DOI: 10.61782/fa.2023.0094

THE SII MODEL HIGHLY UNDERESTIMATES EXTENDED HIGH FREQUENCY SPEECH INFORMATION IN DUTCH DIGITS, WORDS AND SENTENCES

Cas Smits[1,2], Sigrid Polspoel[2,3]
[1]Amsterdam UMC location University of Amsterdam, Otolaryngology-Head and Neck Surgery, Ear and Hearing, Meibergdreef, Amsterdam, The Netherlands
[2]Amsterdam Public Health research institute, Quality of Care, Amsterdam, The Netherlands
[3]Otolaryngology-Head and Neck Surgery, Section Ear and Hearing, Amsterdam UMC location Vrije Universiteit Amsterdam, Amsterdam, De Boelelaan, The Netherlands

ABSTRACT

Standard pure tone audiometry measures hearing thresholds up to 8 kHz. Recent studies have shown that speech information above 8 kHz ("extended high frequencies"; EHFs) improves speech recognition. However, it is unclear whether the EHF benefit depends on the complexity of the speech stimuli. Previously we investigated the added value of EHF information for speech recognition in noise for Dutch digits, words and sentences. Speech stimuli were presented at a fixed signal-to-noise ratio and listening conditions varied only based on available EHF information. The results confirmed findings from other studies and showed a significant benefit of EHF for speech recognition in normal hearing listeners. We have used an approximation of the Speech Intelligibility Index (SII) model to make a rough estimate of the amount of EHF speech information for the different speech materials used. The results suggest that the SII model highly underestimates the importance of EHF in speech recognition.

Keywords: *Consonant–vowel–consonant words, Digits-in-Noise, Extended high frequency, Speech recognition, Speech Intelligibility Index model*

1. INTRODUCTION

The role of extended high frequency (EHF; frequencies >8 kHz) speech information for speech recognition has long been considered insignificant or negligible. However, in recent years it has been shown that EHF speech information is indeed important for speech recognition, especially in noisy conditions. Comprehensive overviews of recent publications can be found in [1, 2]. Monson *et al.* [3] demonstrated a decrease in speech recognition for speech in two-talker babble with co-located talker and maskers with mismatched head orientations, when stimuli were low-pass (LP) filtered at 8 kHz. Motlagh Zadeh *et al.* [4] found significant better speech recognition scores when using LP filtered noise compared to broadband noise for digit-triplets in steady state noise.

The results of these studies raise the question of whether standard models such as the speech intelligibility index model (SII model [5]) are accurate. The SII model can be used to compute a physical measure (SII value) which represents the amount of speech information available to the listener depending on the speech signal, noise signal and hearing thresholds of the listener. Essentially, the SII value is calculated by summing the weighted frequency-specific signal-to-noise ratios. A speech dynamic range of 30 dB is assumed (from -15 dB signal-to-noise ratio, SNR, to +15 dB

[*] ***Corresponding author**: c.smits@amsterdamumc.nl*

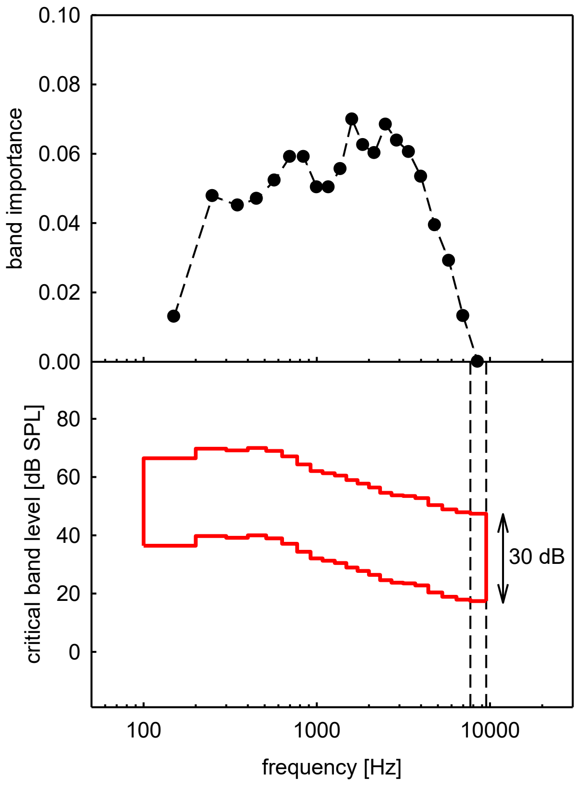

Figure 1. Upper panel shows the SPIN band importance function from the SII model. Lower panel shows the critical band levels for standard speech spectrum. The dashed lines represent the band limits of the highest frequency band in the critical band SII method.

SNR) and a frequency or band importance function is used to weigh the importance of the different frequency regions of the speech. The SII's critical band method uses 21 frequency bands and is the most accurate method in the standard. These bands cover the frequency range from 100 Hz to 9.5 kHz and are essentially the frequency bands 2 to 22 from Zwicker's 1961 classic publication [6, 7]. It is important to emphasize here that the first and last two bands (i.e., bands 1, 23 and 24) are thus considered irrelevant for speech recognition in the SII standard. Figure 1 illustrates the concept of the SII model. The upper panel shows the band importance function for speech material from the SPIN test which is often used when using the SII model for speech-in-noise data. The lower panel shows the critical band levels (± 15 dB) for the standard speech spectrum at an overall level of 62.35 dB SPL, as defined in the SII standard. Because the SPIN band importance is 0 for the highest band, the SII model assumes that only frequencies up to 7.7 kHz contribute to speech recognition for that speech material. Here, and further in the calculations in this paper, it is assumed that the speech level is above the hearing thresholds and that audibility therefore plays no role. This assumption is certainly not correct for all speech materials in the highest critical bands.

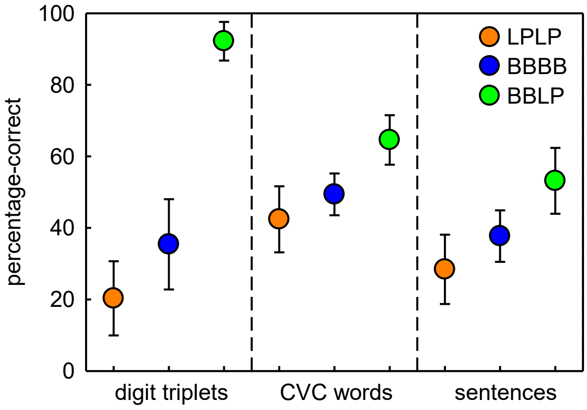

Figure 2. Average score (± SD) of the speech recognition experiment in noise from [8]. BBBB means both speech and noise broadband (BB) (i.e., unfiltered); LPLP means both speech and noise low-pass (LP) filtered at 8 kHz; BBLP means speech BB and noise LP filtered at 8 kHz.

The aim of the current study is to estimate the amount of EHF speech information, represented by the sum of the band importance function across EHFs, for three types of standard Dutch speech material using previously published data [8].

2. SUMMARY OF POLSPOEL *ET AL.* [8]

Dutch digit-triplets, consonant-vowel-consonant words (CVC), and sentences were presented monaurally to twenty-four young adults with normal hearing thresholds (≤ 20 dB HL) up to 16 kHz. Steady-state speech-shaped noises were used as maskers at fixed SNRs of -9 dB (digits triplets), -8 dB (CVC words) and -5 dB (sentences). All three speech materials were presented in three listening conditions that only varied in terms of available EHF information: (1) The BBBB condition, where both speech and noise were unfiltered (broadband; BB); (2) The LPLP condition, where both speech and noise were low-pass (LP) filtered with a cutoff frequency of 8 kHz; and (3) The BBLP condition, where the speech was unfiltered (BB) and the noise LP filtered with a cutoff frequency of 8 kHz. The results showed that for all speech material, the highest scores were achieved

Table 1. Parameters from the speech recognition function [11], average percentage correct scores from [8] and the calculated SII values for each condition.

	Speech recognition function		Percentage-correct			SII		
	S_{50} [%/dB]	SRT [dB SNR]	LPLP	BBBB	BBLP	LPLP	BBBB	BBLP
digit-triplets	17.0	-8.4	20	37	92	0.16	0.19	0.33
CVC words	5.8	-7.3	42	50	65	0.21	0.26	0.34
sentences	10.5	-8.1	28	38	53	0.16	0.19	0.24

in the BBLP condition and the lowest scores in the LPLP condition (see Figure 2). Adding speech frequencies above 8 kHz to the LPLP condition improved the mean recognition scores by 72, 22 and 25 percentage points for digit triplets (triplet scoring), words (phoneme scoring) and sentences (word scoring), respectively.

3. ESTIMATION OF SII VALUES

The transfer function relates SII values to percentage correct scores. Therefore, we can use this function to estimate the SII values for each condition for the three speech materials used in study [5].

As explained and discussed in detail in [9, 10], the transfer function is identical for the speech recognition function in steady-state speech shaped noise where the SNR has been replaced by SII = (SNR+15)/30. The speech recognition functions for digit-triplets, CVC words and sentences were previously determined in a group of normal hearing listeners [11]. Cumulative normal distributions were used to fit the data. The speech recognition threshold (SRT, i.e., the SNR at 50% correct) and slope at 50% correct (S_{50}) are presented in Table 1.

Also shown in Table 1 are the average percentage correct scores for the normal hearing listeners from study [8]. Using the inverse transfer function, the SII values for each condition were calculated and presented in Table 1.

4. ESTIMATION OF THE IMPORTANCE OF EHF

The SII model provides a standardized method to determine speech intelligibility by calculating the audibility in each frequency band and summing the weighted results:

$$SII = \sum_{i=1}^{n} I_i A_i \qquad (1)$$

In study [8], the audibility in each band is either 0 (no speech), 1 (speech present, no noise) or partly masked (in which A can be approximated by (SNR+15)/30). To analyse the results from [8], we only need two bands: <8 kHz and >8 kHz which yields for the three conditions:

$$\begin{aligned} SII_{LPLP} &= I_{<8kHz} \cdot A_{<8kHz} \\ SII_{BBBB} &= I_{<8kHz} \cdot A_{<8kHz} + I_{>8kHz} \cdot A_{>8kHz} \quad (2) \\ SII_{BBLP} &= I_{<8kHz} \cdot A_{<8kHz} + I_{>8kHz} \end{aligned}$$

in which $I_{>8kHz}$ equals the sum of the band importance function >8kHz, i.e., the amount of EHF speech information. Figure 3 illustrates the effect of the different conditions on the audibility of the speech. From the figure it can be determined how any two combinations of conditions can be used to estimate the relative importance of EHF (Table 2).

5. DISCUSSION

Our data analyses from study [8] indicated that the standard band importance functions from the SII model are not appropriate for standard Dutch speech recognition materials. The SPIN band importance function assigns no contribution to frequencies above 8 kHz for speech recognition in noise. Other band importance functions from the SII model indicate some weighting to the highest critical band, but with a maximum of 0.0162, corresponding to less than 2% of the speech information [5]. We estimated that approximately 10 to 20% of the speech information is in the EHF region for the Dutch speech material. It should be noted that these are rough estimates and standard procedures are needed to more

Table 2. Estimated values of the importance of EHF speech information

	$I_{>8\ kHz}$		
	$(SII_{BBBB}-SII_{LPLP})/SII_{BBBB}$	$SII_{BBLP}-SII_{LPLP}$	$30/(15\text{-}SNR)\cdot(SII_{BBLP}-SII_{BBBB})$
digit-triplets	0.20	0.18	0.17
CVC words	0.17	0.13	0.11
sentences	0.18	0.08	0.07

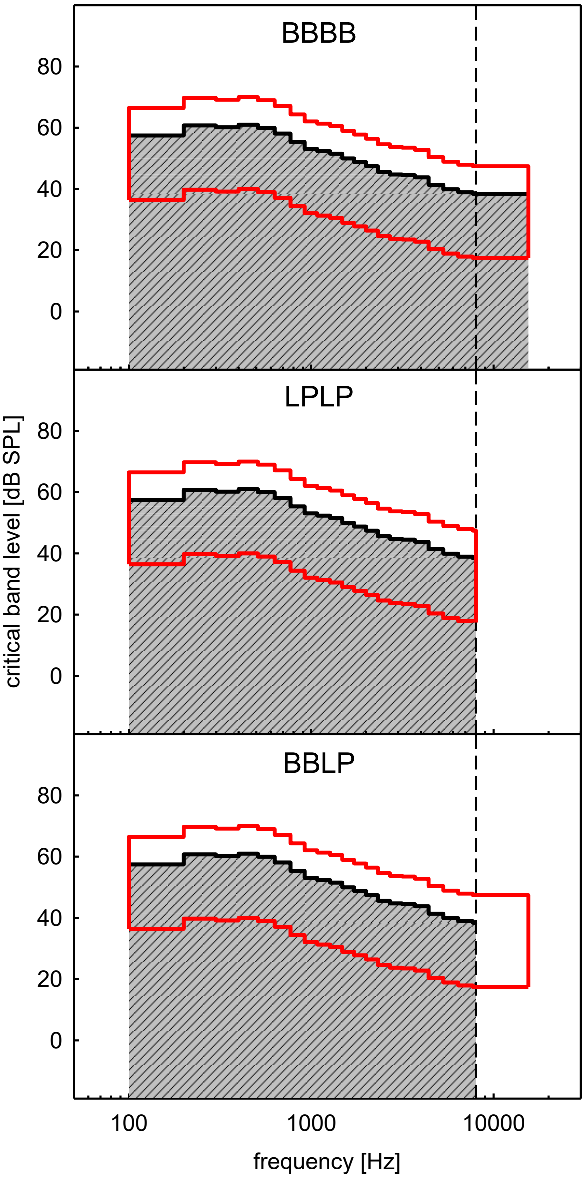

Figure 3. Schematic illustration of the effect of LP filtering at 8 kHz and noise level on the amount of available speech information.

accurately determine the band importance function [12]. The results thus provide strong reasons to revise the current SII model calculations where it is essential to extend the frequency range and (re)determine the weighting functions. To stay as close as possible to the current model, it seems obvious to include Zwicker's two highest critical bands [6] in the SII norm, extending the frequency limit to 15.5 kHz. When using the current SII model, it is highly recommended to lowpass filter the speech material at 9.5 kHz. Monson and Buss [13] analyzed the spectral content of popular English speech corpora used in speech recognition research and emphasized the impact the speech material choice could have on experimental results. They showed large differences in EHF energy between the English speech materials. There are several reasons why the contribution of EHFs to speech recognition has not been extensively studied in the past. Technical limitations play an important role: poor test-retest reliability, difficulties with calibration, poor signal-to-noise ratios and low maximum output due to the low quality of amplifiers [14]. Pollack [15], for example, studied the effect on intelligibility of eliminating the high frequency speech sounds and he reported that the cut-off frequency of their headphones was 7 kHz. Additionally, calibration standards for EHF audiometry were established relatively late [16], and the general belief was that speech signals were mostly inaudible at high frequencies [1].

We did not perform a systematic error analysis, but errors in the estimates of $I_{>8\text{ kHz}}$ are caused by the limited accuracy of the percentage correct scores in [8], and the slope and SRT of the speech recognition function used for the transfer function. It appears that a 2 percent point change in the average speech recognition scores results in changes of the order of 0.02 for the average $I_{>8\text{kHz}}$. Calculations are relatively insensitive to changes in the SRT of the speech recognition function: a 2 dB change in SRT results in changes of approximately 0.01 in $I_{>8\text{KhZ}}$. A change in slope of the speech recognition function is approximately inversely proportional to a change in $I_{>8\text{kHz}}$. Thus, a 10% steeper slope yields a 10% smaller value of $I_{>8\text{kHz}}$.

Tabel 2 shows that $I_{>8\text{ kHz}}$ estimates are lower when the BBLP condition is used in the calculations (third and fourth column compared to second column). The most likely explanation for this finding is the role of audibility in the BBLP condition. Although the spectral level of speech is much lower in the EHFs than in the standard frequencies, the differences in critical band levels are much smaller across all bands (see Fig. 1) [1]. However, because human hearing is less sensitive to EHFs [17], it is likely that the lower part of the 30 dB dynamic range of the EHF speech is inaudible. This will have a particularly large effect in the BBLP condition and for the sentence material which has relatively low speech levels in the EHF range [8]. Thus, we expect that $I_{>8\text{ kHz}}$ values based on the BBBB and LPLP conditions are the most accurate estimates.

6. CONCLUSION AND RECOMMENDATIONS

- Estimates based on speech recognition experiments in noise with standard Dutch digit-triplets, CVC words and sentences show that approximately 10-20% of the speech information is present in the EHF region.
- The current SII model does not accurately account for EHF speech information.
- We recommend to include the two highest critical bands from [6] in the SII norm, extending the frequency limit to 15.5 kHz.
- New band importance functions must be determined for speech material with EHF speech.

7. REFERENCES

[1] Hunter, L.L., B.B. Monson, D.R. Moore, S. Dhar, B.A. Wright, K.J. Munro, L.M. Zadeh, C.M. Blankenship, S.M. Stiepan, and J.H. Siegel, "Extended high frequency hearing and speech perception implications in adults and children," *Hearing research*, 397, pp. 107922, 2020

[2] Monson, B.B. and A. Trine. "Extending the High-Frequency Bandwidth and Predicting Speech-in-Noise Recognition: Building on the Work of Pat Stelmachowicz." *Seminars in Hearing*. Thieme Medical Publishers, Inc., Year

[3] Monson, B.B., J. Rock, A. Schulz, E. Hoffman, and E. Buss, "Ecological cocktail party listening reveals the utility of extended high-frequency hearing," *Hearing Research*, 381, pp. 107773, 2019

[4] Motlagh Zadeh, L., N.H. Silbert, K. Sternasty, D.W. Swanepoel, L.L. Hunter, and D.R. Moore, "Extended high-frequency hearing enhances speech perception in noise," *Proceedings of the National Academy of Sciences*, 116(47), pp. 23753-23759, 2019

[5] ANSI, "S3. 5-1997, Methods for the calculation of the speech intelligibility index," *New York: American National Standards Institute*, 19, pp. 90-119, 1997

[6] Zwicker, E., "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *The Journal of the Acoustical Society of America*, 33(2), pp. 248-248, 1961

[7] Pavlovic, C.V., "Derivation of primary parameters and procedures for use in speech intelligibility predictions," *The Journal of the Acoustical Society of America*, 82(2), pp. 413-422, 1987

[8] Polspoel, S., S.E. Kramer, B. van Dijk, and C. Smits, "The Importance of Extended High-Frequency Speech Information in the Recognition of Digits, Words, and Sentences in Quiet and Noise," *Ear and hearing*, 43(3), pp. 913-920, 2022

[9] Smits, C. and J.M. Festen, "The interpretation of speech reception threshold data in normal-hearing and hearing-impaired listeners: steady-state noise," *The Journal of the Acoustical Society of America*, 130(5), pp. 2987-98, 2011

[10] Smits, C., K.C. De Sousa, and D. Swanepoel, "An analytical method to convert between speech recognition thresholds and percentage-correct scores for speech-in-noise tests," *Journal of the Acoustical Society of America*, 150(2), pp. 1321-1331, 2021

[11] Smits, C., S. Theo Goverts, and J.M. Festen, "The digits-in-noise test: assessing auditory speech recognition abilities in noise," *The Journal of the Acoustical Society of America*, 133(3), pp. 1693-706, 2013

[12] Jin, I.-K., J.M. Kates, K. Lee, and K.H. Arehart, "Derivations of the band-importance function: A cross-procedure comparison," *The Journal of the Acoustical Society of America*, 138(2), pp. 938-941, 2015

[13] Monson, B.B. and E. Buss, "On the use of the TIMIT, QuickSIN, NU-6, and other widely used bandlimited speech materials for speech perception experiments," *The Journal of the Acoustical Society of America*, 152(3), pp. 1639-1645, 2022

[14] Fausti, S.A., R.H. Frey, D.A. Erickson, B. Rappaport, E.J. Cleary, and R.E. Brummett, "A system for evaluating auditory function from 8000–20 000 Hz," *The Journal of the Acoustical Society of America*, 66(6), pp. 1713-1718, 1979

[15] Pollack, I., "Effects of high pass and low pass filtering on the intelligibility of speech in noise," *The Journal of the Acoustical Society of America*, 20(3), pp. 259-266, 1948

[16] ISO/TR 389-5:1998 Acoustics — Reference zero for the calibration of audiometric equipment — Part 5: Reference equivalent threshold sound pressure levels for pure tones in the frequency range 8 kHz to 16 kHz.

[17] Levy, S.C., D.J. Freed, M. Nilsson, B.C. Moore, and S. Puria, "Extended High-Frequency

Bandwidth Improves Speech Reception in the Presence of Spatially Separated Masking Speech," *Ear and hearing*, 36(5), pp. e214-24, 2015