Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise

Koenraad S. Rhebergen, Niek J. Versfeld, and Wouter. A. Dreschler

Citation: The Journal of the Acoustical Society of America **120**, 3988 (2006);
View online: https://doi.org/10.1121/1.2358008
View Table of Contents: http://asa.scitation.org/toc/jas/120/6
Published by the Acoustical Society of America

Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise

Koenraad S. Rhebergen,[a] Niek J. Versfeld,[b] and Wouter A. Dreschler[c]
Department of Clinical and Experimental Audiology, Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

(Received 24 March 2006; revised 4 August 2006; accepted 5 September 2006)

The extension to the speech intelligibility index (SII; ANSI S3.5-1997 (1997)) proposed by Rhebergen and Versfeld [Rhebergen, K.S., and Versfeld, N.J. (2005). J. Acoust. Soc. Am. **117**(4), 2181–2192] is able to predict for normal-hearing listeners the speech intelligibility in both stationary and fluctuating noise maskers with reasonable accuracy. The extended SII model was validated with speech reception threshold (SRT) data from the literature. However, further validation is required and the present paper describes SRT experiments with nonstationary noise conditions that are critical to the extended model. From these data, it can be concluded that the extended SII model is able to predict the SRTs for the majority of conditions, but that predictions are better when the extended SII model includes a function to account for forward masking. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2358008]

PACS number(s): 43.71.An, 43.66.Ba, 43.71.Gv, 43.72.Kb [MSS] Pages: 3988–3997

I. INTRODUCTION

Speech intelligibility decreases due to the presence of a background noise. Parts of the speech signal then are masked by the noise such that not all speech information is available to the listener. French and Steinberg (1947), Fletcher and Galt (1950), and later Kryter (1962a, 1962b) developed a calculation method, known as the articulation index (AI), to predict the speech intelligibility under such masking conditions. The AI calculation scheme was re-examined in the 1980s and early 1990s, which led to a new method accepted as the ANSI S3.5-1997 (1997). Since its revision in 1997, the AI is named the speech intelligibility index (SII). A detailed description of the SII can be found in Pavlovic (1987), and the ANSI S3.5-1997 (1997) standard.

To date, the SII model has been designed and validated only for stationary masking noises. In fluctuating masking noises, speech intelligibility is usually much better for normal-hearing listeners, since the listener is able to take advantage of the relatively silent periods in the noise masker; for hearing impaired listeners this is often not the case (Festen and Plomp, 1990; Houtgast *et al.*, 1992; Versfeld and Dreschler, 2002). However, the SII model does not take into account any fluctuation in the masking noise since it uses only the long term speech and noise spectrum. Therefore, it predicts speech intelligibility inaccurately for these conditions. Since many daily-life background noises do fluctuate strongly over time (Koopman *et al.*, 2001), the SII model is unable to predict speech intelligibility in the majority of real-life situations adequately.

Recently, Rhebergen and Versfeld (2005) proposed an extension to the SII model, in order to improve the predictions for speech intelligibility in fluctuating noise. The basic principle of this approach is that both speech and noise signal are partitioned into small time frames. Within each time frame the instantaneous SII is determined, yielding the speech information available to the listener at that time frame. Next, the SII values of these time frames are averaged, resulting in the SII for that particular speech-in-noise condition. With the aid of various data available for a variety of noise types described in the literature, Rhebergen and Versfeld (2005) have shown that their extension allows a good account for most existing data, dealing with the speech reception threshold (SRT) for sentences. However, there still are conditions where the extended SII (ESII) model is unable to give accurate predictions. First, the ESII model is unable to predict SRTs for sentences in 100% sinusoidally intensity-modulated (SIM) speech noise, as measured by Festen (1987). Although the SRT values predicted by the ESII model yield an improvement over the original SII model, there are still some systematic deviations. Festen found lowest SRTs (i.e., best performance) for modulation frequencies of 16 and 32 Hz, whereas the ESII model predicts the best performance for a modulation frequency of 8 Hz.

Second, Rhebergen *et al.* (2005) measured SRTs with unintelligible interfering speech (foreign language) as a masker played normal and time reversed. By reversing the unintelligible speech masker in time, the SRT worsened about 2.3 dB. Rhebergen *et al.* (2005) argued that this difference could be attributed to differences in the amount of forward masking: The time-reversed speech masker (having a "ramped"-like envelope, i.e., a gradual increase with a sudden offset) provokes more forward masking than a normal speech masker (being more "damped"-like, i.e., a sharp onset followed by a gradual declination). A time-asymmetrical nonspeech-like noise masker may provide more insight into the effects of temporal forward masking on speech intelligibility. The ESII model is, in essence, a time-symmetrical model. It does not account for the differences in forward and

[a]Electronic mail: k.s.rhebergen@amc.uva.nl
[b]Electronic mail: n.j.versfeld@amc.uva.nl
[c]Electronic mail: w.a.dreschler@amc.uva.nl

backward masking. The model predicts the same speech intelligibility with a noise masker played normal and time reversed.

Third, the ESII is a model verified with SRT data described in the literature. To enable a fair comparison between the data obtained in different studies, Rhebergen and Versfeld (2005) restricted themselves to the use of SRT data obtained with one set of speech materials, viz., the Dutch speech corpus of Plomp and Mimpen (1979). Even though the corpus is similar, differences between studies sometimes are substantial: some conditions have been measured abundantly (SRT in stationary speech shaped noise), whereas other conditions have been measured sparsely (SRT in SIM noise) (Festen, 1987). Moreover, SRT data have been collected by different researchers in different experimental settings. This introduces additional variance in the data. For example, SRTs in quiet or interrupted noise differ largely between different papers (de Laat and Plomp, 1983; Festen, 1987; Noordhoek, 2000; Duquesnoy, 1983; Plomp and Mimpen 1979a; 1979b). For a good validation of the ESII model, only SRTs obtained from the same group of subjects and measured under the same experimental conditions should be used for ESII calculations. For instance, the ESII model predicts that SRTs in fluctuating noise, unlike stationary noise, may depend on the subject's absolute threshold, even in normal-hearing listeners.

This paper addresses the problems described above, with the aim to test and, where necessary, refine the ESII model. All experiments have been conducted with normal-hearing subjects.

In the first section, SRT tests are performed for nineteen different noise conditions (test and retest) using the speech material of Versfeld *et al.* (2000). The noise conditions comprise steady state noise, interrupted noise with different modulation frequencies and different duty cycles, SIM noise with different modulation frequencies, and two asymmetrically saw-tooth noises. The noise conditions have been selected such to test the ESII model critically.

In the next section, the observed SRTs are used to evaluate and refine the ESII model. Notably, the ESII calculations are extended by using a function to account for forward masking. Last, predictions and limitations of the finally obtained extended ESII model will be discussed.

II. EXPERIMENT

In this experiment, the SRTs for a number of noise conditions are measured. The results will be used to validate the SII model.

A. Subjects

Twelve normal-hearing subjects (one male, 11 females) participated. Their age ranged from 18 to 29 years and was on average 21.5 years. Subjects were native speakers of the Dutch language and had at least high school education. Each subject had pure-tone thresholds of 15 dB HL or better at octave frequencies from 125 to 8000 Hz (ANSI S3.6, 1996). Table I shows the average pure-tone thresholds.

TABLE I. Pure-tone thresholds averaged across the group of 12 normal-hearing subjects.

Frequency (Hz)	125	250	500	1000	2000	4000	8000
Threshold (dB HL)	8.2	5.9	4.1	0.0	1.8	3.6	5.0
Standard deviation (dB)	6.4	7.0	5.4	4.5	4.0	5.5	8.9

B. Stimuli

The target speech material consisted of short every-day sentences, uttered by a female speaker (Versfeld *et al.*, 2000). The speech material comprises 39 lists of 13 sentences and has been developed for a reliable measurement of the speech intelligibility in noise. The speech was stored at a sample rate of 44.1 kHz and a 16 bits resolution.

All 19 interfering noise conditions are given in Table II. The noise conditions comprise one condition with steady state noise, ten conditions with interrupted noise, two conditions with saw-tooth noise, and six conditions with SIM (sinusoidal intensity modulated) noise. Figure 1 illustrates the wave forms of the noise types. All noise conditions had a long-term average spectrum equal to the long-term average spectrum of the target female speech material (Versfeld *et al.*, 2000). The nonstationary noises were derived from the original stationary masking noise of Versfeld *et al.* (2000), where the envelope was modified with the aid of the MATLAB signal processing toolbox. The interrupted noise conditions were modulated with a duty cycle of 50% and a depth of 100%, and the modulation frequencies were 4, 8, 16, 32, 64, and 128 Hz. Four conditions had a modulation frequency of 8 Hz, but with a duty cycle of 40%, 45%, 55%, and 60%. The SIM noises were generated according to Festen (1987). The modulation frequencies were 4, 8, 16, 32, 64, and 128 Hz, and the modulation depth was 100%. The two saw-tooth noise conditions had a modulation frequency of 8 Hz and the envelope was exponentially increasing or decreasing in time (types 1 and 2, respectively), with a slope of about 40 dB/125 ms.

C. Procedure

Subjects were tested individually in a sound-insulated booth. Signals were played out via an Echo soundcard (Gina 24/96) on a personal computer at a sample frequency of 44.1 kHz, and were fed through a TDT Amplifier (MA2) and a TDT Headphone Buffer (PA4). Subjects received the signals monaurally at their best ear via TDH 39P headphones at a fixed noise level of 65 dB A. After the presentation of a sentence, the subject's task was to repeat the sentence he or she had just been presented. A sentence was scored correct if all words in that sentence were repeated without any error. A list of 13 sentences, unknown to the subject, was used to estimate the signal-to-noise ratio (SNR) at which 50% of the sentences was reproduced without any error, the so-called SRT. For a given condition, the first sentence of the list started far below the expected SRT. The sentence was repeated each time at a 4 dB higher level until the subject was able to reproduce it correctly. The twelve other sentences of that list were presented only once, following a simple up-

TABLE II. Schematic representation of the nineteen noise conditions.

Noise condition	Noise type	Modulation frequency	Modulation depth (%)	Duty cycle (%)	Envelope shape
Int 4 Hz	Interrupted	4	100	50	Square
Int 8 Hz	Interrupted	8	100	50	Square
Int 16 Hz	Interrupted	16	100	50	Square
Int 32 Hz	Interrupted	32	100	50	Square
Int 64 Hz	Interrupted	64	100	50	Square
Int 128 Hz	Interrupted	128	100	50	Square
Int 8 Hz dc40%	Interrupted	8	100	40	Square
Int 8 Hz dc45%	Interrupted	8	100	45	Square
Int 8 Hz dc55%	Interrupted	8	100	55	Square
Int 8 Hz dc60%	Interrupted	8	100	60	Square
Saw-tooth *T*1	Saw-tooth	8	-	-	Exponential increasing
Saw-tooth *T*2	Saw-tooth	8	-	-	Exponential decreasing
SIM 4 Hz	Sinusoidal intensity modulated	4	100	-	Sinusoidal
SIM 8 Hz	Sinusoidal intensity modulated	8	100	-	Sinusoidal
SIM 16 Hz	Sinusoidal intensity modulated	16	100	-	Sinusoidal
SIM 32 Hz	Sinusoidal intensity modulated	32	100	-	Sinusoidal
SIM 64 Hz	Sinusoidal intensity modulated	64	100	-	Sinusoidal
SIM 128 Hz	Sinusoidal intensity modulated	128	100	-	Sinusoidal
Steady state	Stationary	-	-	-	Flat

down procedure with a step size of 2 dB. The SRT was estimated according to the procedure described by Plomp and Mimpen (1979a), i.e., by taking the mean SNR of sentence 5–13 plus the SNR that would have been used for the fourteenth sentence. With each sentence presentation, a random sample of the interfering noise was taken. The noise onset was 1200 ms before the onset of the sentence; it stopped 800 ms after the offset of the sentence. Thus, the duration of the noise was in total 2000 ms longer than the sentence duration.

In total, 19 conditions were tested. The experiment was partitioned into two blocks, a test and a retest block. To avoid confounding of measurement condition order and sentence lists, the order of conditions and sentence lists was counterbalanced across subjects. In total, each subject received 38 lists of 13 sentences preceded by three practice lists.

III. RESULTS

A $19[\text{condition}]\times 2[\text{test/retest}]\times 12[\text{subject}]$ analysis of variance (ANOVA) was performed on the data. Of the main effects, "condition" ($F[18,198]=159.08$, $p<0.001$), and "test/retest" was significant ($F[1,11]=14.87$, $p<0.005$). The SRT of the retest was on average 0.8 dB better. Differences between subjects were not significant ($F[11,15.71]=2.14$, $p>0.05$). Of the interactions, conditions*test ($F[18,198]=1.88$, $p<0.05$) and conditions*subjects ($F[198,198]=1.36$, $p<0.05$) were weakly significant. Therefore, in the remainder of the paper, subjects have not been

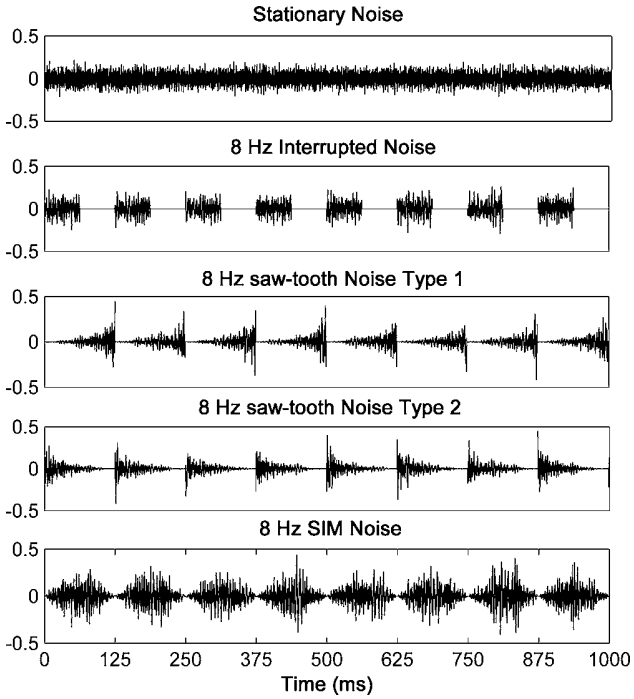

FIG. 1. Illustration of some masking noises used in the present experiment. In this selection all signals have a spectrum equal to the long-term average spectrum of the female target speech (Versfeld *et al.*, 2000) and, with exception of the upper panel, a modulation frequency of 8 Hz. The second panel shows interrupted noise with a duty cycle of 50%; the third panel saw-tooth noise (type 1), the fourth panel saw-tooth noise (type 2, time reversed version of type 1), and at the lower panel a SIM noise.

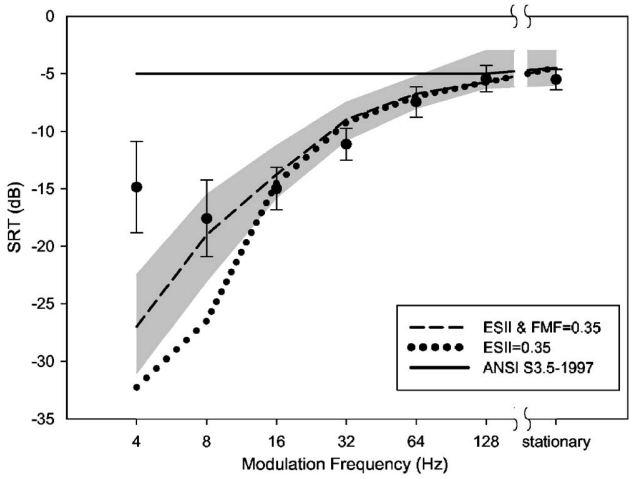

FIG. 2. Speech reception threshold (dB) as function of modulation frequency (Hz) for the steady state noise and the 4–128 Hz interrupted noise conditions. Error bars denote the standard deviations between subjects. The shaded area, the dashed, dotted, and solid lines are model predictions and are explained in Sec. V.

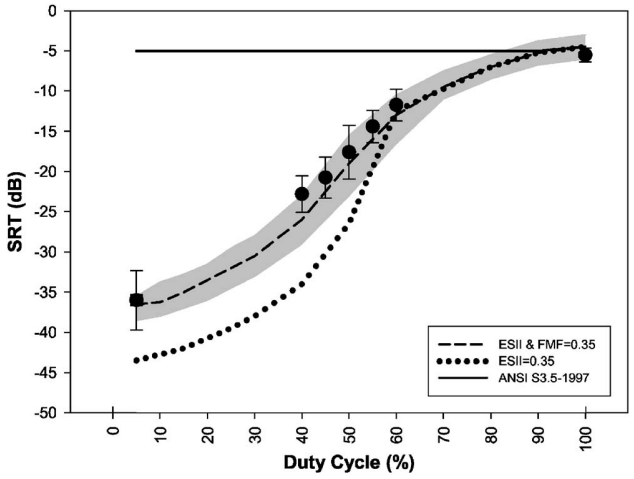

FIG. 3. Speech reception threshold (dB) as function of duty cycle (%) for interrupted noise with a modulation frequency of 8 Hz (circles). Error bars denote the standard deviations between subjects. An additional noise condition (duty cycle 5%, square) will be explained in Sec. IV. The shaded area, the dashed, dotted, and solid lines are model predictions and explained in Sec. V.

entered as a separate factor in the analyses. Because the present paper focuses on model predictions, and not on a possible learning effect, only the results of the retest data have been presented, although test/retest still has been entered as a factor in the data analysis. Below, additional analyses were performed on subsets of the data.

A. Interrupted noise

Figure 2 shows the SRT values (dB) for the retest data, averaged across subjects, as function of modulation frequency (Hz) for interrupted noise with a duty cycle of 50%. Error bars denote the standard deviation between subjects. The shaded area, the dashed, dotted, and solid lines in Fig. 2 are model predictions and are discussed below. A 7[condition]$\times$2[test/retest] ANOVA showed that the main effect of condition was significant ($F[6,1]=123.37$, $p<0.001$). Also, the main effect test/retest was significant ($F[1,1]=8.88$, $p<0.005$). The SRTs of the retest were on average 0.9 dB better. There was no significant interaction. The SRT is lowest with a modulation frequency of 8 Hz (−16.7 dB). This SRT is comparable, but somewhat higher than that obtained by de Laat and Plomp (1983), who found an SRT of −23 dB at a modulation frequency of 10 Hz. The SRT with a 4 Hz interrupted noise is somewhat higher compared to that with an 8 Hz interrupted noise. This may be accounted for by the fact that at these slow modulation rates noise bursts can mask complete words of a sentence. The trend in the present data is consistent with the results of Miller and Licklider (1950), Licklider and Guttman (1957), Gustfsson and Arlinger (1994), Trine (1995), Dubno *et al.* (2002, 2003), and Nelson *et al.* (2003). Bonferroni post hoc tests showed that the 8, 16, 32, and 64 Hz conditions differed significantly from each other. The 64, 128, and stationary noise conditions did not differ significantly.

Figure 3 shows the SRT values (dB) averaged across subjects, for the retest data only, as function of duty cycle (%) for interrupted noise with a modulation frequency of 8 Hz. Error bars denote the standard deviations between subjects. The shaded area, the dashed, dotted, and solid lines in Fig. 3 are model predictions and are discussed below. The duty cycles were 40%, 45%, 50%, 55%, 60%, and 100% (steady state noise). An additional data point, indicated with a filled square, obtained with a duty cycle of 5%, will be discussed below. A 6[condition]$\times$2[test/retest] ANOVA was performed on these data. Of the main effects, difference in condition was significant ($F[5,1]=121.83$, $p<0.001$). Also, the SRT of the retest was on average 1.3 dB better than the test, which was a significant effect ($F[1,1]=9.11$, $p<0.05$). There was no significant interaction between the main effects. Bonferroni post hoc tests showed no significant differences between the 40% and 45% condition, as well as between the 55% and 60% condition. Otherwise, the SRT of all conditions were significantly different. The data show a gradual and almost linear increase in SRT from −21.8 up to −11.6 dB as the duty cycle increases from 40% up to 60%.

B. Saw-tooth noise

For the retest, the mean SRT scores of saw-tooth type 1 and saw-tooth type 2 were −9.3 and −11.9 dB, respectively. A 3[condition]$\times$2[test/retest] ANOVA was performed on the data with the saw-tooth and steady state noise. Of the main effects, differences in condition were significant ($F[2,1]=249.69$, $p<0.001$). The SRT of the retest was on average 0.1 dB better than the test, which was not significant ($F[1,1]=0.368$ $p>0.05$). There was no significant interaction. Bonferroni post hoc tests showed that all three conditions differed significantly from each other.

C. SIM noise

Figure 4 shows the SRT values averaged across subjects, for the retest data, for each of the six conditions for the interfering SIM noises and for the steady state noise. Error bars denote the standard deviations between subjects. A

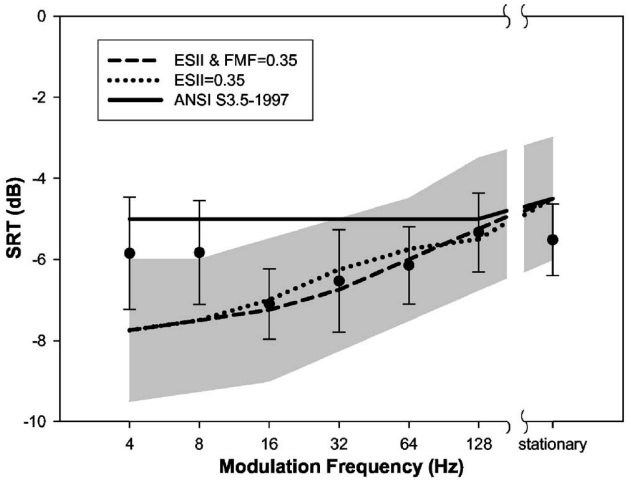

FIG. 4. Speech reception threshold (dB) as function of modulation frequency (Hz) for the conditions with SIM noise. Error bars denote the standard deviations between subjects. The shaded area, the dashed, dotted, and solid lines are model predictions and explained in Sec. V.

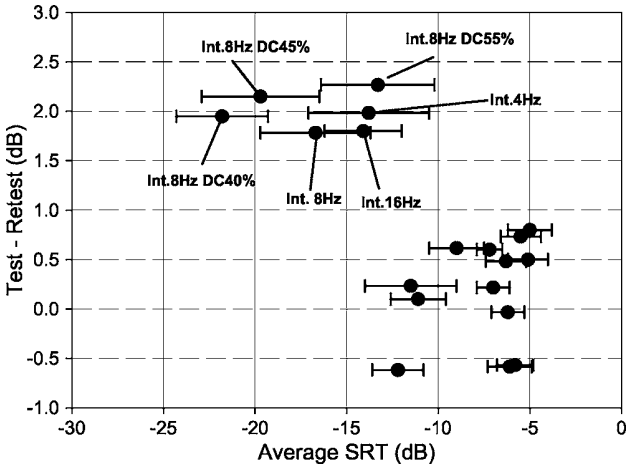

FIG. 5. Test-retest difference (dB) as function of the average SRT for all noise conditions.

7[condition]×2[test/retest] ANOVA showed that of the main effects, only differences in condition were significant ($F[6,1]=8.07$, $p<0.001$). The SRT of the retest was on average 0.1 dB better than the test, which was not significant ($F[1,1]=0.431$, $p>0.05$). There were no significant interactions. Bonferroni post hoc tests showed that the 8, 16, 32, and 64 Hz conditions were not significantly different from each other, and that the 16 Hz condition was significantly different from the 4, 128 Hz condition, and the stationary noise condition. The trends in the present data are not entirely consistent with the results of Festen (1987). He observed best SRTs for the 32 Hz condition, whereas in this study best SRTs were observed for the 16 Hz condition. Furthermore, the SRTs observed by Festen (1987) were 2–3 dB lower than the SRTs observed in this experiment.

IV. DISCUSSION

The results of the present experiment show some interesting aspects.

First of all, the SRTs of the retest were on average 0.8 dB lower than the SRTs of the first test. After separating the SRT data into subgroups, it was clear that this effect was only present in the interrupted noise conditions. A possible explanation may be that listening into the gaps of the interrupted noise requires practice. Hence, the learning effect is expected to be more prominent when the gaps are deeper. In other words: Lower (i.e., better) SRTs are accompanied by larger test-retest differences. Figure 5 displays the average test-retest difference as a function of the mean SRT for all conditions in the present study. The data in Fig. 5 form two subgroups: One subgroup is formed by those conditions where the test-retest difference does not exceed 1 dB, and the mean SRT is higher than about −12 dB. The other subgroup is formed by the conditions with relatively good SRTs of −13 dB or better. Here the test-retest differences are larger than 1.5 dB. The latter group consists of conditions with interrupted noise. It seems that the test-retest difference is related to the gap length and not particularly related to a specific modulation frequency. The lower boundary of the gap length where significant learning effects may occur then is in the order of 50 ms. To what degree the size of the learning effect is related to gap length, and how much practicing is required before thresholds stabilize, is left to future research. It is however important with respect to modeling the data (where one wants to avoid learning as much as possible), as well as with respect to clinical applications with the measurement of the SRT in fluctuating noise conditions (where one wants to reach stable thresholds as quickly as possible).

Second, the two conditions with saw-tooth noise gave a difference of 3.2 dB in SRT. The two conditions are identical, except for that saw-tooth type 2 is the time-reversed counterpart of type 1. It is likely that the difference in SRT is due to differences in forward masking. The temporal envelope of saw-tooth type 2 has a quick onset time (steep slope) and a slow decay (shallow slope), as can be seen in Fig. 1, and produces a plosive-like sound. Such an envelope has resemblance with the envelope of real speech (Rosen, 1992), since the envelope of speech is typically dominated by plosive sounds. The envelope of saw-tooth type 1 is the opposite of type 2: the envelope increases gradually and offsets abruptly. Forward masking is the phenomenon that weak sounds are masked by preceding strong sounds, and is a characteristic of the auditory system that it apparently cannot follow such abrupt offsets accurately. A signal (or speech) that is present directly after the abrupt offset is physically not masked by the saw-tooth noise, but it is masked due to the sluggishness of the auditory system. Hence, a more favorable signal-to-noise ratio is required to result in the same intelligibility as for speech in saw-tooth type 2 noise. Indeed, this is true in the present experiment. It is also true for the data of Rhebergen *et al.* (2005), who described an SRT test with intelligible and unintelligible interfering speech played normal and time-reversed. With Dutch listeners, (unintelligible) Swedish interfering speech gave a rise in SRT of 2.3 dB when played in reverse.

Third, between-subject differences appear to be larger with lower SRT values, as can be seen in Fig. 2. Rhebergen and Versfeld (2005) showed with their extended SII method

that the psychometric function (i.e., SII as a function of SNR) near SII=0.3 is relatively shallow for speech in interrupted noise compared to that for speech in stationary noise. Consequently, a given variation in SII corresponds to a large variation in the SNR with interrupted noise, but to a smaller variation with stationary noise. In the next section (Sec. V), SII calculations will show whether this explanation can fully account for the differences in variance between these conditions.

Fourth, the large difference in SRT (approximately 6 dB) obtained with interrupted noise between the present results and those of de Laat and Plomp (1983) possibly might be explained by differences in absolute threshold, since de Laat and Plomp (1983) found a high correlation between the SRT and the pure-tone average (PTA, averaged across 500, 1000, and 2000 Hz) in their group of subjects. However, the PTA of the present group of subjects is 7.6 dB better (viz., 2.0 dB HL with a standard deviation of 3.6 dB) compared to the average PTA of 9.6 dB HL (with a standard deviation of 3.6 dB) from the subjects of de Laat and Plomp (1983). Moreover, with the present data, a Pearson correlation coefficient was calculated between the individual PTAs and the noise conditions. Correlations were nonsignificant ($r=0.035$, $p>0.05$). The differences in observed SRTs must be due to other factors, such as perhaps the use of different speech corpuses (Plomp and Mimpen, 1979a, versus Versfeld *et al.*, 2000). Although Versfeld *et al.* (2000) found no significant differences between these sets when comparing them in stationary masking noise, van Wijngaarden and Houtgast (2004) did, be it in different listening conditions (such as reverberation). It is known that differences in intelligibility between different speech materials become more apparent under increasingly adverse listening situations (Mullennix *et al.*, 1989). But what properties of the speech signal cause these differences is yet unclear. A similar explanation holds for the differences in SRT with SIM noises obtained by Festen (1987) and the present data.

Next, the increase in SRT with increasing modulation frequency of the interrupted noise (from 8 to 128 Hz) is due to an increase in forward masking. In all conditions, the masker is absent in 50% of the time, but due to a decrease in "gap" duration (62.5 down to 3.9 ms), the forward masking induced by each masker pulse becomes more effective. The SRT with a 128 Hz interrupted noise is even worse than with a steady state noise. In this condition, the gaps are very short in duration, such that they are probably entirely masked. At the same time, the masker pulses must be 3 dB higher in level, in order to have the same long term root-mean-square level as stationary noise. Thus, 128 Hz interrupted noise is a more effective masker than is stationary noise.

The increase in SRT with interrupted noise when changing the modulation frequency from 8 to 4 Hz may be accounted for by the fact that with these slow modulation rates, masking of complete words in a sentence can occur. This phenomenon has also been observed by Miller and Licklider (1950) and Nelson *et al.* (2003), who found optimal performance around modulation rates of 10 and 8 Hz, respectively. Because in the SRT procedure every word of the sentence needs to be repeated correctly, it is unsuitable for these low modulation frequencies (less than 8 Hz).

Finally, a slight change in the duty cycle results in a large change in SRT, cf. Fig. 3. Considering the entire range, a decrease in duty cycle (more pulsed signals with longer gaps) will probably result in increasingly lower thresholds, with the SRT in quiet (absence of a noise masker) as a lower limit. The SRT in quiet is on average about 20 dB A (Duquesnoy, 1983; Duquesnoy and Plomp, 1983; Plomp and Mimpen, 1979b, Noordhoek, 2000), which is, compared to a 65 dB A noise level, equal to an SRT of −45 dB. Additional experiments show that the decrease holds at least to a duty cycle of 5%; resulting in an SRT of −36 dB (see Fig. 3, filled square).

V. SII MODEL PREDICTIONS

A detailed description of the conventional SII model is given in ANSI S3.5-1997 (1997) or Pavlovic (1987), and a detailed description of the ESII model is given in Rhebergen and Versfeld (2005). The basic principle of the conventional SII is that departing from the long term speech spectrum, the long term noise spectrum, and the absolute threshold of hearing, the amount of speech information that exceeds both noise and threshold is calculated. The extension proposed by Rhebergen and Versfeld (2005) is that the long term noise spectrum is replaced by the actual noise signal. Both speech (represented by stationary speech shaped noise) and noise signal are partitioned into small time frames, and within each time frame, the (now instantaneous) conventional SII is calculated, representing the speech information available to the listener at that time frame. The ESII for the condition under investigation is obtained by averaging the instantaneous SII across time. With the ESII, the length of the time frames is frequency dependent, and time constants are adapted from gap detection data (Moore, 1997). The length of a time frame ranges from approximately 35 ms in the lowest frequency band up to 9.4 ms in the highest frequency band. The present paper uses the SPIN 21 critical band weighting function (ANSI S3.5-1997, 1997, Table B.1).

Furthermore, all SII calculations are conducted with the long term speech spectrum of the female target speaker. In order to approach the sound level at the ear drum (as required by the SII model), all signals are filtered with a fifth order finite impulse response filter with the transfer characteristics of the TDH39P headphone that was used in the experiment. Also, if required, the background noise present in the sound proof booth was added to the noise signal. This seems unnecessary, but the SII model defines silence in each band as −50 dB SPL, whereas in a sound proof booth more realistic numbers are between 0 and 10 dB SPL in the mid and high frequencies and 35 to 50 dB SPL in the frequencies below 100 Hz. These levels have almost no effect with most noise types, except for conditions where noises contain relatively long silent periods, such as is the case with interrupted noise.

Since the purpose of the present paper is to evaluate the model, learning effects were eliminated as much as possible

TABLE III. For each condition of the present experiment, the SRTs of the retest and the results of the various SII calculation schemes are given. Lower rows yield the mean and standard deviation of the SII. FMF denotes the use of the forward masking function, asym.w. denotes the use of an asymmetrical integration window, and Lin.w. denotes the use of a fixed integration window of 4 ms.

Noise condition	SRT(dB)	stdv(dB)	Conventional SII	Extended SII (2005)	Extended SII & FMF (asym. w.)	Extended SII & FMF (Lin.w.)
Int 8 Hz	−17.6	3.3	0.000	0.388	0.357	0.359
Int 16 Hz	−15.0	1.8	0.031	0.333	0.291	0.299
Int 32 Hz	−11.1	1.4	0.012	0.271	0.260	0.276
Int 64 Hz	−7.5	1.3	0.239	0.309	0.299	0.292
Int 128 Hz	−5.4	1.1	0.297	0.339	0.299	0.305
Int 8 Hz dc40	−22.8	2.3	0.000	0.465	0.385	0.394
Int 8 Hz dc45	−20.8	2.6	0.000	0.432	0.367	0.376
Int 8 Hz dc55	−14.4	2.0	0.044	0.369	0.354	0.357
Int 8 Hz dc60	−11.7	2.0	0.117	0.365	0.361	0.364
Sawtooth *T*1	−9.3	1.5	0.178	0.389	0.369	0.376
Sawtooth *T*2	−11.9	1.6	0.102	0.310	0.330	0.335
SIM 8 Hz	−5.8	1.3	0.303	0.381	0.384	0.385
SIM 16 Hz	−7.1	0.9	0.260	0.322	0.331	0.335
SIM 32 Hz	−6.5	1.3	0.308	0.318	0.329	0.335
SIM 64 Hz	−6.2	0.9	0.289	0.314	0.315	0.323
SIM 128 Hz	−5.3	1.0	0.319	0.337	0.321	0.326
Steady state	−5.5	0.9	0.314	0.316	0.316	0.317
Mean SII			0.17	0.35	0.33	0.34
Std SII			0.13	0.05	0.04	0.04

by omitting the first test and considering only the average values of the retest. Furthermore, the noise conditions with modulation frequencies below 8 Hz were excluded from the SII calculations. As mentioned earlier, noise conditions with these low modulation frequencies give a rise in SRT due to the masking of complete words.

A. Conventional SII calculations

The fourth column of Table III shows the results of the calculations with the conventional SII model (ANSI S3.5-1997, 1997). By definition, at threshold one would expect the SII to be similar across conditions, since threshold is reached by the availability of a given fixed amount of speech information. Table III shows that for the conventional SII this certainly is not true, which makes the conventional SII model a poor predictor for the speech intelligibility in fluctuating noise. The SII has a mean of 0.17 and a large standard deviation between conditions of 0.13. The SRTs predicted by the conventional SII model are given in Figs. 2–4 by a solid line, where the SII has been kept fixed to 0.35. As can be seen, the predicted SRT is virtually independent on the type of fluctuation of the noise masker—as expected, because the model departs from the long term spectra of speech and noise.

B. Extended SII calculations

The fifth column of Table III shows the extended SII calculations (Rhebergen and Versfeld, 2005). Here, the mean SII value is 0.35 and its standard deviation is equal to 0.05. Predictions of the SRT for the present conditions are plotted in Figs. 2–4 as a dotted curve, where the SII has been kept fixed to 0.35. Predictions with the extended SII model are far better than with the ANSI S3.5-1997 method, described in the previous section. However, it can be seen from the figures that the model still fails to adequately describe the data for conditions with relatively large silent gaps. Apparently, the masking function used in the model underestimates the real amount of masking in these conditions. Also, since the ESII model is a time-symmetric model, it is unable to account for the difference in threshold between the two conditions with the saw-tooth masker. The ESII model always will predict identical thresholds for noises that are each others time reversal. Simple adaptation of the integration time cannot solve the problem, since this results in deviations for the ESII for the other conditions.

In order to overcome these two shortcomings of the model, in the next section a forward masking function is introduced.

C. Implementation of forward masking in the extended SII

A large number of studies have shown the possibility of masking a target signal by a preceding masker (so-called forward masking), and its relationship between masker level and the time interval between masker and target signal (Pollack, 1955; Plomp, 1964; Elliott, 1969; Duifhuis, 1973; Widin and Viemeister, 1979; Jesteadt *et al.*, 1982; Moore and Glasberg, 1983; Kidd and Feth, 1982). These studies show a decrease in masking threshold with increased masker-signal delay. The masking threshold returns in about 200 ms to the level of the unmasked target signal threshold (i.e., when no masker is present), regardless of masker level. When plotting the masking thresholds (dB) as a function of masker-target gap on a logarithmic time scale, a linear relationship exists (e.g., Plomp, 1964). Ludvigsen (1985) has modeled this

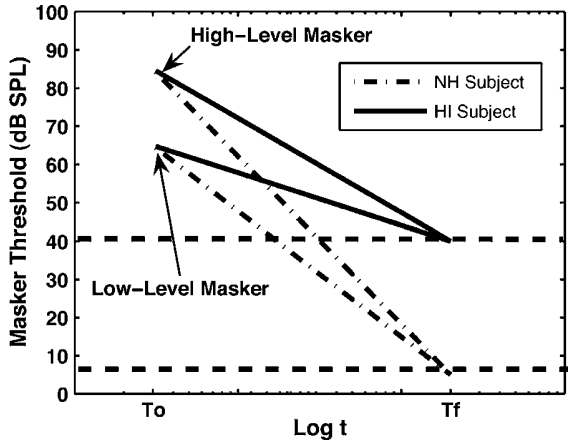

FIG. 6. Masked threshold (dB SPL) plotted as a function of time (on a logarithmic axis) for two masker levels (high-level masker and low-level masker) and for a normal-hearing and a hearing-impaired subject. Horizontal dashed lines indicate the absolute threshold of that subject. (Redrawn from Ludvigsen, 1985, Fig. 4, p. 1277.)

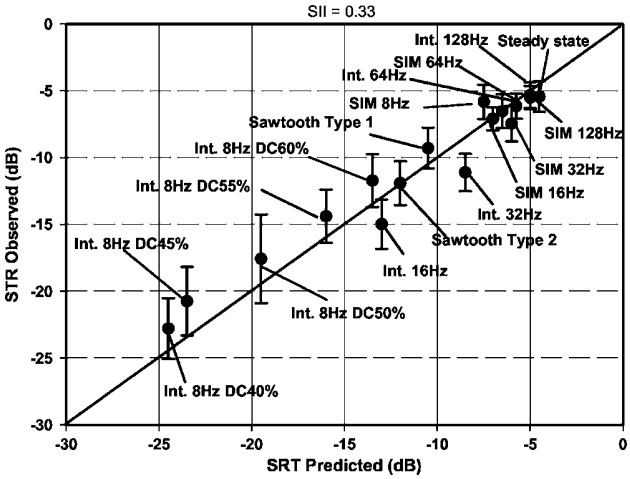

FIG. 7. For all conditions, the observed SRT (dB) is plotted as a function of the predicted SRT (dB), where the prediction has been made with the ESII model with the FMF included. Error bars denote the standard deviation between subjects.

function, which is called in this paper the forward-masking function (FMF). The model parameters were determined from the results of other studies reported in the literature, and the model predicts forward masking very well, not only for the normal hearing, but also for the hearing impaired. Figure 6 displays the FMF, i.e., the masked threshold as a function of time (reprinted from Ludvigsen, 1985) for two masker levels (high-level masker and low-level masker) and for a normal-hearing and a hearing-impaired subject. As can be seen in Fig. 6, the FMF is linear between T_0 and T_f when time is plotted on a logarithmic axis. Also, the duration of the FMF is always equal, regardless of hearing loss or masker level.

The middle portion of the FMF is a simple linear relationship, and is given by

$$E_{\mathrm{FMF}}(t) = E(T_0) - \frac{\log(t/T_0)}{\log(T_f/T_0)} * [E(T_0) - E(T_f)],$$

where at time T_0, the level of the linear portion $E(T_0)$ is equal to the level of the envelope of the masker, and where at time T_f, the linear function intersects with the absolute threshold of hearing $E(T_f)$. The values of the parameters T_0 and T_f are 2 and 200 ms, respectively. To take forward masking into account, the original envelope $E(t)$ is modified according to: $E(t)=\max[E(t),E_{\mathrm{FMF}}(t)]$. In this manner, sharp onsets in the envelope are followed instantaneously, but sharp offsets make that $E_{\mathrm{FMF}}(t)$ take over, resulting in a gradual decline of the envelope (due to forward masking). The FMF does not account for the phenomenon of backward masking, where a soft signal is masked by a louder signal that follows it. Backward masking is still poorly understood (Moore, 1997), and its effect on speech intelligibility is still unclear and probably not very large.

Note that the FMF is similar in the low and high frequency bands, whereas temporal integration is not. Both phenomena act separately on the signal, and probably are situated in different places in the auditory system. So far, forward masking was not modeled separately in the ESII, but rather was taken together with temporal integration. Effectively this gave longer integration times, which can explain that a best fit to the data of Rhebergen and Versfeld (2005) was obtained by multiplication of the time constants from Moore (1997) by a factor of 2.5. However, when forward masking is modeled separately from temporal integration, the original integration times (the original frequency dependent gap detection lengths) should be used.

The sixth column of Table III shows the ESII calculations with the FMF included. This addition of forward masking results in better predictions, i.e., SII values are closer together as can be deduced from the lower standard deviation (0.04). Indeed, the SII for those conditions with relatively long silent periods has decreased. In addition, the SII values for the two saw-tooth conditions are closer together.

When the calculation scheme is simplified by taking all integration times equal to 4 ms, the standard deviation in SII for all noise conditions remains the same (0.04), see the seventh column of Table III. This is because the time constants of the FMF are an order of magnitude larger than those of gap detection. Predictions of the ESII model with the FMF included are denoted in Figs. 2–4 with dashed lines, where the SII has been kept fixed to 0.35. Indeed, especially for conditions with larger silent gaps, this SII model predicts the data better. The shaded areas in Figs. 2–4 indicate the range of the predicted SRT when the SII value ranges between 0.30 and 0.40. With the exception of those noise conditions with a very low modulation frequency, all observed SRTs are close to the dashed line, and are situated in the shaded area. The benefit of the FMF is less clear in the SIM noise conditions (Fig. 4). Due to the fact that there are no complete silent periods in the SIM noise conditions (see Fig. 1 for comparison between SIM and interrupted noise), the differences between the ESII model with or without FMF is less clear.

Figure 7 displays for all conditions described in the previous section the relationship between the observed SRT and the SRT as predicted by the ESII model with the FMF. Predictions were made under similar assumptions as described

above, and the SII was taken equal to 0.33, the average value of the SII in Table III, column 6. If the data of Fig. 7 are considered in detail, some predicted SRTs lie above the diagonal. Since these conditions appear to be the conditions in which a learning effect was observed (see Fig. 5), it is expected that after all learning effects are overcome, SRTs will become better and, hence, lie closer to the diagonal.

Several conditions, especially those with low SRTs, show large between-subject differences. As argued above, these large differences are due to the shallowness of the psychometric function (SII as a function of SNR). When converting the observed individual SRTs to SII values, differences between subjects are comparable for all conditions. This indicates that no factors other than differences in the steepness of the psychometric function play a role. Alternatively, the width of the shaded area in Figures 2–4 (especially in Fig. 2) is clearly related to the width of the error bars.

VI. EXTENSIONS TO AND LIMITATIONS OF THE SII MODEL

The addition of a forward masking function has increased the predictive power of the ESII model, at least for normal-hearing subjects. However, experiments thus far dealt with presentation near 65 dB A only. It is known that with increasing level, the excitation pattern broadens, hence spectral resolution decreases, and temporal resolution increases. Although in principle the extended SII model can account for the increase in temporal resolution (Ludvigsen, 1985), it cannot account for the broadening of the auditory filters, since filter bandwidths are fixed in the model. A future extension to the SII model thus may be to implement a more realistic, level dependent, auditory filterbank.

For listeners with normal hearing or with a mild hearing loss, the SII model in its present form (ANSI S3.5-1997, 1997) is able to predict the speech intelligibility in stationary noise (Pavlovic, 1987; Noordhoek, 2000). However, the model is not meant to predict the speech intelligibility for listeners with moderate to severe hearing loss (Rankovic, 1998; Ching *et al.*, 1998, 2001; Hogan and Turner, 1998; Noordhoek, 2000). With these hearing losses the SII model overestimates performance. Hearing-impaired subjects often require SII values that exceed 0.33, which indicates that correction only for audibility is not sufficient. One reason may be that auditory processing is less than optimal (so-called suprathreshold deficits) (Noordhoek, 2000). Additionally, decreased spectrotemporal resolution may play a role. The latter factor may be accounted for by the extended SII model by parameter adjustment.

As mentioned above, the SRT paradigm is not particularly well suited for sentences in modulated noise where the modulation frequency is low and masking of whole words may occur. It is possible that SRTs are measured that do not reflect the actual masking situation, since the SRT procedure yields a correct score only when the whole sentence is scored correctly. Thus, with these low modulation frequencies, the increase in threshold is actually an experimental artifact. Indeed, when using a different experimental paradigm (Trine, 1995) thresholds do not increase when lowering the modulation frequency. With the present speech materials (Versfeld *et al.*, 2000), the speech rate is about 4–5 syllables per second, hence, a 4-Hz modulated signal may mask half of a word or more. Apparently, this already causes deterioration in intelligibility. Thus, there is a lower limit in the modulation frequency to which the SRT procedure is valid.

VII. SUMMARY

The present paper describes a validation study of the extension to the SII (ANSI S3.5-1997, 1997) proposed by Rhebergen and Versfeld (2005). The extended SII model was validated with SRT experiments with listeners with normal-hearing in nonstationary (or fluctuating) noise conditions that are critical to the extended SII model. From these data, it can be concluded that the extended SII model is able to predict the SRTs for the majority of conditions, but that predictions are better when the extended SII model includes a function to account for forward masking.

ACKNOWLEDGMENTS

Tammo Houtgast, Joost Festen, Gaston Hilkhuysen, and Erwin George are acknowledged for the inspiring discussions on the topic. László Körössy is especially acknowledged for his help with the computer programming and his help with the SRT computer program. The editor Dr. M. Sommers, and two anonymous reviewers are acknowledged for there useful comments.

ANSI (**1996**). "ANSI S3.6-1996, American national standard methods for specification for audiometers" (American National Standards Institute, New York).

ANSI (**1997**). "ANSI S3.5-1997, American national standard methods for calculation of the speech intelligibility index" (American National Standards Institute, New York).

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.

Ching, T. Y. C., Dillon, H., Katsch, R., and Byrne, D. (**2001**). "Maximising effective audibility in hearing aid fitting," Ear Hear. **22**, 212–224.

de Laat, J. A. P. M., and Plomp, R. (**1983**). "The reception threshold of interrupted speech for hearing-impaired listeners," in *Hearing—Physiological Bases and Psychophysics*, edited by R. Klinke and R. Hartman (Springer-Verlag, Berlin), pp. 359–363.

Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (**2002**). "Benefit of modulated maskers for speech recognition by younger and older adults with normal-hearing," J. Acoust. Soc. Am. **111**, 2897–2907.

Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (**2003**). "Recovery from prior stimulation: Masking of speech by interrupted noise for younger and older adults with normal-hearing," J. Acoust. Soc. Am. **113**, 2084–2094.

Duifhuis, H. (**1973**). "Consequences of peripheral frequency selectivity for nonsimultaneous masking," J. Acoust. Soc. Am. **54**, 1471–1488.

Duquesnoy, A. J. (**1983**). "Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons," J. Acoust. Soc. Am. **74**, 739–743.

Duquesnoy, A. J., and Plomp, R. (**1983**). "The effect of a hearing aid on the speech-reception threshold of hearing-impaired listeners in quiet and in noise," J. Acoust. Soc. Am. **73**, 2166–2173.

Elliott, L. L. (**1969**). "Masking of tones before, during, and after brief silent periods in noise," J. Acoust. Soc. Am. **45**, 1277–1279.

Festen, J. M. (**1987**). "Speech-perception threshold in a fluctuating background sound and its possible relation to temporal resolution," in *The Psychophysics of Speech Perception*, edited by M. E. H. Schouten (Martinus Nijhoff, Dordrecht), pp. 461–466.

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.

Fletcher, H., and Galt, R. H. (**1950**). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, 89–151.
French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.
Gustafsson, H. A., and Arlinger, S. D. (**1994**). "Masking of speech by amplitude-modulated noise," J. Acoust. Soc. Am. **95**, 518–529.
Hogan, C. A., and Turner, C. W. (**1998**). "High-frequency audibility: Benefits for hearing-impaired listeners," J. Acoust. Soc. Am. **104**, 432–441.
Houtgast, T., Steeneken, H. J., and Bronkhorst, A. W. (**1992**). "Speech communication in noise with strong variations in the spectral or the temporal domain," *Proceedings of the 14th International Congress on Acoustics*, Vol. 3, pp. H2–H6.
Jesteadt, W., Bacon, S. P., and Lehman, J. R. (**1982**). "Forward masking as a function of frequency, masking level, and signal delay," J. Acoust. Soc. Am. **71**, 950–962.
Kidd, G., and Feth, L. L. (**1982**). "Effects of masker duration in pure-tone forward masking," J. Acoust. Soc. Am. **75**, 1384–1386.
Koopman, J., Franck, B. A., and Dreschler, W. A. (**2001**). "Toward a representative set of "real-life" noises," Audiology **40**, 78–91.
Kryter, K. D. (**1962a**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697.
Kryter, K. D. (**1962b**). "Validation of the articulation index," J. Acoust. Soc. Am. **34**, 1698–1702.
Licklider, J. C. R., and Guttman, N. (**1957**). "Masking of speech by line-spectrum interference," J. Acoust. Soc. Am. **29**, 287–296.
Ludvigsen, C. (**1985**). "Relations among some psychoacoustic parameters in normal and cochlearly impaired listeners," J. Acoust. Soc. Am. **78**, 1271–1280.
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., and Glasberg, B. R. (**1983**). "Growth of forward masking for sinusoidal and noise maskers as a function of signal delay; implications for suppression in noise," J. Acoust. Soc. Am. **73**, 1249–1259.
Moore, B. C. (**1997**). *An Introduction to the Psychology of Hearing*, 4th ed. (Academic, London).
Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**, 365–378.
Nelson, P. B., Jin, S. H., Carney, A. E., and Nelson, D. A. (**2003**). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **113**, 961–968.
Noordhoek, I. M. (**2000**). "Intelligibility of narrow-band speech and its relation to auditory functions in hearing-impaired listeners," Doctoral thesis, Free University, Amsterdam.
Pavlovic, C. V. (**1987**). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. **82**, 413–422.
Plomp, R. (**1964**). "Rate of decay of auditory sensation," J. Acoust. Soc. Am. **36**, 277–282.
Plomp, R., and Mimpen, A. M. (**1979a**). "Improving the reliability of testing the speech reception threshold for sentences," Audiology **18**, 43–52.
Plomp, R., and Mimpen, A. M. (**1979b**). "Speech-reception threshold for sentences as a function of age and noise level," J. Acoust. Soc. Am. **66**, 1333–1342.
Pollack, I. (**1955**). "Masking by a periodically interrupted noise," J. Acoust. Soc. Am. **27**, 353–355.
Rankovic, C. M. (**1998**). "Factors governing speech reception benefits of adaptive linear filtering for listeners with sensorineural hearing loss," J. Acoust. Soc. Am. **103**, 1043–1057.
Rhebergen, K. S., and Versfeld, N. J. (**2005**). "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**, 2181–2192.
Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2005**). "Release from informational masking by time reversal of native and non-native interfering speech," J. Acoust. Soc. Am. **118**, 1274–1277.
Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. London, Ser. B **336**, 367–373.
Trine, T. D. (**1995**). "Speech recognition in modulated noise and temporal resolution: Effects of listening bandwidth," Doctoral dissertation, University of Minnesota, Twin Cities (unpublished).
van Wijngaarden, S. J., and Houtgast, T. (**2004**). "Effect of talker and speaking style on the Speech Transmission Index," J. Acoust. Soc. Am. **115**, 38–41.
Versfeld, N. J., Daalder, L., Festen, J. M., and Houtgast, T. (**2000**). "Method for the selection of sentence materials for efficient measurement of the speech reception threshold," J. Acoust. Soc. Am. **107**, 1671–1684.
Versfeld, N. J., and Dreschler, W. A. (**2002**). "The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners," J. Acoust. Soc. Am. **111**, 401–408.
Widin, G. P., and Viemeister, N. F. (**1979**). "Intensive and temporal effects in pure-tone forward masking," J. Acoust. Soc. Am. **66**, 388–395.