Speech intelligibility in background noise with ideal binary time-frequency masking

DeLiang Wang[a)]
Department of Computer Science & Engineering and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

Ulrik Kjems, Michael S. Pedersen, and Jesper B. Boldt
Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark

Thomas Lunner
Oticon Research Centre Eriksholm, Kongevejen 243, DK-3070 Snekkersten, Denmark and Department of Clinical and Experimental Medicine and Technical Audiology, Linköping University, S-58183 Linköping, Sweden

(Received 4 March 2008; revised 21 January 2009; accepted 27 January 2009)

Ideal binary time-frequency masking is a signal separation technique that retains mixture energy in time-frequency units where local signal-to-noise ratio exceeds a certain threshold and rejects mixture energy in other time-frequency units. Two experiments were designed to assess the effects of ideal binary masking on speech intelligibility of both normal-hearing (NH) and hearing-impaired (HI) listeners in different kinds of background interference. The results from Experiment 1 demonstrate that ideal binary masking leads to substantial reductions in speech-reception threshold for both NH and HI listeners, and the reduction is greater in a cafeteria background than in a speech-shaped noise. Furthermore, listeners with hearing loss benefit more than listeners with normal hearing, particularly for cafeteria noise, and ideal masking nearly equalizes the speech intelligibility performances of NH and HI listeners in noisy backgrounds. The results from Experiment 2 suggest that ideal binary masking in the low-frequency range yields larger intelligibility improvements than in the high-frequency range, especially for listeners with hearing loss. The findings from the two experiments have major implications for understanding speech perception in noise, computational auditory scene analysis, speech enhancement, and hearing aid design. [DOI: 10.1121/1.3083233]

PACS number(s): 43.71.Gv, 43.66.Dc [KWG] Pages: 2336–2347

I. INTRODUCTION

Human speech communication typically takes place in complex acoustic backgrounds with environmental sound sources, competing voices, and ambient noise. It is remarkable that human speech understanding remains robust in the presence of such interference. This perceptual ability is thought to involve the process of auditory scene analysis (Bregman, 1990), by which the auditory system first analyzes a noisy input into a collection of sensory elements in time and frequency, also known as segments (Wang and Brown, 2006), and then selectively groups segments into auditory streams which correspond to sound sources.

It is well known that listeners with hearing loss have greater difficulty in speech perception in background noise. A standard way to quantify speech intelligibility in noise is a speech-reception threshold (SRT), which is the mixture signal to noise ratio (SNR) required to achieve a certain intelligibility score, typically 50%. Hearing-impaired (HI) listeners need 3–6 dB higher SNR than normal-hearing (NH) listeners in order to perform at the same level in typical noisy backgrounds (Plomp, 1994; Alcantara *et al.*, 2003). For speech-shaped noise (SSN) which is a steady noise with a long-term spectrum matching that of natural speech, the SRT increase for HI listeners is from 2.5 to 7 dB (Plomp, 1994). For fluctuating noise or competing speech, the increase is considerably higher (Festen and Plomp, 1990; Hygge *et al.*, 1992; Eisenberg *et al.*, 1995; Peters *et al.*, 1998); for a single competing talker, the increase is as much as 10–15 dB (Carhart and Tillman, 1970; Festen and Plomp, 1990; Peters *et al.*, 1998). Note that, for typical speech materials, a 1 dB increase in SRT leads to a 7%–19% reduction in the percent correct score, and a 2–3 dB elevation creates a significant handicap for understanding speech in noisy listening conditions (Moore, 2007).

Although modern hearing aids improve the audibility and comfort of noisy speech, their ability to improve the intelligibility of noisy speech is unfortunately very limited (Dillon, 2001; Alcantara *et al.*, 2003). Extensive research has been made to develop noise reduction algorithms in order to close the SRT gap between HI and NH listeners. Monaural speech enhancement algorithms, such as Wiener filtering and spectral subtraction, perform statistical analysis of speech and noise and then estimate clean speech from noisy speech (Lim, 1983; Benesty *et al.*, 2005). Although these algorithms produce SNR improvements, they have not led to increased speech intelligibility for human subjects (Levitt, 2001;

[a)]Author to whom correspondence should be addressed. Electronic mail: dwang@cse.ohio-state.edu

 0001-4966/2009/125(4)/2336/12/$25.00

Moore, 2003b; Edwards, 2004). Attempts have also been made to directly enhance speech cues, especially formants which are spectral peaks of speech (Bunnell, 1990; Simpson *et al.*, 1990). This processing results in clearer formant structure; however, listening tests with both NH and HI listeners show little improvement in speech intelligibility (Baer *et al.*, 1993; Alcantara *et al.*, 1994; Dillon, 2001). Unlike monaural speech enhancement, beamforming (spatial filtering) with a microphone array has been demonstrated to achieve significant speech intelligibility improvements, particularly with large arrays (Kates and Weiss, 1996; Levitt, 2001; Schum, 2003). On the other hand, practical considerations of hearing aid design often limit the size of an array to two microphones, and the effectiveness of beamforming degrades in the presence of room reverberation (Greenberg and Zurek, 1992; Levitt, 2001; Ricketts and Hornsby, 2003). Additionally, to benefit from spatial filtering target speech and interfering sounds must originate from different directions.

Recent research in computational auditory scene analysis (CASA) has led to the notion of an ideal binary time-frequency mask as a performance upper bound to measure how well CASA algorithms perform (Wang and Brown, 2006). With a two-dimensional time-frequency (*T-F*) representation or decomposition of the mixture of target and interference, where elements in the representation are called *T-F* units, an ideal binary mask (IBM) is defined as a binary matrix within which 1 denotes that the target energy in the corresponding *T-F* unit exceeds the interference energy by a predefined threshold and 0 denotes otherwise. The threshold is called the local SNR criterion (*LC*), measured in decibels. More specifically, IBM is defined as

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

where $s(t,f)$ denotes the target energy within the unit of time t and frequency f and $n(t,f)$ the noise energy in the *T-F* unit, with both $s(t,f)$ and $n(t,f)$ measured in decibels. The mask is considered ideal because its construction requires access to the target and masker signals prior to mixing, and under certain conditions the IBM with $LC=0$ dB has the optimal SNR gain among all the binary masks (Wang, 2005; Li and Wang, 2009). As a separation technique, applying the IBM with $LC=0$ dB to the mixture input retains the *T-F* regions of the mixture where target energy is stronger than interference energy while removing the *T-F* regions where target energy is weaker than interference energy.

Varying *LC* results in different IBMs. Recently, Brungart *et al.* (2006) tested the effects of IBM with different *LC* values on speech mixtures with one target utterance and 1–3 competing utterances of the same talker, where the sound levels of all the utterances are set to be equal. Their experimental results show that, when $0\ \mathrm{dB} \geq LC \geq -12$ dB, IBM produces nearly perfect intelligibility scores, which are dramatically higher than in a control condition where speech mixtures are presented to listeners without processing. They suggest that the choice of $LC=-6$ dB, which lies near the center of the performance plateau, may be better than the commonly used 0 dB *LC* for intelligibility improvement. Furthermore, they attribute the intelligibility improvement to the removal of informational masking which occurs when the listener is unable to successfully extract or segregate acoustically detectable target information from the mixture. Anzalone *et al.* (2006) investigated the intelligibility improvements of a related version of IBM, defined by a comparison between target energy and a threshold rather than a comparison between target energy and interference energy. Using mixtures of speech and SSN, they found that IBM leads to substantial SRT reductions: more than 7 dB for NH listeners and more than 9 dB for HI listeners. In addition they reported that, while NH listeners benefit from ideal masking in both the low-frequency (LF) and high-frequency (HF) ranges, HI listeners benefit from ideal masking only at LFs (up to 1.5 kHz). Li and Loizou (2007) used the IBM to generate "glimpses," or *T-F* regions with stronger target energy, to study several factors that influence glimpsing of speech mixed with babble noise. Their results show that it is important to generate glimpses in the LF to mid-frequency range (up to 3 kHz) that includes the first and the second formant of speech, but not necessary to glimpse a whole utterance; high intelligibility is achieved when the listener can obtain glimpses in a majority of time frames. More recently, Li and Loizou (2008b) extended the findings of Brungart *et al.* (2006) to different types of background interference, including speech babble and modulated SSN. Moreover, they evaluated the impact of deviations from the IBM on intelligibility performance and found that there is a gradual drop as the amount of mask errors increases. A subsequent study by Li and Loizou (2008a) shows that NH listeners obtain significant intelligibility improvements from IBM processing with as few as 12 frequency channels, and IBM processing in the LF to mid-frequency range that includes the first and the second formant appears sufficient.

In this paper, we evaluate the effects of IBM processing on speech intelligibility with two kinds of background noise: SSN and cafeteria noise, using both NH and HI listeners. While SSN is commonly used in the literature, the cafeteria noise we use contains a conversation between two speakers in a cafeteria background and it resembles the kind of noise typically encountered in everyday life. Our study adopts the standard IBM definition with a comparison between target and interference and measures speech intelligibility by SRT at the 50% level. As suggested by the findings of Brungart *et al.* (2006), we set *LC* to −6 dB in IBM construction. Intrigued by the observation of Anzalone *et al.* (2006) that HI listeners derive little benefit from IBM in the HF range, we conduct an experiment to test whether ideal masking in the HF range is indeed not important for HI subjects. Unlike Anzalone *et al.* (2006) who applied a constant gain to compensate for the hearing loss of their HI subjects, we apply gain prescriptions to fit individual HI listeners.

In what follows, Sec. II details IBM processing. Section III describes an experiment that tests the effects of ideal masking on mixtures of speech with SSN or cafeteria noise. Section IV describes an experiment that compares the effects of ideal masking in LF, HF, and all-frequency (AF) ranges. Further discussion is given in Sec. V. Finally, Sec. VI concludes the paper.

II. IDEAL BINARY MASKING

The concept of IBM in CASA is directly motivated by the auditory masking phenomenon which, roughly speaking, refers to the perceptual effect that a louder sound renders a weaker sound inaudible within a critical band (Moore, 2003a). So keeping noise in *T-F* units with stronger target energy as done in the standard IBM definition with 0 dB *LC* should not reduce speech intelligibility, and this is indeed what was found by Drullman (1995). On the other hand, IBM processing removes all the *T-F* units with stronger interference energy as the target energy in these units is assumed to be masked by the interference. Removing these masker-dominated units also serves to remove informational masking, which is a dominant factor for reduced speech intelligibility in speech and other modulated maskers (Brungart, 2001). Hence IBM processing, as a form of ideal time-frequency segregation, is expected to yield larger speech intelligibility improvement in a modulated noise condition than in a steady noise condition (Brungart *et al.*, 2006).

Like earlier studies (Brungart *et al.*, 2006; Anzalone *et al.*, 2006), we use a gammatone filterbank to process a stimulus and then time windowing to produce a cochleagram which is a two-dimensional *T-F* presentation (Wang and Brown, 2006). Specifically, we use a 64-channel filterbank that is equally spaced on the equivalent rectangular bandwidth (ERB) rate scale with center frequencies distributed from 2 to 33 ERBs (corresponding to 55–7743 Hz). The bandwidth of each filter is 1 ERB. We note that this filterbank is similar to the one used in Anzalone *et al.* (2006) whereas Brungart *et al.* (2006) used a 128-channel filterbank covering the frequency range of 80–5000 Hz. The response of each filter is divided into 20 ms frames with a frame shift of 10 ms, hence generating a two-dimensional matrix of *T-F* units. The cochleagram of a stimulus is simply the two-dimensional graph of response energy within all the *T-F* units. For a given mixture of target signal and background noise, the IBM is calculated by comparing whether the local SNR within a *T-F* unit is greater than *LC*. As mentioned in Sec. I, we fix $LC=-6$ dB in this study as suggested by Brungart *et al.* (2006). Such a choice of negative *LC* retains certain *T-F* units where the target energy is weaker but not much weaker than the interference energy, in accordance with Drullman's observation that weaker speech energy below the noise level still makes some contribution to speech intelligibility (Drullman, 1995). Indeed, a pilot test with 0 dB *LC* indicates that SRT improvements are not as high as those produced with $LC=-6$ dB. More generally, in order to produce large auditory masking, the masker needs to be stronger than the masked signal (Moore, 2003a).

Given an IBM, the waveform output of IBM can be resynthesized from the mixture input by weighting the mixture cochleagram by the IBM and correcting phase shifts introduced during gammatone filtering (see Wang and Brown, 2006). Such an output can then be played to a listener as a stimulus in our experiments. Figure 1 illustrates IBM for a mixture of a speech utterance and a cafeteria background. The SNR of the mixture is 0 dB. Figure 1(a) shows the cochleagram of the target speech, Fig. 1(b) that of the background noise, and Fig. 1(c) that of the mixture. Figure 1(d) displays the IBM with $LC=-6$ dB, and Fig. 1(e) the cochleagram of the resynthesized result of ideal masking with the IBM in Fig. 1(d). The ideally masked mixture in Fig. 1(d) is clearly more similar to the target speech shown in Fig. 1(a) than the original mixture shown in Fig. 1(c) is. As a comparison, Fig. 1(f) shows the IBM with $LC=0$ dB, and Fig. 1(g) the cochleagram of the corresponding ideal masking output. With the increased *LC*, the IBM has fewer 1's and retains less mixture energy.

III. EXPERIMENT 1: EFFECTS OF IDEAL BINARY MASKING ON SPEECH-RECEPTION THRESHOLD

This experiment was designed to quantify the SRT effects of IBM for both NH and HI listeners. Sentences from the Dantale II corpus (Wagener *et al.*, 2003) were used as target speech, and tests were conducted with two different backgrounds: SSN and cafeteria noise.

A. Methods

1. Stimuli

The Dantale II corpus (Wagener *et al.*, 2003) comprises sentences recorded by a female Danish speaker. Each sentence has five words with a fixed grammar (name, verb, numeral, adjective, and object), for example, "Linda bought three lovely presents" (English translation). Each word in a sentence is randomly chosen from ten equally meaningful words. As a result, recognizing a subset of words in a sentence does not help with the recognition of the remaining words. There are a total of 15 test lists, and each list has ten sentences with no repeating word. There are a few seconds of silence between sentences within each list to allow a listener time to report what has been heard. Similar to the Swedish sentence test (Hagerman, 1982), the closed set corpus was designed for repeated use, and training effects are minimal after familiarization with a few lists (Wagener *et al.*, 2003). We use the speech-shaped noise included with the Dantale II corpus, which is produced by superimposing the speech material in the corpus. The cafeteria noise employed is a recorded conversation in Danish between a male and female speaker that took place in a cafeteria background (Vestergaard, 1998). To emphasize temporal modulation effects, the long-term spectrum of this noise was shaped to match that of the Dantale II speech material (Johannesson, 2006). Target speech and background noises are all digitized at 20 kHz sampling frequency.

A speech utterance and a background noise are first processed separately by a 64-channel gammatone filterbank (see Sec. II), which produces a flat frequency response within the frequency range of the filterbank. Filter responses are then windowed into 20 ms rectangular frames with a 50% overlap between consecutive frames, resulting in a two-dimensional cochleagram. This 100 Hz frame rate is frequently used in speech processing (Rabiner and Juang, 1993). For a given mixture of a Dantale II list and a background noise, the mixture SNR is calculated during the intervals that contain speech energy. To account for the forward masking of the continuously present noise that occurs between two consecu-

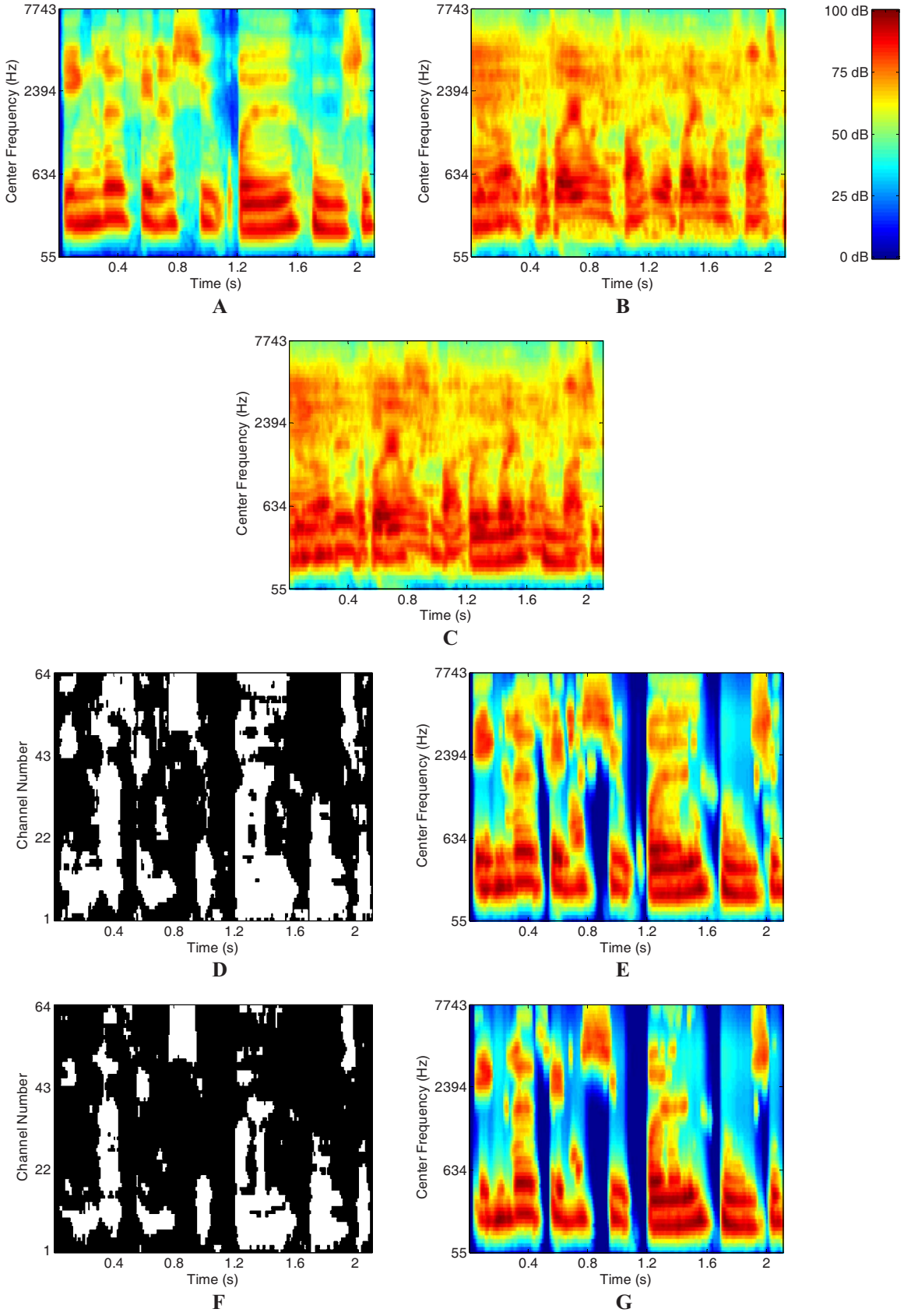

FIG. 1. (Color online) Illustration of IBM (A) Cochleagram of a target speech utterance. (B) Cochleagram of a cafeteria background. (C) Cochleagram of a 0 dB mixture of the speech and the background shown in A and B. (D) IBM with $LC=-6$ dB, where 1 is indicated by white and 0 by black. (E) Cochleagram of the segregated mixture by the IBM in D. (F) IBM with $LC=0$ dB. (G) Cochleagram of the segregated mixture by the IBM in (F).

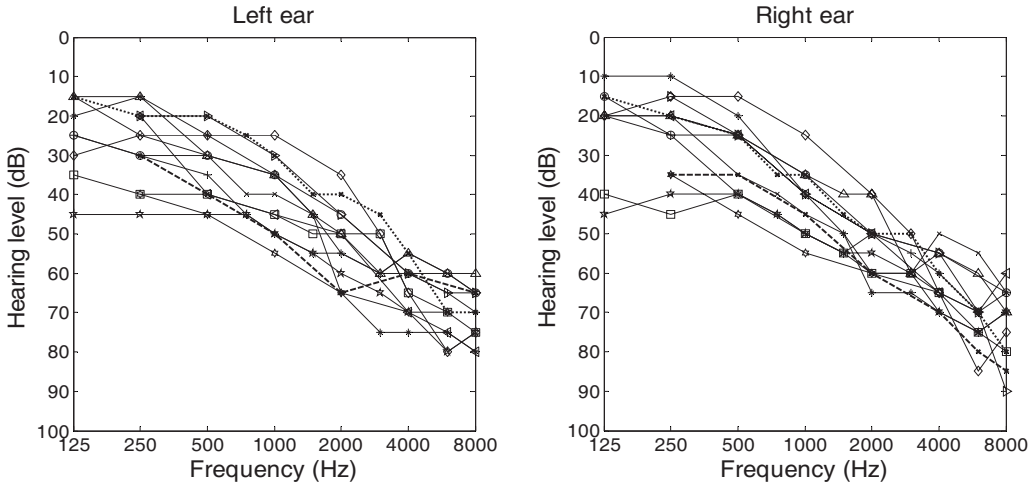

FIG. 2. Audiograms of the 13 HI listeners who participated in the experiments. The dashed line indicates the subject who only participated in Experiment 1, and the dotted line the subject who only participated in Experiment 2.

tive sentences (Moore, 2003a), a period of 100 ms is added before the onset of a sentence for mixture SNR calculation. For a mixture input with a specified SNR, IBM is constructed from the cochleagrams of the target speech and the background noise with *LC* fixed at −6 dB. The IBM is then used to resynthesize a waveform stimulus from the mixture cochleagram. Note that, as a result, the masker signals in between sentences are removed by IBM processing because during such intervals there is only masker energy.

As control conditions, mixtures of speech and background noise were also presented to listeners without segregation. To incorporate filtering effects and any distortions that might be introduced during cochleagram analysis and synthesis, a mixture in an unsegregated condition is processed through an all-1 binary mask or the IBM with the *LC* of negative infinity, therefore including all the *T*-*F* units in the resynthesis.

2. Listeners

A total of 12 NH listeners and a total of 12 listeners with sensorineural hearing loss participated in this experiment. All subjects were native Danish speakers. The NH listeners had hearing thresholds at or below 20 dB HL from 125 Hz to 8 kHz, and their ages ranged from 26 to 51 with the average age of 37. The NH listeners had little prior experience with auditory experiments, and were not informed of the purpose or design of the experiment.

The 12 HI listeners had a symmetric, mild-to-moderate, sloping hearing loss. The audiograms of these listeners are shown in Fig. 2. They had an age range from 33 to 80 with the average age of 67. All the HI listeners were experienced hearing aid wearers. The tests were performed with their hearing aids taken off, and compensation was applied to each HI subject individually. Specifically, a gain prescription was computed from an individual's audiogram using the NAL-RP procedure (Dillon, 2001), and then used to produce amplification with appropriate frequency-dependent shaping. The hearing losses in the left ear and the right ear were compensated for separately. The subjects had participated in Dantale II listening tasks before, but were not told of the purpose and design of this experiment.

3. Procedure

There are a total of four test conditions in this experiment: two ideal masking conditions with SSN and cafeteria noise and two control conditions with unsegregated mixtures. Three Dantale II lists with a total of 30 sentences were randomly selected from the corpus for each test condition. Subjects were instructed to repeat as many words as they could after listening to each stimulus that corresponded to one sentence, and they were not given any feedback as to whether their responses were correct or not. To familiarize them with the test procedure, subjects were given a training session at the beginning of the experiment by listening to and reporting on three lists of clean sentences. The order of the four conditions was randomized but balanced among the listeners (Beck and Zacharov, 2006). A subject test with the four conditions and a training session together took less than 1 h, and a short break was given roughly halfway through the test.

The Dantale II test employs an adaptive procedure in order to find the 50% SRT. The procedure is to present test sentences at SNR that is continuously adapted according to the number of correctly reported words in the previous sentence (Hansen and Ludvigsen, 2001). In a test condition with 30 sentences, the first 10 sentences are used to reach a steady 50% SRT level and the final SRT is determined by averaging the SNR levels for the last 20 sentences.

Speech and noise were both set to the same initial sound pressure level (SPL) for NH listeners. For HI listeners, the initial SPL of speech was set to 5 dB higher than the noise SPL in Experiment 1, and to the same SPL of noise in Experiment 2. In unsegregated conditions, the noise level was fixed while the speech level was adjusted during the adaptive procedure. In ideal masking conditions, as input SNR drops IBM becomes sparser with fewer 1's. To ensure that ideally masked stimuli remain audible at very low SNRs, the speech

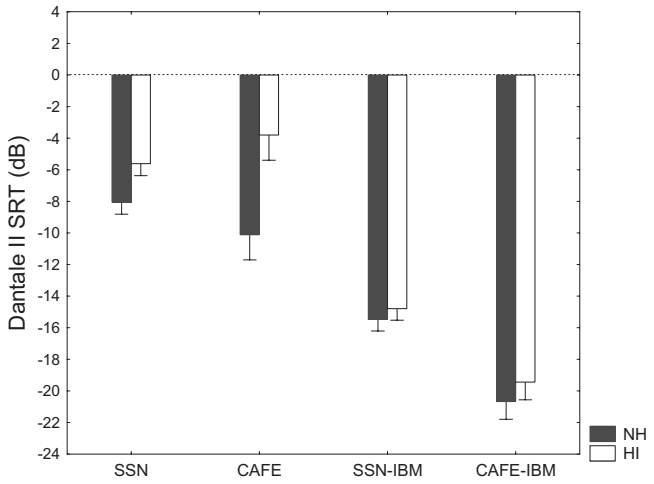

FIG. 3. SRTs for different conditions of Experiment 1 for NH and HI listeners. A more negative SRT corresponds to better performance. Error bars indicate 95% confidence intervals of the means.

level was fixed while the noise level was adjusted in all IBM conditions. As a result, with fewer retained *T-F* units their sound levels became higher even though the levels of the speech signals within these units were unchanged, and the loudness of a processed mixture was thus kept within a small range. This way of adjusting input SNR ensured that the stimuli in all four conditions were comfortably audible.

During a test, a subject was seated in a sound attenuating booth. Test stimuli were generated using the built-in sound card (SoundMAX) in a control computer (IBM ThinkCenter S50) and then presented diotically to a listener through headphones (Sennheiser HD 280 Pro). For HI listeners, an external amplifier (Behring Powerplay HA4000) was used to increase the sound level so that the stimuli within the test range were all audible and yet not uncomfortably loud. The amplification level was adjusted once for each HI listener before the test began.

4. Statistical analysis and power

During the planning phase of the study, the experiment was statistically powered to detect a within-subject between-condition difference of 1.0 dB on mean scores across conditions on the Dantale II test described subsequently for $p<0.05$ at 80% power. This required at least ten complete data sets per condition. Analysis of variance (ANOVA) was performed on all of the data from NH and HI subjects, with within-subject factors of type of processing (IBM or unsegregated) and of type of noise (SSN or cafeteria noise), and a between-subject factor of subject type (NH and HI). *Post hoc* tests were the Bonferroni test and/or the Fisher least-significant-difference (LSD) test, applied where appropriate. The Bonferroni test was used as the most conservative test to indicate differences between means, while the Fisher LSD test was used as the most conservative test for a null result. All statistics were performed using STATISTICA version 7 (StatSoft, 2007).

B. Results and discussion

Figure 3 shows the SRT results of all four test conditions: SSN, CAFE, SSN-IBM, and CAFE-IBM, for both NH and HI listeners. For NH listeners, the mean SRT for unsegregated mixtures with SSN (SSN) is −8.15 dB, for unsegregated mixtures with cafeteria noise (CAFE) is −10.25 dB, for ideal masking with SSN (SSN-IBM) is −15.56 dB, and for ideal masking with cafeteria noise (CAFE-IBM) is −20.70 dB. The ANOVA for NH subjects showed that the main effects of processing type and noise type were significant [$F(1,11)=606.1$, $p<0.001$, and $F(1,11)=78.1$, $p<0.001$, respectively], and there was also a significant interaction between processing type and noise type [$F(1,11)=32.3$, $p<0.001$]. The Bonferroni *post hoc* tests indicated that all means were significantly different [$p<0.005$] from one another. The results show that ideal masking leads to lower (better) SRT compared to unsegregated mixtures regardless of background noise, that the cafeteria background yields a lower SRT than the SSN, and that ideal masking has a greater effect on the cafeteria background. The SRT for the unsegregated SSN condition is comparable to the reference level of −8.43 dB for the Dantale II task (Wagener *et al.*, 2003). The lower SRT for the cafeteria background is consistent with previous studies showing that NH listeners exhibit higher intelligibility in fluctuating backgrounds (Festen and Plomp, 1990; Peters *et al.*, 1998).

For the SSN background, IBM produces an average SRT improvement of 7.4 dB. This level of improvement is consistent with what was found by Anzalone *et al.* (2006) using the HINT test (Nilsson *et al.*, 1994), but higher than the 5 dB improvement reported by Brungart *et al.* (2006) using the CRM task (Bolia *et al.*, 2000). The main difference between our experiment and Brungart *et al.* (2006) lies in different *LC* values: their test uses 0 dB *LC* whereas *LC* is set to −6 dB in our study. As reported in Brungart *et al.* (2006) the choice of $LC=-6$ dB seems better than $LC=0$ dB in terms of speech intelligibility (see also Sec. II).

For the cafeteria background, ideal masking lowers SRT by 10.5 dB on average, which represents a larger gain than for the SSN background. Unlike SSN, the cafeteria background contains significant spectral and temporal modulations which contribute to better intelligibility in the unsegregated condition. We stress that the larger SRT improvement for this background is achieved on top of the better performance of listening to unsegregated mixtures.

For HI listeners, the mean SRTs are −5.61, −3.80, −14.79, and −19.44 dB for the SSN, CAFE, SSN-IBM, and SSN-CAFE conditions, respectively. The ANOVA where both NH and HI subjects were included showed that the main effects of subject type, processing type, and noise type were significant [$F(1,22)=17.2$, $p<0.001$; $F(1,22)=1959.0$, $p<0.001$; and $F(1,22)=100.6$, $p<0.001$, respectively], and there were also significant interaction effects between-subject type and processing type, subject type and noise type, and processing type and noise type [$F(1,22)=49.9$, $p<0.001$; $F(1,22)=19.2$, $p<0.001$; and $F(1,22)=163.9$, $p<0.001$ respectively], as well as a three-way interaction between subject type, processing type, and noise type [$F(1,11)=19.7$, $p<0.001$]. The Bonferroni as well as the Fisher LSD *post hoc* tests on the three-way interaction indicated that all means were significantly different ($p<0.006$)

except for the SSN-IBM and CAFE-IBM conditions where the differences between NH and HI listeners were insignificant ($p > 0.05$). The *post hoc* results show that ideal masking produces lower SRT compared to unsegregated mixtures regardless of noise type, and has a greater effect for the cafeteria background. No difference, however, was revealed between the NH subjects and the HI subjects in the two IBM conditions by either the more conservative Bonferroni test or the less conservative Fisher LSD test. The elevated levels of SRT in the two unsegregated conditions show that HI listeners perform worse in speech recognition in noisy environments, and the levels of SRT increment, 2.5 dB for the SSN condition and 6.5 dB for the CAFE condition, are broadly compatible with previous findings of HI listeners' increased difficulty in speech understanding in noise (see Sec. I). IBM lowers SRT substantially. The SRT gain resulting from ideal masking is 9.2 dB for the SSN background, and this level of improvement is compatible with that reported in Anzalone *et al.* (2006). For the cafeteria background, ideal masking produces a very large SRT improvement of 15.6 dB.

By comparing NH and HI results in Fig. 3, it is clear that HI listeners benefit from ideal masking even more than NH listeners, particularly for the cafeteria background. The results suggest that, after IBM processing, the intelligibility performance is comparable for HI and NH listeners in both SSN and cafeteria backgrounds. It is remarkable that the speech intelligibility of HI listeners becomes statistically indistinguishable from that of NH listeners after ideal masking.

IV. EXPERIMENT 2: EFFECTS OF BAND-LIMITED IDEAL BINARY MASKING ON SPEECH-RECEPTION THRESHOLD

The results of Experiment 1 show large SRT improvements resulting from IBM processing. As mentioned in Sec. I, a main finding reported by Anzalone *et al.* (2006) is that, while NH listeners benefit from IBM in both the LF and HF ranges, HI listeners benefit from ideal masking only in the LF range. This finding is significant because it suggests that, to alleviate the hearing loss of HI listeners, one need not worry about performing *T-F* masking in the HF range; speech segregation at HFs tends to be more difficult than at LFs (Wang and Brown, 2006). Although their interpretation based on the upward spread of masking is reasonable, the fact that they apply constant amplification with no spectral shaping to compensate for the sloping hearing loss of their subjects may suggest a simpler interpretation: the lack of the IBM benefit in the HF range may be partially accounted for by the potentially less compensated hearing loss at HFs. Experiment 2 was primarily designed to assess the importance of IBM processing at HFs for HI listeners as compared to NH listeners. In this experiment, we compensated for the hearing loss of individual listeners based on their audiograms. We compare the intelligibility performance in three setups: IBM in the LF range only, ideal masking in the HF range only, and ideal masking in the AF range. Both SSN and cafeteria backgrounds are used. Consequently, there are a total of six test conditions in this experiment.

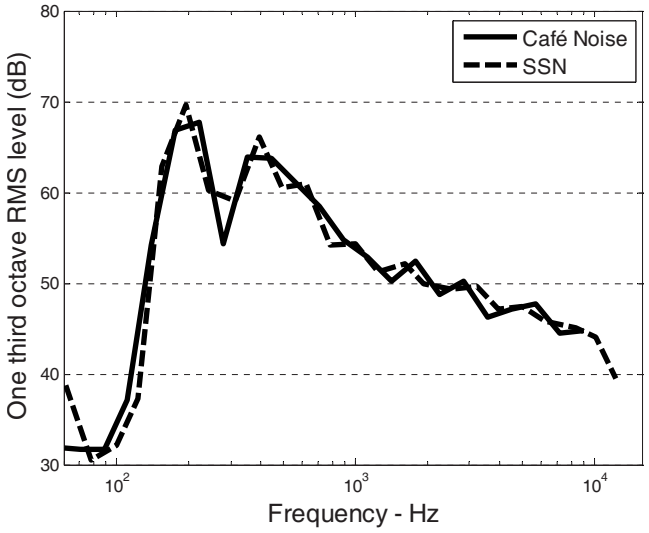

FIG. 4. Long-term spectrum of the SSN in Dantale II (redrawn from Wagener *et al.*, 2003). The spectrum is expressed as root mean square levels in one-third octave bands. Also shown is the long-term spectrum of the cafeteria noise.

A. Methods

1. Stimuli

As in Experiment 1, Dantale II sentences were used as target speech, and SSN and cafeteria noise were used as two different backgrounds. The IBM processing in the AF condition was the same as in Experiment 1. For the LF condition, the same IBM processing as in Experiment 1 was used in the lower 32 frequency channels while an all-1 mask was applied to the higher 32 frequency channels. This way of processing produces no segregation in the HF range. In the HF condition, the reverse was done: IBM was applied to the higher 32 channels while an all-1 mask was applied to the lower 32 channels (hence no segregation in the LF range). This equal division of the 64-channel gammatone filterbank yields a frequency separation boundary approximately at 1.35 kHz. Note that this boundary separating LF and HF ranges is a little lower than the 1.5 kHz boundary used in Anzalone *et al.* (2006). Our choice was partly motivated by the consideration that both the speech material and the SSN in the Dantale II corpus have energy distribution heavily tilted toward LFs so that IBM processing below 1 kHz likely removes significantly more noise than IBM processing above 1 kHz. The long-term spectrum of the SSN (Wagener *et al.*, 2003) is shown in Fig. 4, along with the long-term spectrum of the cafeteria noise. With the 1.5 kHz boundary, the NH results from Anzalone *et al.* (2006) show that the SRT in their LF condition is a little lower than the SRT in their HF condition.

2. Listeners

12 NH listeners and 12 HI listeners participated in this experiment. The pool of NH listeners was the same as that participated in Experiment 1 except for one. This substitution lowered the average age from 37 to 36 without altering the age range. The pool of HI listeners also remained the same as in Experiment 1 except for one. This substitution (see Fig.

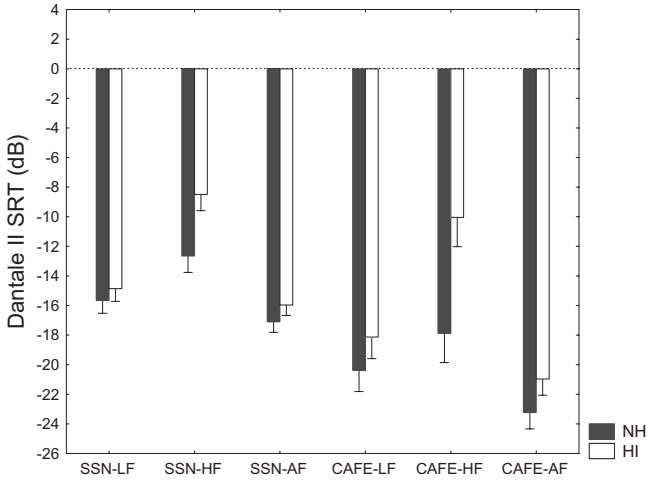

FIG. 5. SRTs for different conditions of Experiment 2 for NH and HI listeners. Error bars indicate 95% confidence intervals of the means.

2), plus a listener whose birthday occurred between the two experiments, changed the average age from 67 to 66 without changing the age range. Again, subjects were naïve to the purpose and design of the experiment. NH listeners were familiar with the Dantale II sentences by virtue of participating in Experiment 1, and as noted in Sec. III A 2, HI listeners had experience listening to Dantale II sentences prior to Experiment 1. Due to the limited number of test lists (15) available in the Dantale II corpus, the same lists used in Experiment 1 were also employed in Experiment 2. It is worth mentioning that the corpus was designed for repeated usage (Wagener *et al.*, 2003; see also Sec. III A 1).

3. Procedure and statistical analysis

The procedure of this experiment is the same as in Experiment 1 except for the following. To vary the input SNR, the noise level was adjusted while the speech level was fixed as in the ideal masking conditions of Experiment 1. In the LF and HF conditions, there is no segregation in half of the frequency channels. As the input SNR decreases in the negative range, the sound level of a stimulus in these conditions is dominated by the background noise in the unsegregated frequency range and hence becomes increasingly louder. To ensure that LF and HF stimuli are not too loud for NH listeners who have very low SRTs, the speech level was fixed at a lower volume than in Experiment 1. Despite this change of sound level, all test stimuli were still comfortably audible for NH listeners. Note that this change did not impact HI listeners as the amount of amplification was individually set for them. ANOVA was performed similarly on all the data from NH and HI subjects as in Experiment 1, with within-subject factors of type of processing (LF, HF, or AF) and of type of noise (SSN or CAFE), and a between-subject factor of subject type (NH or HI).

B. Results and discussion

Figure 5 shows the SRT results of all six test conditions in Experiment 2: SSN-LF, SSN-HF, SSN-AF, CAFE-LF, CAFE-HF, and CAFE-AF, for both NH and HI listeners. The ANOVA for NH subjects showed that the main effects of processing type and noise type were significant [$F(2,22)=255.5$, $p<0.001$, and $F(1,11)=231.2$, $p<0.001$, respectively], and there was also a significant interaction between processing type and noise type [$F(2,22)=4.4$, $p<0.05$]. The Bonferroni tests indicated that all NH means were significantly different ($p<0.006$) from one another, except between the SSN-AF and the CAFE-HF condition. For the SSN background, the mean SRT is −15.66 dB in the LF condition, −12.65 dB in the HF condition, and −17.10 dB in the AF condition. The results show that NH listeners perform better when IBM processing is applied in the LF range than in the HF range, and the difference in SRT is approximately 3 dB. This SRT difference is larger than the SRT difference of slightly more than 1 dB reported by Anzalone *et al.* (2006), despite the fact that the boundary separating LFs and HFs is 1.35 kHz in our processing and 1.5 kHz in their processing. Even with the lower frequency boundary we find that, with the same input SNR, the HF condition leaves more noise than the LF condition since the noise energy is distributed mostly in the LF range (see Fig. 4). The discrepancy is likely due to different ways of IBM processing used in the two studies. The AF condition yields the lowest SRT, which is about 1.6 dB lower than in the LF condition.

For the cafeteria background, the mean SRT is −20.37 dB in the LF condition, −17.88 dB in the HF condition, and −23.24 dB in the AF condition. Clearly NH subjects perform better in this background than in SSN, consistent with the results of Experiment 1. Again, NH listeners benefit more from IBM processing at LFs than at HFs and the relative benefit is 2.5 dB. The AF condition also gives the lowest SRT, which is about 2.9 dB lower than in the LF condition. That NH subjects performed better in the AF condition than in the LF condition for both the SSN and cafeteria backgrounds suggest that they do benefit from IBM in the HF range, even though the benefit is not as high as from the LF range.

The ANOVA where both HI and NH subjects were included showed that the main effects of subject type, processing type, and noise type were significant [$F(1,22)=19.1$, $p<0.001$; $F(2,44)=255.4$, $p<0.001$; and $F(1,22)=317.2$, $p<0.001$, respectively], and there were also significant interaction effects between subject type and processing type, subject type and noise type, and processing type and noise type [$F(2,44)=31.2$, $p<0.001$; $F(1,22)=18.3$, $p<0.001$; and $F(2,44)=14.2$, $p<0.001$, respectively], as well as a three-way interaction between subject type, processing type, and noise type [$F(2,44)=5.4$, $p<0.01$]. Table I shows the Fisher LSD *post hoc* tests. As seen in the table, all conditions were significantly different ($p<0.05$) from one another within the NH subjects (conditions {1}–{6} contrasted against each other) and within the HI subjects (conditions {7}–{12} contrasted against each other). However, the differences between NH and HI were insignificant for the conditions of SSN-LF and SSN-AF.

For HI listeners, the mean SRTs for the SSN background are −14.85, −8.49, and −15.96 dB for the LF, HF, and AF conditions, respectively. The SRT advantage of the LF condition over the HF condition is 6.4 dB, whereas the advantage of the AF condition over the LF condition is only

TABLE I. Fisher LSD *post hoc* significance tests for the three-way interaction of subject type, processing type, and noise type. Significance levels above $p > 0.05$ are given in boldface.

Subject type	Processing type	Test condition	{1}	{2}	{3}	{4}	{5}	{6}	{7}	{8}	{9}	{10}	{11}
NH	SSN-LF	{1} −15.66											
	SSN-HF	{2} −12.65	0.00										
	SSN-AF	{3} −17.10	0.00	0.00									
	CAFE-LF	{4} −20.37	0.00	0.00	0.00								
	CAFE-HF	{5} −17.88	0.00	0.00	**0.07**	0.00							
	CAFE-AF	{6} −23.24	0.00	0.00	0.00	0.00	0.00						
HI	SSN-LF	{7} −14.85	**0.45**	0.05	0.04	0.00	0.01	0.00					
	SSN-HF	{8} −8.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00				
	SSN-AF	{9} −15.96	**0.77**	0.00	**0.29**	0.00	**0.08**	0.00	0.01	0.00			
	CAFE-LF	{10} −18.13	0.03	0.00	**0.34**	0.04	**0.81**	0.00	0.00	0.00	0.00		
	CAFE-HF	{11} −10.05	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	CAFE-AF	{12} −20.96	0.00	0.00	0.00	**0.58**	0.01	0.04	0.00	0.00	0.00	0.00	0.00

1.1 dB. These data are generally comparable with those in Anzalone *et al.* (2006). The results suggest that HI listeners derive considerably more benefit from ideal masking at LFs than at HFs, and the SRT difference is much larger than for NH listeners (see Fig. 5). Although part of the larger gap may be caused by a larger SRT gain (9.2 dB) in HI listeners than that (7.4 dB) in NH listeners due to IBM processing, the fact that the relative advantage of the AF condition over the LF condition for HI listeners is even a little smaller than for NH listeners (1.1 dB versus 1.6 dB) strongly indicates that IBM processing in LF is to a greater extent responsible for the SRT improvement of ideal masking in HI listeners than in NH listeners. In other words, almost all the benefit of IBM can be obtained by IBM only in the LF range. This, of course, is not to say that ideal masking in HF does not improve speech intelligibility compared to no segregation. As illustrated in Fig. 3, IBM processing at all frequencies results in a 9.2 dB SRT improvement compared to no segregation, and the AF condition produces a 7.5 dB relative advantage over the HF condition. This comparison suggests that ideal masking at HFs produces some improvement in speech intelligibility.

For the cafeteria background, the SRTs in the LF, HF, and AF conditions are −18.13, −10.05, and −20.96 dB, respectively (see Fig. 5). The SRT advantage of LF processing over HF processing is 8.1 dB and that of AF over LF is 2.8 dB. These results show a similar pattern as for the SSN background, even though the SRT difference of 2.8 dB between the LF and AF conditions clearly reaches statistical significance (see Table I), and HF processing yields a significant SRT improvement over no segregation as suggested by comparing with the data in Experiment 1. The use of the fluctuating cafeteria background reinforces the conclusion that ideal masking in LF produces a much stronger benefit than that in HF, and this effect is greater in HI listeners than in NH listeners.

The two AF conditions for the SSN and cafeteria backgrounds are the same as the corresponding ideal masking conditions in Experiment 1. The NH performances in Experiment 2 are somewhat better than in Experiment 1. A comparison between Fig. 5 and Fig. 3 shows that the discrepancies are 1.5 dB for SSN and 2.5 dB for cafeteria noise. The only difference in stimuli is the sound level; as pointed out in Sec. IV A 3, the sound level is softer in Experiment 2 than in Experiment 1. For example, at the input SNR of −10 dB, the sound level in Experiment 1 is about 63 dB(A) SPL for the SSN background and 75 dB(A) for the cafeteria background, while the corresponding levels in Experiment 2 are 51 and 51 dB(A), respectively. Studies suggest that softer sound can produce better recognition under certain conditions (Hagerman, 1982; Studebaker *et al.*, 1999). To examine whether the sound volume was a factor in the performance differences, we performed a follow-up experiment with the same pool of the NH listeners who participated in Experiment 2. The follow-up experiment was to simply check subjects' percent correct scores at the sound levels used in the two experiments when the input SNR was fixed at one of the SRTs (alternating between subjects) already obtained in the experiments. The cafeteria background noise was used. The scores are 50.6% with the louder level of Experiment 1 and 58.6% with the softer level of Experiment 2. The 8% difference is statistically significant $[t(11)=3.31,\ p<0.01]$, but unlikely large enough to explain the 2.5 dB SRT difference. Perhaps more important is a learning effect. Unlike HI listeners who were experienced with the Dantale II task, NH listeners used in this investigation had little prior experience with auditory experiments before participating in Experiment 1. When they participated in the second experiment, the familiarity with the Dantale II task acquired during Experiment 1 likely contributed to their better performance. In the predecessor to Dantale II—the Hageman sentence test—Hagerman and Kinnefors (1995) found a training effect of about 0.07 dB per ten sentences, which may explain the differences between Experiments 1 and 2. This interpretation is consistent with the observation that the corresponding performance differences between Experiment 1 and Experiment 2 are smaller for HI listeners; one-third of the mean performance differences is accounted for by the replacement of one HI listener from Experiment 1 to Experiment 2 (see Sec. IV A 2).

V. GENERAL DISCUSSION

The robustness of speech recognition in noise by NH listeners is commonly attributed to the perceptual process of glimpsing, or "listening in the dips," which detects and gathers *T-F* regions of a sound mixture where target speech is relatively stronger compared to interference (Miller and Licklider, 1950; Howard-Jones and Rosen, 1993; Assmann and Summerfield, 2004; Li and Loizou, 2007). As glimpsing involves grouping, this account is closely related to the ASA account that applies to both speech and nonspeech signals (Bregman, 1990). Poorer performance of listeners with hearing loss in fluctuating backgrounds is generally explained as their inability to take advantage of temporal and spectral dips, perhaps caused by reduced frequency selectivity and temporal resolution (Moore, 2007). IBM could be understood as producing glimpses or performing ASA for the listener. The fact that ideal masking also improves intelligibility of NH listeners suggests that even listeners without hearing loss can fail to make full use of the speech information available in a noisy input. The less-than-ideal performance in noisy environments is probably caused by the failure in detecting a glimpse—a *T-F* region with relatively strong target energy—or grouping detected glimpses. This failure becomes more acute with hearing loss. Because ideal masking does an "ideal" job of glimpsing for the auditory system, it helps to nearly equalize the performances of HI and NH listeners (see Fig. 3).

The results of Experiment 1 demonstrate that listeners with or without hearing loss benefit more from IBM processing in the cafeteria background than in the SSN background. The cafeteria background has temporal and spectral modulations, and as a result the amount of informational masking caused by target-masker similarity is expected to be higher than that in SSN. Indeed, some listeners voluntarily commented after the experiment that the conversation in the background distracted their attention, making it harder to concentrate on target utterances. The larger SRT improvement observed for the cafeteria background is thus consistent with the interpretation that ideal masking removes or largely attenuates informational masking (Brungart *et al.*, 2006). In a situation extremely conducive to informational masking, namely, the mixtures of speech utterances of the same talker, Brungart *et al.* (2006) found that the effect of ideal masking is tantamount to a 22–25 dB improvement in input SNR. The 10.5 dB SRT improvement obtained through ideal masking in the cafeteria background, although greater than that obtained in the SSN background, is much smaller than that obtained in mixtures of same-talker utterances. The improvement is also smaller than those obtained in mixtures of different-talker utterances (Chang, 2004), although the gap is not quite as big as in same-talker mixtures. One can therefore expect even larger SRT improvements when interference is one or several competing talkers, a kind of background that produces very large performance gaps between NH and HI listeners as reviewed in Sec. I.

The results of Experiment 2 are on the whole consistent with the related findings of Anzalone *et al.* (2006) even though we used individual gain prescriptions to compensate for listeners' hearing loss. The results are also qualitatively consistent with the findings of Li and Loizou (2007) illustrating that glimpses in the LF to mid-frequency range are more beneficial for speech intelligibility than those in the HF range. However, a few differences between our results and the results of Anzalone *et al.* (2006) are worth noting. First, although considerably smaller than LF processing, there is a benefit from ideal masking in the HF range for HI listeners in our study whereas their study did not show a significant benefit. A possible reason is the individual gain prescription employed in our study that makes segregated speech relatively louder in the HF range than the constant gain applied in their study. Second, we find a relatively greater LF benefit in NH listeners than in their study. The main reason, we believe, is that LF processing removes more background noise than HF processing for a given input SNR. With negative input SNRs (see Fig. 5), the residual noise in the HF condition is in the LF range while that in the LF condition is in the HF range, and the background noises used in our experiments have energy distributed mostly in the LF range, as shown in Fig. 4. This explanation, not considered by Anzalone *et al.*, gives a partial account for the larger LF benefit for listeners with hearing loss. The large SRT gap between LF and HF processing for HI listeners (see Fig. 5), however, cannot be fully explained this way as the gap is substantially larger—to the extent that the SRT performance in LF processing is almost the same as in AF processing. Another likely reason is upward spread of masking (Anzalone *et al.*, 2006) which listeners with sensorineural hearing loss are especially susceptible to (Jerger *et al.*, 1960; Gagne, 1988; Klein *et al.*, 1990). Upward spread of masking is a more prominent factor in the HF condition because of no segregation in the LF range. Also, with more hearing loss at HFs (see Fig. 2), HI listeners are less able to utilize audible HF speech information in recognition compared to NH listeners (Dubno *et al.*, 1989; Ching *et al.*, 1998; Hogan and Turner, 1998). This could also contribute to a steeper performance decline of HF processing relative to AF processing for HI listeners than for NH listeners.

Despite different definitions of IBM, the SRT improvements observed in our study and in Anzalone *et al.* (2006) are very close for the SSN background. It is all the more remarkable considering that their IBM is generated on a sample-by-sample basis while ours is generated on a frame-by-frame basis, which has a drastically lower temporal resolution, and that, in their experiments, IBM-determined gains take the values of 1 and 0.2 while the gains take the values of 1 and 0 in our experiments. The use of two-valued gains is a key similarity between the studies. The most important difference is, of course, that our definition is based on a comparison between target and interference energy and theirs is between target energy and a fixed threshold. The local SNR based IBM is arguably easier to estimate computationally, as many speech segregation algorithms compute binary time-frequency masks by exploring local SNR explicitly or implicitly (Divenyi, 2005; Wang and Brown, 2006). Also, there is little basis in a noisy signal to identify those *T-F* regions of significant target energy where interference is much stronger.

The results from our experiments have major implications for CASA and speech enhancement research aiming to improve speech intelligibility in noisy environments. In addition to affirming the general effectiveness of IBM as a computational goal, our data provide direct evidence that a choice of *LC* at −6 dB for IBM construction, first suggested by Brungart *et al.* (2006), is effective for improving human speech recognition. A comparison between the data of Brungart *et al.* (2006) and ours for the SSN background indicates that the IBM with −6 dB *LC* yields larger SRT improvement than commonly used 0 dB *LC*. Compared to 0 dB *LC*, the choice of −6 dB *LC* retains those *T-F* units where local SNR falls between 0 and −6 dB in ideal masking (see Fig. 1). From the standpoint of SNR, such inclusion will lower the overall SNR of the segregated signal. In other words, if the objective is to improve the SNR of the output signal, the choice of −6 dB *LC* is a poorer one compared to that of 0 dB *LC*. This discussion casts further doubt on the suitability of traditional SNR as a performance metric to evaluate sound separation systems, and at the same time, could shed light on why monaural speech enhancement algorithms often improve SNR but not speech intelligibility (see Sec. I). Another strong implication of our results (see also Anzalone *et al.*, 2006) is that performing speech separation in the LF range is a great deal more important than in the HF range, particularly for improving speech intelligibility of HI listeners.

Our results point to a very promising direction for hearing aid design to improve speech intelligibility in noise of listeners with hearing loss, that is, by designing hearing aids that function in similar ways to IBM. IBM processing improves SRT by a large margin, and HI listeners derive larger benefit than NH listeners. Equally important, the profile of improvement with respect to different kinds of background noise seems to match that of typical hearing impairment. We consider it a highly significant result that ideal masking almost equalizes the intelligibility performances of HI and NH listeners (see Fig. 3). Of course, facing a noisy input IBM cannot be directly constructed and algorithms must be developed to estimate IBM. Encouraging effort has been made in CASA with the explicit goal of IBM estimation (Wang and Brown, 2006), and in limited conditions high-quality estimates are obtainable (see, e.g., Roman *et al.*, 2003). However, computing binary masks close to the IBM in unconstrained acoustic environments remains a major challenge. On the other hand, the extent of intelligibility gain for HI listeners produced by IBM processing much more than fills the SRT gap from NH listeners; Experiment 1 shows a gap of 2.5 dB for the SSN background and a gap of 6.5 dB for the cafeteria background while the ideal masking improvements for HI listeners are 9.2 and 13.8 dB for the two backgrounds, respectively. Hence, perfect IBM estimation is not necessary to bring the performance of HI listeners to the same level as that of NH listeners.

VI. CONCLUSION

The present study was designed to evaluate the impact of IBM on speech intelligibility in noisy backgrounds for both NH and HI listeners. Two experiments were conducted and the main results are summarized below.

- For NH listeners, IBM processing resulted in 7.4 dB SRT reduction for SSN and 10.5 dB reduction for cafeteria noise.
- For HI listeners, IBM processing resulted in 9.2 dB SRT reduction for SSN and 15.6 dB reduction for cafeteria noise.
- After IBM processing, the intelligibility performances for HI listeners and NH listeners were comparable.
- For NH listeners, IBM processing at LFs produced greater SRT reduction than at HFs. The differences were 3 dB for SSN and 2.5 dB for cafeteria noise.
- For HI listeners, IBM processing at LFs produced greater SRT reduction than at HFs. The differences were 5.5 dB for SSN and almost 8 dB for cafeteria noise.

ACKNOWLEDGMENTS

We thank the Associate Editor Ken Grant, and two anonymous reviewers for their helpful comments. The work was conducted while D.W. was a visiting scholar at Oticon A/S. The authors are grateful to M. Schlaikjer, L. Bramsløw, and M. Hartvig, for their assistance in the experiments, and Y. Li for his assistance in figure preparation. D.W. was supported in part by an AFOSR grant (F49620-04-01-0027) and an NSF grant (IIS-0534707).

Alcantara, J. I., Dooley, G., Blamey, P., and Seligman, P. (**1994**). "Preliminary evaluation of a formant enhancement algorithm on the perception of speech in noise for normally hearing listeners," Audiology **33**, 15–27.

Alcantara, J. I., Moore, B. C. J., Kuhnel, V., and Launer, S. (**2003**). "Evaluation of the noise reduction system in a commercial digital hearing aid," Int. J. Audiol. **42**, 34–42.

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.

Assmann, P., and Summerfield, A. Q. (**2004**). "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer, New York) pp. 231–308.

Baer, T., Moore, B. C. J., and Gatehouse, S. (**1993**). "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times," J. Rehabil. Res. Dev. **30**, 49–72.

Beck, S., and Zacharov, N. (**2006**). *Perceptual Audio Evaluation: Theory, Method and Application* (Wiley, Chichester, NY).

Benesty, J., Makino, S., and Chen, J., eds. (**2005**). *Speech Enhancement* (Springer, New York).

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).

Brungart, D. S. (**2001**). "Information and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Brungart, D., 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.

Bunnell, H. T. (**1990**). "On enhancement of spectral contrast in speech for hearing-impaired listeners," J. Acoust. Soc. Am. **88**, 2546–2556.

Carhart, R. C., and Tillman, T. W. (**1970**). "Interaction of competing speech signals with hearing losses," Arch. Otolaryngol. **91**, 273–279.

Chang, P. (**2004**). "Exploration of behavioral, physiological, and computational approaches to auditory scene analysis," M.S. thesis, The Ohio State University Department of Computer Science and Engineering, Columbus, OH; http://www.cse.ohio-state.edu/pnl/theses.html (Last viewed September 2008).

Ching, T. Y. C., Dillon, H., and Byrne, D. (**1998**). "Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification," J. Acoust. Soc. Am. **103**, 1128–1140.
Dillon, H. (**2001**). *Hearing Aids* (Thieme, New York).
Divenyi, P., ed. (**2005**). *Speech Separation by Humans and Machines* (Kluwer Academic, Norwell, MA).
Drullman, R. (**1995**). "Speech intelligibility in noise: Relative contribution of speech elements above and below the noise level," J. Acoust. Soc. Am. **98**, 1796–1798.
Dubno, J. R., Dirks, D. D., and Ellison, D. E. (**1989**). "Stop-consonant recognition for normal-hearing listeners and listeners with high-frequency hearing loss. I: The contribution of selected frequency regions," J. Acoust. Soc. Am. **85**, 347–354.
Edwards, B. (**2004**). "Hearing aids and hearing impairment," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer, New York).
Eisenberg, L. S., Dirks, D. D., and Bell, T. S. (**1995**). "Speech recognition in amplitude-modulated noise of listeners with normal and listeners with impaired hearing," J. Speech Hear. Res. **38**, 222–233.
Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.
Gagne, J.-P. (**1988**). "Excess masking among listeners with a sensorineural hearing loss," J. Acoust. Soc. Am. **83**, 2311–2321.
Greenberg, J. E., and Zurek, P. M. (**1992**). "Evaluation of an adaptive beamforming method for hearing aids," J. Acoust. Soc. Am. **91**, 1662–1676.
Hagerman, B. (**1982**). "Sentences for testing speech intelligibility in noise," Scand. Audiol. **11**, 79–87.
Hagerman, B., and Kinnefors, C. (**1995**). "Efficient adaptive methods for measurements of speech reception thresholds in quiet and in noise," Scand. Audiol. **24**, 71–77.
Hansen, M., and Ludvigsen, C. (**2001**). "Dantale II—Danske Hagermann sætninger (Dantale II—Danish Hagermann sentences)," Danish Speech Audiometry Materials (Danske Taleaudiomaterialer), Værløse, Denmark.
Hogan, C. A., and Turner, C. W. (**1998**). "High-frequency audibility: Benefits for hearing-impaired listeners," J. Acoust. Soc. Am. **104**, 432–441.
Howard-Jones, P. A., and Rosen, S. (**1993**). "Uncomudulated glimpsing in 'checkerboard' noise," J. Acoust. Soc. Am. **93**, 2915–2922.
Hygge, S., Ronnberg, J., Larsby, B., and Arlinger, S. (**1992**). "Normal-hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds," J. Speech Hear. Res. **35**, 208–215.
Jerger, J. F., Tillman, T. W., and Peterson, J. L. (**1960**). "Masking by octave bands of noise in normal and impaired ears," J. Acoust. Soc. Am. **32**, 385–390.
Johannesson, R. B. (**2006**). "Output SNR measurement method," Report No. 052-08-04, Oticon Research Centre Eriksholm, Snekkersten, Denmark.
Kates, J. M., and Weiss, M. R. (**1996**). "A comparison of hearing-aid array-processing techniques," J. Acoust. Soc. Am. **99**, 3138–3148.
Klein, A. J., Mills, J. H., and Adkins, W. Y. (**1990**). "Upward spread of masking, hearing loss, and speech recognition in young and elderly listeners," J. Acoust. Soc. Am. **87**, 1266–1271.
Levitt, H. (**2001**). "Noise reduction in hearing aids: A review," J. Rehabil. Res. Dev. **38**, 111–121.
Li, N., and Loizou, P. C. (**2007**). "Factors influencing glimpsing of speech in noise," J. Acoust. Soc. Am. **122**, 1165–1172.
Li, N., and Loizou, P. C. (**2008a**). "Effect of spectral resolution on the intelligibility of ideal binary masked speech," J. Acoust. Soc. Am. **123**, EL59–EL64.
Li, N., and Loizou, P. C. (**2008b**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.
Li, Y., and Wang, D. L. (**2009**). "On the optimality of ideal binary time-frequency masks," Speech Commun. **51**, 230–239.
Lim, J., ed. (**1983**). *Speech Enhancement* (Prentice-Hall, Englewood Cliffs, NJ).
Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.
Moore, B. C. J. (**2003a**). *An Introduction to the Psychology of Hearing*, 5th ed. (Academic, San Diego, CA).
Moore, B. C. J. (**2003b**). "Speech processing for the hearing-impaired: Successes, failures, and implications for speech mechanisms," Speech Commun. **41**, 81–91.
Moore, B. C. J. (**2007**). *Cochlear Hearing Loss*, 2nd ed. (Wiley, Chichester, UK).
Nilsson, M., Soli, S., and Sullivan, J. A. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.
Peters, R. W., Moore, B. C. J., and Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**, 577–587.
Plomp, R. (**1994**). "Noise, amplification, and compression: Considerations of three main issues in hearing aid design," Ear Hear. **15**, 2–12.
Rabiner, L. R., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).
Ricketts, T., and Hornsby, B. W. (**2003**). "Distance and reverberation effects on directional benefit," Ear Hear. **24**, 472–484.
Roman, N., Wang, D. L., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**, 2236–2252.
Schum, D. J. (**2003**). "Noise-reduction circuitry in hearing aids, II: Goals and current strategies," Hear. J. **56**, 32–41.
Simpson, A. M., Moore, B. C. J., and Glasberg, B. R. (**1990**). "Spectral enhancement to improve the intelligibility of speech in noise for hearing-impaired listeners," Acta Oto-Laryngol. **469**, 101–107.
StatSoft, Inc. (**2007**). STATISTICA (data analysis software system), version 7, http://www.statsoft.com (Last viewed February 2008).
Studebaker, G. A., Sherbecoe, R. L., McDaniel, D. M., and Gwaltney, C. A. (**1999**). "Monosyllabic word recognition at higher-than-normal speech and noise levels," J. Acoust. Soc. Am. **105**, 2431–2444.
Vestergaard, M. (**1998**). "The Eriksholm CD 01: Speech signals in various acoustical environments," Report No. 050-08-01, Oticon Research Centre Eriksholm, Snekkersten, Denmark.
Wagener, K., Josvassen, J. L., and Ardenkjær, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. **42**, 10–17.
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 Academic, 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/IEEE, New York).