, 5–36.
Wang, Y., Han, K., and Wang, D. L. (**2013**). "Exploring monaural features for classification-based speech segregation," IEEE Trans. Audio. Speech Lang. Process. **21**, 270–279.
Wang, Y., and Wang, D. L. (**2013**). "Towards scaling up classification-based speech separation," IEEE Trans. Audio. Speech Lang. Process. **21**, 1381–1390.
Warren, R. M., Riener, K. R., Bashford, J. A., Jr., and Brubaker, B. S. (**1995**). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," Percept. Psychophys. **57**, 175–182.