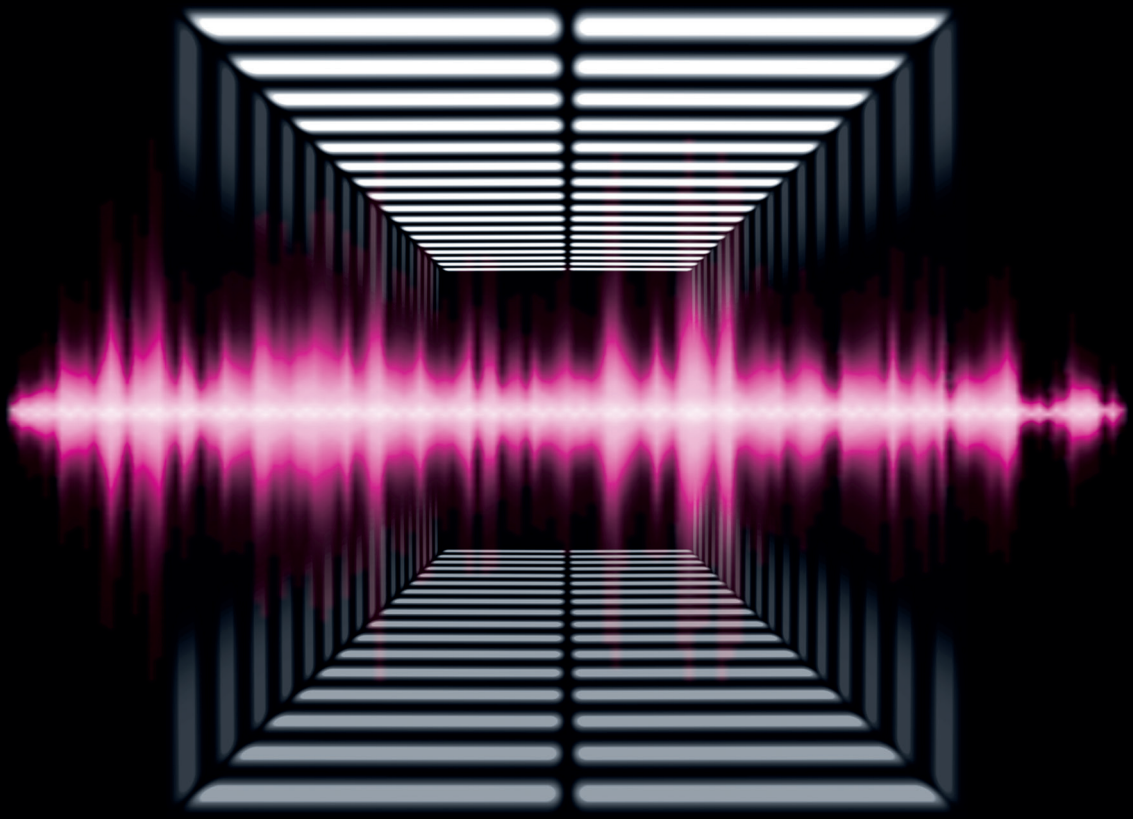
THE EXTENDED SPEECH TRANSMISSION INDEX

Predicting speech intelligibility in non-stationary noise and reverberation



Jelmer van Schoonhoven

THE EXTENDED SPEECH TRANSMISSION INDEX

Predicting speech intelligibility in non-stationary noise and reverberation

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 9 november 2023, te 13.00 uur

door Jelmer van Schoonhoven
geboren te Wonseradeel

Colofon

The Extended Speech Transmission Index:
Predicting speech intelligibility in non-stationary noise and reverberation
Jelmer van Schoonhoven

Design/lay-out
Promotie In Zicht | www.promotie-inzicht.nl

Print
Drukkerij Walden | www.walden.nl

Promotiecommissie

Promotor:	prof. dr. ir. W.A. Dreschler	AMC-UvA
Copromotor:	dr. K.S. Rhebergen	UMC Utrecht
Overige leden:	prof. dr. ir. J.C.M. Smits	AMC-UvA
	dr. ir. P. Brienesse	AMC-UvA
	prof. dr. P.P.G. Boersma	Universiteit van Amsterdam
	prof. dr. P. van Dijk	Rijksuniversiteit Groningen
	prof. dr. ir. T. Houtgast	Vrije Universiteit Amsterdam

Faculteit der Geneeskunde

Table of contents

CHAPTER 1

GENERAL INTRODUCTION

1.1 Perception of sound

"A sound is said to exist if a disturbance propagated through an elastic material causes an alteration in pressure or displacement of the particles of the material which can be detected by a person or by an instrument" (Beranek and Mellow, 2012). When the sound in question is speech, the elastic material is often air (although the own voice is partly perceived via bone conduction). The vocal cords of the speaker cause air molecules to vibrate, leading to small, local pressure variations relative to the atmospheric pressure. Due to repeated collisions with neighboring molecules, these pressure variations propagate through the air. Molecule displacements that lead to perception of sound can be smaller than one nanometer, and local pressure variations can be as small as one billionth of the atmospheric pressure.

The hearing organ is sensitive to these tiny pressure variations. The human eardrum vibrates in response to the colliding air molecules and efficiently relays the sound energy via the middle ear ossicles to the inner ear. Here, in the cochlea, which is basically a fluid filled cylinder with two compartments, the transduction of the mechanical vibration to an action potential takes place. The mechanical vibration sets the basilar membrane in motion, which is tonotopically organized due to the variation in its width and its stiffness along the length axis. This tonotopicity causes the amplitude of vibration on the basilar membrane to be largest at the location that corresponds to the frequency of the presented sound. In this way, the cochlea behaves like a frequency analyzer, where the sharpness of the basilar membrane peak determines the specificity of the analyzer. The active contribution of the outer hair cells leads to amplification of soft sounds, but also to improved spectral and temporal resolution of the inner ear. The inner hair cells are activated by the motion of the basilar membrane and eventually they cause activation of the synapses of the auditory nerve fibers. The resulting electrical activity is now relayed via the cochlear nucleus and superior olivary nucleus to the auditory cortex. The ipsi- and contralateral pathways of the auditory system are combined at the level of the brainstem and at higher levels.

1.2 Speech perception

In clinical practice, detection of sound is usually tested using pure tones. A pure tone consists of one frequency and therefore excites a small region on the basilar membrane, depending on the sound level of the tone. Due to their simplicity, pure tones are ideal for determining the hearing threshold for

specific frequencies. However, detecting soft, pure tones does not fully represent hearing and listening in daily life. As opposed to pure tones, the speech signal is a highly complex and dynamic broadband sound, strongly modulated in time and frequency. Speech intelligibility is not a matter of mere detection, but the result of a complex analysis of the spectral and temporal properties of the speech signal.

Two important characteristics of a speech signal are the temporal envelope and the temporal fine structure. The contribution of the temporal envelope to speech perception has been studied extensively (Houtgast and Steeneken, 1985; Stone *et al.*, 2010; Fogerty, 2011) and several intelligibility models were based on the modulations of the speech envelope (e.g., Houtgast and Steeneken, 1978; Jørgensen and Dau, 2011; Taal *et al.*, 2011). Drullman (1995) investigated the influence of the envelope and the fine structure on the intelligibility of speech. He found that speech with random fine structure and an intact envelope was perfectly intelligible, whereas intelligibility dropped to 17% when the fine structure was intact with a random temporal envelope.

Decoding of the speech signal is a bottom-up process in the auditory system, but top-down processing is also of vital importance. Higher order factors like linguistic skills, contextual information, a priori knowledge, auditory attention and expectations of the listener contribute to speech perception (e.g., Bronkhorst *et al.*, 1993). Most people take the intelligibility of speech for granted when no problems occur. When difficulties arise — due to difficult circumstances and/or hearing problems — it becomes clear what a difficult and energy consuming task speech intelligibility can be. This is the result of the extra resources that are needed for the top-down processes, to be able to compensate for the degraded input signal of the auditory system.

1.2.1 Distorted speech

There are various ways to distort sounds. Several nonlinear methods are — but not limited to — deterministic envelope reduction (Noordhoek and Drullman, 1997), envelope compression (Drullman, 1995; Hohmann and Kollmeier, 1995; Rhebergen *et al.*, 2009), peak & center clipping (Steeneken and Houtgast, 2002), and spectral subtraction (Ludvigsen *et al.*, 1993; Dubbelboer and Houtgast, 2007). However, the current thesis only deals with room acoustics and therefore distortion types will be limited to noise and reverberation.

1.2.1.1 Stationary noise

Different definitions for *noise* may be found in the dictionary. Of these definitions "any sound that is undesired or interferes with one's hearing of something" (Merriam-Webster.com, January 24, 2023) probably defines the concept best as it is used in the current text. What is a pleasant conversation or agreeable music to one person, is troublesome noise to another, for instance if it hinders the intelligibility of speech.

Stationary noise is probably the most used type of distortion in clinics and research. Often, noise is used with a gaussian distribution and a frequency spectrum that matches the spectral content of the speech material. Stationary noise can mask speech, making certain speech sounds inaudible. This type of masking is traditionally referred to as energetic masking (EM). Speech is a robust, highly redundant signal. Due to this redundancy, speech can be distorted quite a lot before intelligibility deteriorates for a normally hearing person. For monaurally presented Dutch sentences, the sound level of stationary speech-shaped noise (SSN) can be about 5 dB higher than the speech, while a healthy listener is still able to repeat 50% of the sentences correctly (e.g., Versfeld *et al.*, 2000; Rhebergen *et al.*, 2006). Although this signal-to-noise ratio or SNR of approximately −5 dB is valuable in a clinical setting, in normal circumstances with noise present, people usually function in SNRs between 0 and 15 dB (Olsen, 1998; Wu *et al.*, 2018). At these SNRs the intelligibility of the above sentence material is higher than 95% (Versfeld *et al.*, 2000). Besides this, most real-life background noises are not stationary, but are temporally modulated (Koopman *et al.*, 2001).

1.2.1.2 Non-stationary noise

When the envelope of the noise is temporally modulated, listeners make use of the gaps in the noise and intelligibility increases when compared to SSN with the same sound level (e.g., Festen and Plomp, 1990). This fluctuating masker benefit or FMB is sometimes compared to the 'picket fence' theory (Miller and Licklider, 1950); a visual analogy when looking to a landscape through a picket fence. The brain restores the visual fragments and forms a complete image, instead of a fragmented version of the landscape. By using similar forms of top-down restoration, the speech fragments are tied together in order to increase intelligibility.

However, when a masker is modulated, the modulations can also *interfere* with the speech modulations and hamper intelligibility as a result. This form of masking is often referred to as modulation masking (MM). Especially when the modulation spectra of the speech and noise overlap, the fluctuating masker benefit is counteracted (Fogerty *et al.*, 2016). Houtgast (1989) concluded that the modulation-detection threshold is highest when the test modulation frequency overlaps with the masker modulation band. Dau *et al.* (1997a; 1997b) also conducted experiments regarding amplitude modulation detection and could quantitively explain the results using a model based on a modulation filterbank.

Whether or not noise is purely stationary can be quite arbitrary. In fact, Stone *et al.* (2012) refers to SSN as *notionally* steady background noise, since random fluctuations exist and might be an important contributor to the masking of speech (Dubbelboer and Houtgast, 2008; Jørgensen and Dau, 2011). Drullman (1995) concluded that these random fluctuations in SSN cause spurious modulations that interfere with the perception of relevant speech modulations. Stone *et al.* (2011) concluded, based on experiments with vocoded speech and noise, that the random fluctuations in SSN have a large effect on speech intelligibility. Besides this, the introduction of 8 Hz modulations in certain (vocoded) test conditions resulted in *lower* intelligibility. This opposes the more traditional view of the fluctuating masker benefit. The authors argue that the masking effect of notionally steady state noise may primarily be the result of MM and not so much of EM.

1.2.1.3 Competing speaker

In the examples above, background noise was assumed without any fine structure. When the distorting noise is a competing speaker, the listener still benefits from the gaps in the noise. However, the informational content of the background noise distracts the listener and causes a decrease in intelligibility (Rhebergen *et al.*, 2005; Durlach, 2006). Although a formal definition is missing, this form of masking is generally referred to as informational masking or IM. Note that a non-intelligible, but speech-like signal like the International Speech Test Signal or ISTS (Holube *et al.*, 2010) does not contain any semantic information, but still causes some degree of informational masking, estimated at 4.6 dB when listening to a female Dutch speaker (Francart *et al.*, 2011).

Auditory attention plays an important role in speech intelligibility, especially in complex listening environments and when competing speakers are present. Object formation, object selection and stream segregation are crucial in understanding speech in the presence of a competing speaker. Shinn-Cunningham (2008) defines an object as "a perceptual estimate of the content of a discrete physical source", and streaming as "grouping of short-term auditory objects across longer time scales".

1.2.1.4 Reverberation

Besides noise, another type of distortion that hinders intelligibility is reverberation. Soundwaves reflect upon a surface and reach the ear of listener ear slightly later than the direct sound waves. Depending on the ratio between the direct and the reverberant sound, the quality of the speech signal decreases. Note that early reflections (within 50–100 ms after the direct sound) might aid intelligibility (e.g., Lochner and Burger, 1964; Boothroyd, 2004; Warzybok *et al.*, 2013). The ratio

between direct and reverberant sound is mostly influenced by the room acoustics (e.g., room size and reflecting versus absorbing surfaces) and by the distance between the speaker and the listener.

Like other room acoustics parameters, the characteristics of the reverberation can be derived from the impulse response (Hak *et al.*, 2012). It is classically measured by recording the reverberant sound energy after an impulse sound. This probe signal can be a gun shot or a popping balloon. However, a more robust and controlled measurement method uses a sweep as a probe signal, which is a sine wave with increasing frequency as a function of time (ISO3382-2, 2008). By deconvolving the recorded signal using the original sweep signal, the impulse response is obtained. The impulse response provides information about the degree of reverberation and can also be used to estimate the Modulation Transfer Function (MTF) as a result of reverberation (Schroeder, 1978; 1981). Inversely, it can also be used to convolve an existing speech signal in order to simulate certain acoustical conditions (e.g., George *et al.*, 2008).

Reverberation generally leads to poorer speech intelligibility. The depth of the temporal modulations in speech decreases, since the gaps are filled with reverberant sound energy. Since modulations in the speech signal carry information, this modulation reduction reduces the quality of the speech signal (Houtgast and Steeneken, 1973). Nabelek *et al.* (1989) also discuss overlap-masking, where a consonant can mask a subsequent segment of the speech, and self-masking, resulting in temporal smearing of the sound energy within each consonant. This type of distortion is primarily caused by the effect of reverberation on the fine structure of the speech.

1.2.2 Spatial separation

When the sources of noise and speech are spatially separated, intelligibility generally increases. This is referred to as spatial release of masking (Plomp, 1976; Bronkhorst, 2015). Depending on the location of the speech and noise sources relative to the listener, this might be caused by the head shadow effect, leading to a more favorable SNR at the ear opposite of the noise source. Besides the head shadow effect, another mechanism that leads to increased intelligibility is binaural squelch, caused by the phase difference between the ears of the presented signals (Dieudonne and Francart, 2019). For normally hearing subjects and speech masked by SSN, the head shadow effect is roughly 5 – 8 dB and binaural squelch 2 – 5 dB (Bronkhorst and Plomp, 1989; Dieudonne and Francart, 2019). Since the current work primarily focuses on the monaural presentation of speech and noise, no in-depth description of the effects of spatial separation is provided here.

1.2.3 Context

A topic that is important in speech intelligibility and that is widely used in the current work is *context*. Miller *et al.* (1951) found that the intelligibility of monosyllabic words increased as the size of the test vocabulary decreased. They also saw an increase in intelligibility when a word was presented in a sentence rather than in isolation. Both the a priori knowledge the listener has about the test vocabulary (the correct answer can only be one of N alternatives), and the syntax and semantics of the sentence are examples of context. Especially in challenging listening environments, listeners often cannot identify all of the speech elements using sensory information alone. In this case context is used to infer what elements were missed. This process is easier when the topic of the conversation is known and when the listener is fluent in the language being spoken. In both cases the listener is better able to use context to aid intelligibility. Several methods are available that aim to model the effects of context. Most notable are the models by Boothroyd and Nittrouer (1988) and by Bronkhorst *et al.* (1993), which will be discussed more thoroughly in chapter 4.

1.3 Modelling speech intelligibility

1.3.1 Overview

The first method that was used to estimate intelligibility was the Articulation Index or AI (French and Steinberg, 1947; Fletcher and Galt, 1950; Kryter, 1962; ANSI-S3.5, 1969). The AI was developed at Bell Labs in the early days of telephone communication as result of the ambition to improve intelligibility due to the poor signal quality of telephone systems. Analyses of communication channels using the AI were a lot cheaper than performing speech intelligibility tests. The AI uses the premise that the articulation error in nonsense speech is the product of the articulation error in lowpass and high-pass filtered speech (e_L and e_L respectively). This relation generalizes for K frequency bands (often 21 critical bands). An important extension was provided by French and Steinberg (1947) when they related the articulation error to the SNR per frequency band.

The AI was used to improve telephone communication channels and British pilot-crew communications during the second world war. The AI eventually evolved into the Speech Intelligibility Index or SII (Pavlovic, 1987; ANSI-S3.5, 1997) and can be viewed as the proportion of the total speech information that is audible for the listener. It is a value between zero and one, which can be converted to the intelligibility of a specific speech corpus via a transfer function. For example, under the same conditions, a higher SII (or AI) is needed for the recognition of words in sentences with low predictability than for words in sentences with high predictability (Sherbecoe and Studebaker, 1990; Bell *et al.*, 1992). This is an effect of context and as consequence, a different transfer function is needed for each type of speech corpus.

An important application of the SII is the evaluation of communication channels. However, when it concerns room acoustics, noise is not the only factor that is detrimental to speech intelligibility. Reverberation also needs to be accounted for. In the seventies and eighties, the Speech Transmission Index (STI) was developed (Houtgast and Steeneken, 1973; Houtgast *et al.*, 1980; Houtgast and Steeneken, 1985; IEC60268-16, 2011). Where the SII was based on *audibility* of speech, the STI was based on the *modulations* that are present in the speech. Both noise and reverberation lead to a reduction in modulation depth and a decrease in intelligibility. This modulation reduction can be converted to an index between 0 and 1 using several calculation steps. The STI is often used as a measurement method in room acoustics, like in the design of auditoria. In this situation, no prior knowledge about the noise or acoustics is needed. Alternatively, like the SII, it can also be used as a model for the estimation of the intelligibility of a certain speech corpus.

Over the years, various other models have been developed that aim to predict speech intelligibility. Examples that are derived from the SII are the Extended SII (Rhebergen *et al.*, 2005; 2006) to better deal with non-stationary noises and the Binaural Speech Intelligibility Model or BSIM (Beutelmann *et al.*, 2010) for binaural hearing and spatial separation of sound sources. Models that use the speech envelope to predict intelligibility are for example the Envelope Power Spectrum Model or EPSM (Jørgensen and Dau, 2011), the Short Time Objective Intelligibility Measure or STOI (Taal *et al.*, 2011) and different versions of the STI using speech as a probe signal (e.g., Payton and Braida, 1999; 2002). Lastly, various models that were derived from automated speech recognition systems were developed over the recent years. See Karbasi and Kolossa (2022) for a review.

1.3.2 Speech Transmission Index (STI)

1.3.2.1 Overview

The current thesis primarily focusses on the Speech Transmission Index. A detailed description is provided in chapter 3, and in IEC60268-16 (2011) and (IEC60268-16, 2020)[i]. In short, the STI aims to calculate the modulation transfer function (MTF) as a result of noise and reverberation. Classically, this is done by

i In the current work, IEC60268-16 (2011) was primarily referenced. However, since 2020 a revised edition of the standard is available (IEC60268-16, 2020). This will be discussed briefly in chapter 7.

presenting 98 separate modulated test signals by using all combinations of seven octave frequency bands and 14 modulation frequencies. With an average of 10 seconds per test signal, the full traditional STI measurement requires about 15 minutes. The properties of the transmission channel (e.g., a room with background noise and reverberation) cause changes in the signal modulations. After recording the signal and calculating the modulation reduction for each combination of octave band and modulation frequency band, the MTF can be calculated. The MTF is eventually converted to the STI.

1.3.2.2 Strengths and limitations

The main advantage of the STI is its easy applicability and elaborate validation (Houtgast and Steeneken, 1978; Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985; Steeneken and Houtgast, 1999; 2002). Using a calibrated system with a loudspeaker and a microphone, a reasonably accurate result can be obtained in a relatively short time period. When using derivations such as STIPA or RASTI (now obsolete), the measurements even take up less time (IEC60268-16, 2011). Over the years the STI has been widely used, validated and updated, which makes it a reliable method in room acoustics.

One disadvantage is the sensitivity of the classic, direct measurement method to fluctuations in the background noise. The STI assumes that modulation reduction only occurs as a result of reverberation and stationary noise. When noise is non-stationary, reliability drops (IEC60268-16, 2011). When using the full STI and the direct measurement method it takes 15 minutes to do one measurement. Although this is relatively short, the probability is high that some noise peaks occur during the measurement (e.g., a slamming door or an interfering speaker).

An alternative option is the indirect measurement method, where the MTFs as a result of noise (MTF_{SNR}) and as a result of reverberation (MTF_{rev}) are measured separately. The MTF_{SNR} can easily be calculated by recording the background noise for any desired period of time. When the speech level is known, the SNR can be used to calculate the MTF_{SNR}. The MTF_{rev} can be calculated using the impulse response (Schroeder, 1978; Houtgast and Steeneken, 1985). This indirect measurement method is less sensitive to fluctuations in the noise. However, it is not clear under what circumstances this method can be used reliably.

A second disadvantage of the STI also concerns fluctuations in the noise. As mentioned earlier, when noise is non-stationary, the fluctuating masker benefit causes intelligibility to increase. If the STI measurement can be done reliably in non-stationary noise, the next question is how the resulting STI-value is related to intelligibility. There is a high probability that the STI-value will be too low, since the model assumes stationarity of the noise.

1.4 Outline of this thesis

The goal of the current thesis was to increase applicability of the STI in non-stationary background noise. To achieve this goal, both the measurement method and the fluctuating masker benefit were investigated. A secondary objective was to deviate as little as possible from the original calculation scheme. In doing so, the complexity of the model would not increase more than necessary, and the applicability would remain similar.

The first major drawback of the traditional STI is that the classical, direct measurement method is not suitable for non-stationary background noise. The indirect measurement method is more robust, but it is unclear under which circumstances performance is optimal. **Chapter 2** focused on the background conditions that were necessary to reliably apply the indirect measurement method. The second major drawback of the traditional STI is that the fluctuating masker benefit is not accounted for. To deal with this shortcoming, an extension of the STI was proposed in **chapter 3**. This extension was based on the calculation and averaging of STI-values for short time windows. The new model was named the Extended STI or ESTI and was evaluated using new and existing sentence intelligibility data in various non-stationary background noises. The ESTI outperformed the classic STI in various noise and reverberation conditions, but intelligibility predictions of speech in noises with low modulation frequencies (< 8 Hz) were still inaccurate. The authors hypothesized that this effect might be related to the context of the speech material, since the probability that meaningful parts in the speech are fully masked by the noise bursts increases under these conditions. To account for this aspect, context was added to the ESTI-model in **chapter 4** and evaluated using existing intelligibility data of meaningful monosyllabic words. In **chapter 5** new intelligibility data was measured in normally hearing subjects using meaningful and nonsense monosyllabic words. This data was then used for the additional evaluation of the context-based ESTI or cESTI. Finally, **chapter 6** focused on the evaluation of the cESTI-model using the sentence material of chapter 3.

CHAPTER 2

TOWARDS MEASURING THE SPEECH TRANSMISSION INDEX IN FLUCTUATING NOISE: ACCURACY AND LIMITATIONS

Van Schoonhoven, J., Rhebergen, K.S., Dreschler, W.A. (2017)
Journal of the Acoustical Society of America 141(2): 818-827

2.1 Abstract

In the field of room acoustics, the modulation transfer function (MTF) can be used to predict speech intelligibility in stationary noise and reverberation and can be expressed in one single value: the Speech Transmission Index (STI). One drawback of the classical STI measurement method is that it is not validated for fluctuating background noise. As opposed to the classical measurement method, the MTF due to reverberation can also be calculated using an impulse response measurement. This indirect method presents an opportunity for STI measurements in fluctuating noise, and a first prerequisite is a reliable impulse response measurement. The conditions under which the impulse response can be measured with sufficient precision were investigated in the current study. Impulse response measurements were conducted using a sweep stimulus. Two experiments are discussed with variable absorption, different levels of stationary and fluctuating background noise, and different sweep levels. Additionally, simulations with different types of fluctuating noise were conducted in an attempt to extrapolate the experimental findings to other acoustical conditions. The experiments and simulations showed that a minimum impulse-to-noise ratio of +25 dB in fluctuating noise was needed.

2.2 Introduction

2.2.1 Background of the original STI

The concept of the modulation transfer function (MTF) in the field of room acoustics was introduced by Houtgast and Steeneken (1973). Since this first publication, the MTF concept has been used to evaluate the speech transmission from a talker to a listener in a room (Houtgast *et al.*, 1980; Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985). The MTF can be used to describe and predict speech intelligibility in stationary noise and reverberation. The acoustical analysis of the MTF is described in detail in Houtgast and Steeneken (1985), Houtgast and Steeneken (2002), Houtgast et al. (1980), and IEC60268-16 (2011). In short, the original and transmitted signals are filtered in seven octave bands (125 – 8000 Hz). The intensity envelopes of the filtered signals are then used to determine the modulation spectrum as a function of 14 modulation frequencies (0.63 – 12.5 Hz, in 1/3 octave bands). The modulation reduction is defined as the ratio between the modulation depth of the input signal and the modulation depth of the output signal. The modulation reduction can be a result of reverberation and/or background noise. A full MTF analysis is based on a 7 × 14 matrix of modulation reduction values. Each value is converted to an apparent signal-to-noise ratio (SNR), after which all values for a given octave band are averaged, clipped (between –15 and 15 dB), and normalized to calculate a modulation transfer index (MTI) for that octave band. A weighted sum of the MTI-values across all octave bands finally results in the Speech Transmission Index (STI): a number between 0 and 1 to indicate the quality of the transmission of the signal.

Houtgast and Steeneken hypothesized that the MTF is related to the Articulation Index or AI (ANSI-S3.5, 1969); today called the Speech intelligibility Index or SII (ANSI-S3.5, 1997). Both indices (AI and SII) reflect the significance of the SNR with respect to speech intelligibility, whereas the STI approach integrates the effects of both noise and reverberation.

The final steps of the STI calculation are equal to the SII scheme. In fact, the ANSI-S3.5 (1997) recommends the use of the MTF in the SII scheme for predicting the speech intelligibility in reverberating conditions (Sec. 5.2). Both methods predict the same "effective audibility" (and therefore the same level of speech intelligibility) in the same listening conditions (Steeneken, 2002). However, the SII is a theoretical model that can only be used if the input speech and noise levels (i.e., SNRs) are known, whereas the STI measuring device can be used to directly measure the MTF with a test a) signal or original speech signal (Payton and Braida, 1999).

The STI was originally validated for normal hearing listeners, but through the years other subjects were studied as well. Duquesnoy and Plomp (1980) found that subjects with different degrees of presbycusis need a higher STI to achieve 50% intelligibility and that the STI remains stable for increasing reverberation times. Plomp and Duquesnoy (1980) further investigated the effect of reverberant conditions on hearing-impaired elderly subjects. They state that, for equal performance, the reverberation time must be decreased by a factor 0.75 – 0.82 per dB deterioration in intelligibility in noise, depending on the proximity of the speaker. Van Wijngaarden *et al.* (2004) concluded that normally hearing listeners need a 1 – 7 dB higher SNR for 50% speech intelligibility in a non-native language, depending on the proficiency in that language.

2.2.2 Indirect measurement of the STI

The MTF can also be calculated indirectly, as opposed to the direct method described above. Schroeder (1978; 1981) described the MTF as a result of reverberation as the Fourier transform of the squared impulse response, normalized by the energy of the squared impulse response [first term between square brackets in Eq. (2-1)]. The MTF as a result of reverberation can thus be characterized completely by the impulse response. In the case of stationary noise and no reverberation, the modulation reduction is described by the SNR only [second term between square brackets of Eq. (2-1)]. Combining the contributions of noise and reverberation therefore leads to the following expression:

$$m(F) = \left[\frac{\left|\int_0^\infty h^2(t)e^{-i2\pi Ft}\mathrm{d}t\right|}{\int_0^\infty h^2(t)\mathrm{d}t}\right]\left[1+10^{-\frac{SNR}{10}}\right]^{-1} \qquad (2\text{-}1)$$

with $\boldsymbol{m}$ as the MTF as a function of the octave band center frequency ($\boldsymbol{F}$) and $\boldsymbol{h}$ as the impulse response as a function of time [$\boldsymbol{t}$, see also (Houtgast *et al.*, 1980; Houtgast and Steeneken, 1985)].

A consequence of this approach is that the indirect method requires two different measurements. The first being the impulse response measurement and the second being the measurement of the background noise at the location of the listener's ear. An impulse response measurement takes a few seconds and the additional noise measurement can take as long as the experimenter desires.

2.2.3 Using the indirect method in fluctuating noise

The original STI concept assumes that the noise is stationary during the measurement. Fluctuations of background noise introduce additional modulations during the STI measurement. As a result, the determined MTF is not solely

based on the modulation reduction due to reverberation and stationary noise, but is also determined by the modulations of the background noise, if it is not stationary. In some cases, this might lead to an *increase* in modulation depth. Direct STI measurements in fluctuating noise can therefore introduce inaccuracies (IEC60268-16, 2011). The risk of fluctuations in the background noise is increased due to the long measurement time of the traditional STI (up to 15 min). In order to reduce these effects, alternatives like the RASTI (Houtgast and Steeneken, 1985) were developed that allow doing a screening in 10 – 15 s. Still, to avoid fluctuating background noise, the STI is often measured in quiet (after work/school time) and afterward the effect of stationary background noise is introduced using a theoretical approach. This method does not account for the effects of the people in the room on the acoustics and does not take into account real life ambient noise. Rhebergen and Versfeld (2005) and Rhebergen *et al.* (2006) extended the Speech Intelligibility Index model (ESII) by calculating the instantaneous SII in time frames. These instantaneous SII-values were then averaged over a certain period of time in order to obtain a single ESII-value. This concept works fine in different types of non-stationary background noise in normal hearing listeners (Rhebergen *et al.*, 2006; 2008) and hearing-impaired listeners (George *et al.*, 2006; Desloge *et al.*, 2010; Rhebergen *et al.*, 2010; Rhebergen *et al.*, 2014). George *et al.* (2010) and George *et al.* (2012) successfully used the theoretical ESII concept to predict the SNR at 50% intelligibility (cSNR) in fluctuating noise and reverberation.

The indirect measurement method offers an opportunity to calculate the STI per time frame, analogous to the ESII. When using this concept, the $\boldsymbol{SNR}$ and $\boldsymbol{m(F)}$ in Eq. (2-1) become time-dependent, and for each time frame an instantaneous value for the STI can be calculated. These values can then be averaged in order to calculate the ESTI (extended Speech Transmission Index). With this approach one can possibly account for fluctuations in background noise.

One prerequisite of using the ESTI method described here is that the impulse response can be measured in fluctuating noise with such precision that the STI can be accurately determined. Hak *et al.* (2012) concluded that an impulse-to-noise ratio (INR) of at least +15 dB is required to reliably measure the STI in stationary noise. The INR (Hak *et al.*, 2008) is defined as the ratio between the peak of the impulse response and the sound pressure level of the background noise. It is unclear whether the criterion of +15 dB holds for fluctuating background noise. We therefore investigated the noise conditions under which the STI can be accurately determined.

2.3 Materials and methods

Two experiments were done in order to investigate under which conditions the impulse response can be reliably measured in order to calculate the STI. Experiment 1 was conducted in a room with variable absorption and different levels of background noise. Experiment 2 was conducted in a room with fixed absorption and background noise level, but with different stimulus levels. The two experiments are described separately. Besides the two experiments, simulations were done to extrapolate the experimental findings to other conditions.

Impulse response measurements were conducted by playing and recording an exponential sweep in different acoustical conditions using Dirac software (version 5, Brüel & Kjær type 7841, Nærum, Denmark), according to ISO3382-1 (2009) and IEC60268-16 (2011). The recorded sweep was then deconvolved with the original sweep in order to obtain the impulse response. All measurements were done in quiet and in stationary and fluctuating noise. The frequency of the sweep is defined by

$$f(t) = f_0 \beta^t \qquad (2\text{-}2)$$

with β = 9.5 and f_0 = 0.1 Hz[ii]

We conducted all sound level measurements using a Brüel & Kjær Sound Level Meter (type 2250, Brüel & Kjær, Nærum, Denmark) at 1.2m from the floor in the center of the room. The sweep was played via a laptop through a JBL Control 2P active, 2-way loudspeaker (JBL, Northridge, US). The noise was played via a separate laptop through a Samson Servo 120a amplifier (Samson Technologies, Hicksville, US) and a Tannoy Reveal passive loudspeaker (Tannoy Ltd., Coatbridge, Scotland). Both loudspeakers have a flat frequency response (+/– 3 dB) between ≤ 100 Hz and 12 kHz, which is sufficient since the STI is calculated using the octave bands between 125 Hz and 8000 Hz (IEC60268-16, 2011). Recordings were done in the center of the room using the same sound level meter via a Brüel & Kjær USB audio interface (ZE 0948, Brüel & Kjær, Nærum, Denmark) on a laptop using Dirac software (version 5, Brüel & Kjær type 7841, Nærum, Denmark) and CoolEdit (version 2000, Adobe Systems, San Jose, US) software. MATLAB (version 2010a, MathWorks Inc., Natick, US) was used for the simulations and Dirac software for the analyses of these simulations. In order to perform the experiments using a realistic approximation of a single talker, the International Speech Test Signal or ISTS (Holube *et al.*, 2010) was used as fluctuating noise. The stationary noise used had the same spectral characteristics as the ISTS. The ISTS is a 60 s long, non-intelligible speech signal created by segmenting and mixing running speech in six different languages. It is shaped according to the Long Term Average Speech Spectrum (Byrne *et al.*, 1994).

ii This low frequency was the default setting in the Dirac software and was therefore also used in the simulations. This means that, although the sweep length was 5.46 s, the relevant STI frequencies were played between 3.0 s and 5.3 s. Consequently, the "effective sweep length" was approximately 2.3 s.

2.3.1 Experiment 1

Experiment 1 was conducted a room of 6.6m long, 6.2m wide, and 4.8m high. The acoustical conditions of the room are summarized in Table 2-1. The absorption properties were altered using absorbing panels and curtains, resulting in conditions A1, A2, and A3 with reverberation times (early decay time, EDT) of 1.6 s, 1.1 s, and 0.6 s (averaged over 500, 1000, and 2000 Hz), respectively. See Bronkhorst and Plomp (1990) for a schematic diagram of the test room. Both loudspeakers were directed at two non-opposing corners of the room and were positioned ~1.5 m from each wall and from the floor, The recording microphone was located in the center of the room and therefore ~2.4 m from the rear of each loudspeaker. An exponential sweep was played in quiet, in stationary noise, and in fluctuating noise. The noise level in the center of the room was 52, 67, or 82 dB (A) for each absorption condition, respectively. The measurement was repeated three times for each condition. Besides this, a 30 s long recording was made of the stationary and fluctuating noises at 67 dB (A) for each reverberation condition. The sweep gain was fixed to yield a level of 85 dB (A) in the most reverberant condition. The sweep level decreased with decreasing reverberation time. In this paper, we use the EDT instead of the T_{30} as outcome measure for the reverberations since the EDT has a higher correlation with the MTF than the T_{30} (Houtgast, 1978).

Table 2-1: Different acoustical conditions of the room. In all conditions measurements were done in quiet. The EDT and T30 noted are the average values, measured in the octave bands with center frequency of 500, 1000 and 2000 Hz. The corresponding STI-values are given in the rightmost column. The standard deviations are indicated between brackets.

Condition	Sweep level	EDT (s) +/– s.d.	T_{30} (s) +/– s.d.	STI +/– s.d.
A1	85 dB (A)	1.6 (0.04)	1.6 (0.14)	0.54 (0.001)
A2	84 dB (A)	1.1 (0.06)	1.1 (0.03)	0.61 (0.002)
A3	81 dB (A)	0.6 (0.06)	0.5 (0.07)	0.75 (0.001)

2.3.2 Experiment 2

Experiment 2 was conducted using nine different sweep levels between 47 and 77 dB (A). The room was 5.8 m long, 5.2 m wide, and 2.7 m high and had a fixed reverberation time (EDT = 0.3 s, averaged over the octave bands with frequency of

500, 1000, and 2000 Hz). Both loudspeakers were directed at two non-opposing corners of the room and were positioned ~1.5 m from each wall and from the floor. The recording microphone was located in the center of the room and therefore ~1.8 m from the rear of each loudspeaker. The sweep was played in quiet, in stationary noise, and in fluctuating noise (same noise conditions as experiment 1). The sound level of the noise at 1 meter from the source was 65 dB (A). The measurement was repeated three or six times depending on the sweep level [six times for the levels between 52 and 62 dB (A)]. Besides this, a 30 s long recording was made of the stationary and fluctuating noise at 65 dB (A).

2.3.3 Simulations

To extrapolate the findings in experiments 1 and 2, simulations were done in MATLAB. This was done by summing a sweep and a noise signal and then bandpass filtering the resulting signal using an eighth-order IIR filter to mimic the loudspeaker response (with 80 Hz and 20 kHz as the –3 dB points). The filtered signal was then convolved with an impulse response in order to simulate the acoustics of the room. The filter and impulse response used here were the same for the sweep and the noise. This does not entirely account for the different positions of the two loudspeakers in the experiments. However, since the microphone was positioned in the center of the room, the positions of the loudspeaker with respect to the microphone were nearly identical. The next step was a deconvolution of the resulting signal with the original sweep signal using circular deconvolution in order to estimate the original impulse response. This estimated impulse response was then used to calculate the STI and the INR, equivalently to the experiments.

The room acoustics and the types of noise were varied (see Table 2-2). In general, four categories of noise were used: stationary, ISTS, interrupted noise, and tonal noise. The interrupted noise was speech-shaped noise, either modulated by a square wave of 2, 4, or 8 Hz, or by a time-scaled maximum-length sequence (MLS) with a value of either zero or unity. The tonal noise was a pure tone carrier wave of 500 Hz, 1 kHz, 2 kHz, or 4 kHz, modulated by a sine wave of 2, 4, or 8 Hz. To approximate a realistic measurement, the starting point of the sweep relative to the noise was randomly varied. To obtain information about different starting points of the sweep, 22 retests per condition were done. The SNR was increased in 5 dB steps from –40 up to +25 dB and –50 and –60 dB were added to account for extremely poor conditions. Synthetic impulse responses were generated by multiplying white noise with an exponential decay envelope[iii] (George *et al.*, 2008). The EDT-values were 0.1 s, 0.2 s, 0.4 s, 1.0 s, and 1.4 s.

iii The EDT and the time constant τ of exponential decay $e^{-t/\tau}$ are related: $EDT = \tau \ln 1000$.

Besides the synthetic impulse responses, three realistic impulse responses were used from experiment 1, with EDT-values of 0.6, 1.1, and 1.6 s. A sampling rate of 44.1 kHz was used in all simulations.

Table 2-2: Different noise types that were used in the simulations. The second column gives the category which will be used to refer to the noises in the remainder of the text

Noise Type	Noise category
Quiet	Quiet
Stationary speech-shaped noise (SSN)	Stationary noise
International Speech Test Signal (ISTS)	ISTS
SSN modulated with a square wave of 2 Hz	Interrupted noise
SSN modulated with a square wave of 4 Hz	Interrupted noise
SSN modulated with a square wave of 8 Hz	Interrupted noise
SSN modulated with MLS with 2 Hz as the dominant modulation frequency	Interrupted noise
SSN modulated with MLS with 4 Hz as the dominant modulation frequency	Interrupted noise
SSN modulated with MLS with 8 Hz as the dominant modulation frequency	Interrupted noise
Pure tone of 500 Hz modulated with a 2 Hz sine wave	Tonal noise
Pure tone of 500 Hz modulated with a 4 Hz sine wave	Tonal noise
Pure tone of 500 Hz modulated with an 8 Hz sine wave	Tonal noise
Pure tone of 1 kHz modulated with a 2 Hz sine wave	Tonal noise
Pure tone of 1 kHz modulated with a 4 Hz sine wave	Tonal noise
Pure tone of 1 kHz modulated with an 8 Hz sine wave	Tonal noise
Pure tone of 2 kHz modulated with a 2 Hz sine wave	Tonal noise
Pure tone of 2 kHz modulated with a 4 Hz sine wave	Tonal noise
Pure tone of 2 kHz modulated with an 8 Hz sine wave	Tonal noise

2.4 Results

The goal of the experiments and simulations was to determine under which conditions the impulse response can be estimated with sufficient accuracy in order to determine the STI. Fig. 2-1 shows the EDT as a function of octave center frequency for acoustical conditions A1, A2, and A3 from experiment 1. The top left panel in Fig. 2-1 shows a clear difference in EDT measured in quiet between the three acoustical conditions in the frequency range between 250 and 4000 Hz. The calculated STI-values in quiet for conditions A1, A2, and A3 are 0.54, 0.61, and 0.75, respectively.

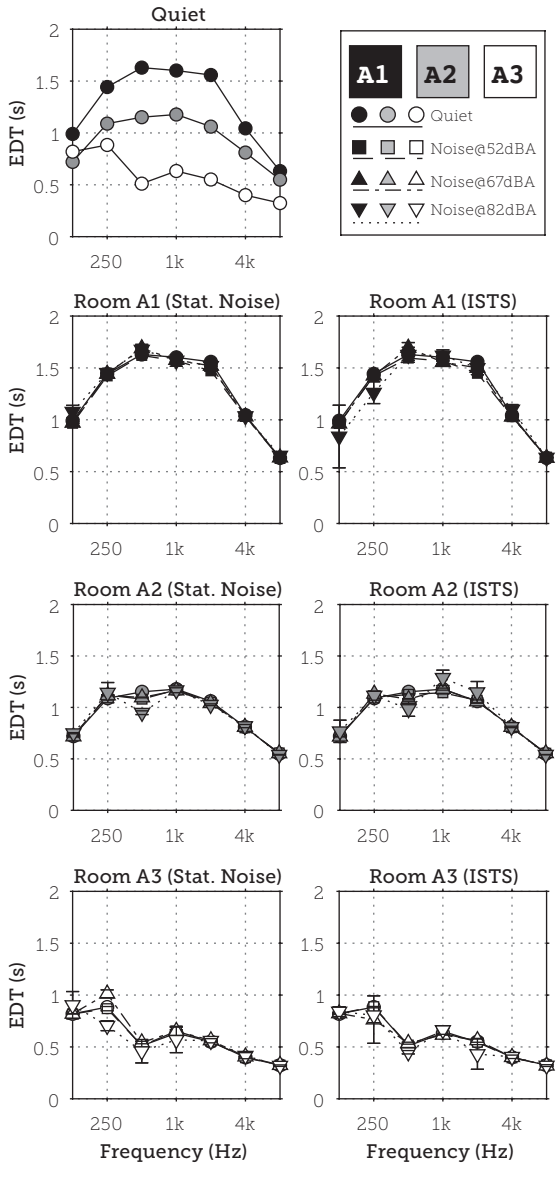


Fig. 2-1: The EDT as a function of the octave bands with center frequency for the different acoustical conditions of the room. (Top) Measurements in quiet for the three reverberation conditions. (Left) Measurements in stationary noise and (Right) measurements in fluctuating noise. The results from the quiet measurements for each reverberation condition are repeated in the corresponding noisy-measurement plots.

The other six panels in Fig. 2-1 show the EDT as a function of center frequency for conditions A1, A2, and A3 measured in quiet, stationary noise, and fluctuating noise with levels of 52, 67, or 82 dB (A). These panels show that EDT measured in stationary and fluctuating noise are closely related to the EDT in quiet.
There are some deviations between the EDT in quiet and in noise but these differences are relatively small. To determine if these differences in EDT have an effect on the STI measurements, we determined the impulse response in all noise conditions and discarded the influence of the noise. Mathematically, this is equivalent to setting the SNR at +∞ such that the second term in Eq. (2-1) is unity and does not affect $m(F)$:

$$m_{SNR=\infty}(F) = \left[\frac{\left|\int_0^\infty h^2(t)e^{-i2\pi Ft}\mathrm{d}t\right|}{\int_0^\infty h^2(t)\mathrm{d}t}\right] \tag{2-3}$$

Using this approach, the STI can be calculated based on reverberation measurements in noise, as if these measurements were done in quiet. Since the actual STI in quiet is known for all acoustical conditions, the ΔSTI can be calculated:

$$\Delta STI = STI_{SNR=\infty} - STI_{quiet} \tag{2-4}$$

According to IEC60268-16 (2011) a STI measurement is reliable when three measurements fall within the range of 0.03 STI units. In the following results, we therefore classify a STI estimation as successful when it deviates 0.015 STI units or less from the true STI in quiet.
The filled symbols in the two top panels of Fig. 2-2 represent the absolute ΔSTI-values of experiments 1 and 2 as a function of the INR. The lines in the graph represent the 95th percentile of these |STI|-values. It can be seen that, for INR-values above +25 dB, the |STI| is smaller than 0.015 STI units. Below an INR of +25 dB the deviation from zero is larger, especially for the measurements conducted in fluctuating noise (ISTS). For stationary noise this INR limit is closer to +15 dB.
The four bottom panels show the |STI| as a function of INR for the simulations. For clarity, only the P_{95} lines are shown. Inspection of the middle two panels shows that, with a few exceptions, the |STI| falls within the range of 0.015 STI units if the INR is larger than +25 dB. This is similar to the results of the experiments. For tonal and interrupted noise (bottom two panels) deviations occur at an INR of around +20 dB, but the pattern is similar to that of the other panels. For stationary noise deviations start occurring at an INR closer to +15 dB, which is also in agreement with the experimental results.

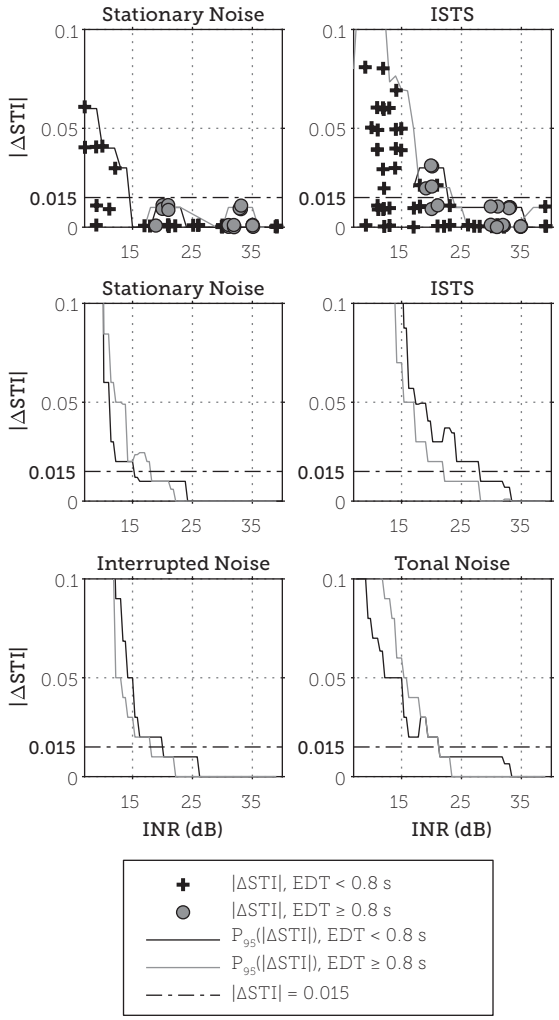


Fig. 2-2: Plot of the |ΔSTI| as a function of the INR. The horizontal dashed-dotted lines represent the bandwidth in which a STI estimation is classified as reliable (< 0.015). The solid lines represent the 95th percentile of the measured data (calculated using a 5 dB window for the experiments and a 2 dB window for the simulations). Due to the large amount of data, the simulation plots only show the P_{95} lines for clarity purposes

The large dataset that was generated using the simulations presents the opportunity to investigate the INR as a classifier of the accuracy of the STI and construct Receiver Operating Characteristic (ROC) curves. A positive result of

the classifier was defined as a |STI|-value larger than 0.015 (i.e., the result is classified as inaccurate). The cutpoint of the classifier (INR) was varied to obtain sensitivity and specificity values, which resulted in ROC curves for the different noise categories (see Fig. 2-3). To identify an appropriate value of the classifier, we set the minimum value of the sensitivity on 99%, allowing 1 out of every 100

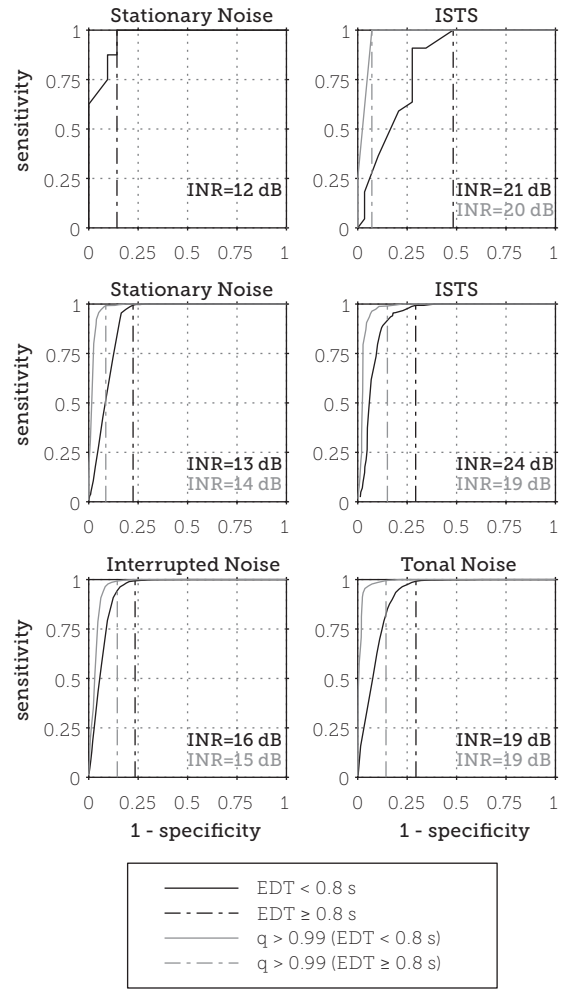


Fig. 2-3: ROC curves with INR as classifier to judge the outcome variable (ΔSTI). There was not enough data available from the experiments to calculate the ROC curve for higher reverberation times in stationary noise. The INRs in the plots represent the values for which the sensitivity is larger than 99%.

correctly identified STI-values to be inaccurate. The authors consider this as an acceptable chance for an error to occur. The INR, sensitivity, and specificity values obtained using this criterion are depicted in Table 2-3. Using this criterion, an INR of +15 dB in stationary noise and of +25 dB in fluctuating noise appear to be suitable indicators for the reliability of the impulse response measurements defined above.

Table 2-3: Required values of the INR using the sensitivity criterion ($q \geq 99\%$). See text for explanation.

	Required INR		Sensitivity (q)		Specificity (p)	
	EDT < 0.8 s	EDT ≥ 0.8 s	EDT < 0.8 s	EDT ≥ 0.8 s	EDT < 0.8 s	EDT ≥ 0.8 s
Stat. noise (exp)	12	-	100%	-	85.7%	-
ISTS (exp)	21	20	100%	100%	51.7%	92.9%
Stat. noise (sim)	13	14	99.4%	99.2%	71.5%	80.9%
ISTS (sim)	24	19	99.3%	99.0%	61.6%	77.3%
Int. noise (sim)	16	15	99.4%	99.3%	66.6%	77.3%
Tonal noise (sim)	19	19	99.2%	99.5%	52.6%	79.0%

The INR itself is a useful criterion during the measurement, but does not provide information about the necessary sweep levels needed in practice. Fig. 2-4 shows the relation between the broadband SNR (in this case the sweep-to-noise ratio) and the INR, averaged over all simulations. Median values and the 95th percentile lines are shown. An SNR of –5 dB (range –15 to +1.7 dB) corresponds to an INR of +15 dB. An SNR of 7.5 dB (range –4 to +15 dB) corresponds to an INR of +25 dB.

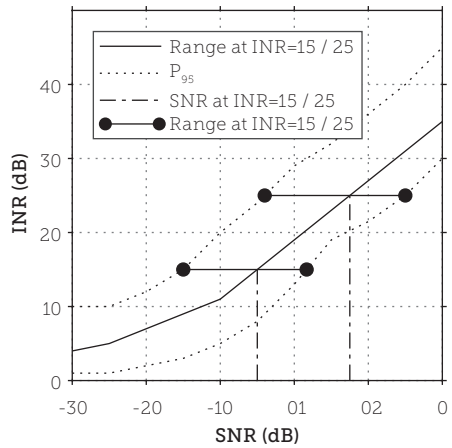


Fig. 2-4: The relation between the INR and the SNR. Median values are shown (solid), together with the 95th percentile lines (dotted). The dashed-dotted lines represent the median SNRs that correspond to an INR of +15 dB and +25 dB. These SNR-values are –5 dB (range –15 dB to +1.7 dB) and +7.5 dB (range –4 dB to +15 dB), respectively

2.5 General discussion and conclusions

The classical method to estimate the STI is widely used to quantify the quality of speech transmission in a certain environment. An inherent problem of this method is measuring the STI in fluctuating noise. The goal of the current study was to determine under which conditions the impulse response can be measured in fluctuating noise in order to make a reliable estimation of the STI. This will facilitate the use of the STI in conditions with fluctuating background noise.
When using a calibrated system during the measurements, it is straightforward to adjust the sweep level in order to accomplish a minimum SNR. In reality, it will occur that the background noise level is unknown. Hak *et al.* (2012) described the calculation of the INR. This is the ratio between the peak level of the impulse response and the noise floor. According to Hak *et al.* (2008) the INR has to be larger than +15 dB in order to reliably calculate the STI in stationary noise. Table 2-3 shows that this number indeed holds for stationary noise, but that this limit shifts to +25 dB for the fluctuating noises used in the current study. Since the INR can always be assessed during the measurements, one can decide during the measurement if the level of the sweep must be increased or not.

The difference in criteria between fluctuating and stationary noise can be explained as follows. When stationary pink noise is used as background sound of the exponential sweep, the SNR in each frequency band is identical, since the frequency spectra of both signals are the same. In the current experiments speech-shaped noise was used, but the differences between frequency bands are still quite small and, more importantly, no significant temporal effects are present. When using fluctuating noise certain frequencies of the sweep coincide with peaks in the noise, which leads to a poorer SNR at those specific frequencies. However, this detrimental effect is not compensated by a positive effect of valleys in the noise that coincide with other frequencies. After all, one can assume that above a certain threshold the reliability is stable and does not increase further with increasing SNR. The same long-term root-mean-square (rms) of the background noise does therefore not lead to equal results for stationary and fluctuating noise.

Various noises were used in the simulations. We chose modulated tonal noise to simulate a condition where all the energy was concentrated in one octave band. The interrupted noises were chosen to simulate a condition with an extremely rapid on- and offset. These conditions will be encountered in real life only in exceptional cases, but were used to mimic extreme measurement conditions. Nonetheless, based on Table 2-3 the ISTS remains the most disadvantageous background noise. Tonal noise led to poor conditions in only one octave band, but the resulting error is largely averaged out in the overall estimation. On the other hand, during the peaks of the interrupted noise energy is distributed over all frequency bands. The SNR per octave band is therefore relatively high, leading to robust measurements at lower SNRs. Resembling natural speech, the intensity peaks of the ISTS vary in frequency content over the course of one measurement, increasing the probability of low SNRs in multiple frequency bands. The peaks and values in the signal lead to the fact that the instantaneous SNR can be much higher or lower than the long-term SNR. The ISTS has therefore the disadvantages of both the amplitude modulated tonal noise and the interrupted broadband noise. This decreases the accuracy of the measurement for noise with the same spectrum as the long-term average spectrum of speech.

In Fig. 2-4 the relation between the INR and broadband SNR is depicted. An INR of +15 corresponds to an SNR of –5 dB and an INR of +25 dB to an SNR of +7.5 dB. Taking into account the P_{95} lines, Fig. 2-4 shows that for 95% of the simulations an INR of +25 dB corresponds to an SNR between –4 dB and +15 dB. In other words, in roughly 97.5% of the cases an INR of +25 dB corresponds to a broadband SNR lower than +15 dB. Picard and Bradley (2001) reviewed several studies that reported about levels of background noise in classrooms.

For traditional classrooms equivalent levels up to 75 dB (A) were measured. Shield and Dockrell (2004) found that the average exposure in London urban area schools was 72 dB (A). They reported large variations in background noise levels between and within schools. The current findings suggest that in these cases a sweep level of 90 dB (A) should be used for a reliable STI measurement in fluctuating noise. This sweep level is quite high and might not be appropriate for daily practice due to hardware limitations and/or subjective factors. The INR of +25 dB, the SNR of +15 dB, and the noise level of 75 dB (A) are safe estimations. Taking into account that the noises in daily life have fewer gaps than the noises used in this study, we expect that in most conditions lower sweep levels can be used. Most importantly, the user gets direct feedback in the form of the INR and can repeat the measurement when the INR is too low.

In the current study, we chose to display the majority of the results as STI measurements. To display more results as reverberation times would have gained the reader more insight in the actual consequences of adding noise to the measurements. However, this was not the purpose of the current paper.

We aimed at doing the experiments with a realistic signal with sufficient fluctuations. Assuming that background noise in, e.g., a classroom is mainly the consequence of talking, a noise with multiple talkers (e.g., babble noise) would be more realistic. However, this noise type has got relatively few temporal gaps and is therefore close to stationary noise. On the other hand, interrupted noise will have larger troughs than a single speaker, but is less often encountered in real-life measurements. The ISTS is a widely available approximation of a single talker, which makes it most suitable for the current study.

The disadvantage of doing real-life experiments is the limited number of conditions one can test. We chose a single fluctuating noise type and several SNRs. Besides this, in experiment 1 the absorption of the room was varied. In order to extend the number of conditions, the simulations were done, which yielded similar results as to the actual experiments. The number of conditions can always be extended by using more types of noise, more SNRs, and more absorption characteristics. However, it is our opinion that the current number of conditions provides enough evidence that the reverberation time can be estimated with sufficient precision in order to make a reliable calculation of the STI in conditions with fluctuating noise. Strictly, the conclusions drawn in this paper are only valid for the noise types used in the experiments and simulations. Currently, the STI has limited value in fluctuating noise and this study is a first step toward applying the STI in non-stationary background conditions. One step further is the application for people with sensorineural hearing loss. Due to degraded temporal and spectral resolution speech reception in noise is worse than in normal hearing subjects, especially in fluctuating background (e.g.,

Festen and Plomp, 1990; Versfeld and Dreschler, 2002). To incorporate this into the ESTI-model and to be able to predict performance of hearing-impaired subjects will therefore be a challenge.
The experiments and simulations that were done in the current study show that the reverberation can be measured with sufficient accuracy in fluctuating noise to reliably calculate the STI. A minimum INR of +25 dB is required. The next step will be to do speech intelligibility measurements under different acoustical conditions and in noise types, calculate the STI per time frame, and investigate the use of the ESTI.

2.6 Acknowledgements

The authors would like to thank Tammo Houtgast and Joost Festen for fruitful discussions, Joost Festen for the use of the variable acoustic room at the VU University Medical Center Amsterdam, Constant Hak and Han Vertegaal from Acoustics Engineering for technical support regarding the B&K Dirac software. J.S. and K.S.R. contributed equally to this work.

CHAPTER 3

THE EXTENDED SPEECH TRANSMISSION INDEX: PREDICTING SPEECH INTELLIGIBILITY IN FLUCTUATING NOISE AND REVERBERANT ROOMS

Van Schoonhoven, J., Rhebergen, K.S., Dreschler, W.A. (2019)
Journal of the Acoustical Society of America 145(3): 1178-1194

3.1 Abstract

The Speech Transmission Index (STI) is used to predict speech intelligibility in noise and reverberant environments. However, measurements and predictions in fluctuating noises lead to inaccuracies. In the current paper, the Extended Speech Transmission Index (ESTI) is presented in order to deal with these shortcomings. Speech intelligibility in normally hearing subjects was measured using stationary and fluctuating maskers. These results served to optimize model parameters. Data from the literature were then used to verify the ESTI-model. Model outcomes were accurate for stationary maskers, maskers with artificial fluctuations, and maskers with real life non-speech modulations. Maskers with speech-like characteristics introduced systematic errors in the model outcomes, probably due to a combination of modulation masking, context effects, and informational masking.

3.2 Introduction

3.2.1 History

The Speech Transmission Index (STI) is used since the 1970s to determine speech transmission quality and to predict speech intelligibility in different acoustic environments (Houtgast and Steeneken, 1973; Houtgast *et al.*, 1980; Houtgast and Steeneken, 1985). The concept is based on the modulation transfer function (MTF), which describes the modulation reduction of a modulated signal caused by noise and/or reverberation. The basic assumption of the model is that reduced modulations in speech lead to deteriorated intelligibility.

Classically, the intensity modulated test signals are filtered into seven octave bands between 125 and 8000 Hz, after which the modulation reduction m is determined for 14 modulation frequency bands between 0.63 and 12.5 Hz per octave band. This results in a 7×14 matrix of m-values. Each m-value is corrected for the hearing threshold and upward spread of masking, after which an apparent signal-to-noise ratio (SNR) is calculated. All apparent SNRs are then clipped between –15 and 15 dB and divided by 30. The resulting transmission indices are then averaged over all modulation frequency bands within each octave frequency band. The weighted sum of these seven mean transmission indices finally results in the STI: a number between 0 and 1, which represents the preservation of modulations in the transmitted signal. A detailed description can be found in Houtgast and Steeneken (2002) and IEC60268-16 (2011).

3.2.2 Fluctuating maskers

Speech intelligibility in fluctuating background noise is usually better when compared to stationary noise with the same long-term root-mean-square (rms) value (e.g., Festen and Plomp, 1990). This is caused by the ability to "listen in the gaps". Depending on the properties of the masker, informational masking (IM) and modulation masking (MM) might counteract this fluctuating masker benefit (FMB). IM occurs when "the signal and masker are both audible but the listener is unable to disentangle the elements of the target signal from a similar-sounding distracter" (Brungart, 2001), and is most prominent when an interfering talker is used as a masker. The effect of MM is smaller and is suggested to be present when the modulation spectra of speech and masker overlap (Stone and Moore, 2004; Apoux and Bacon, 2008).

Also, with overlapping modulation spectra the amount of context in the speech material gets increasingly important. By using redundancies in speech stimuli, listeners are, to a certain extent, able to reconstruct temporally interrupted sentences and/or words (Warren, 1970; Howard-Jones and Rosen, 1993). However, when larger meaningful entities (e.g., words or syllables in a sentence) are masked by

peaks in the noise, this reconstruction process becomes more difficult, especially when the amount of context in the speech material is low (Boothroyd and Nittrouer, 1988). Calculating a long-term STI-value for fluctuating background noise by using the classic STI-method does not account for these phenomena and leads to an inaccurate prediction of speech intelligibility.

To incorporate the FMB in the STI, Bronkhorst and Houtgast (1990) already suggested to average the modulation reduction over a certain time interval instead of averaging the signal and noise intensities. However, an instantaneous calculation of the modulation reduction using the classic STI-method is not possible. Kates (1987) also used a temporal approach by suggesting the short-time articulation index (AI) for adaptive noise reduction systems. Rhebergen and Versfeld (2005) and Rhebergen *et al.* (2006)Rhebergen and Versfeld (2005); Rhebergen *et al.* (2006) calculated the Speech Intelligibility Index (ANSI-S3.5, 1997) as a function of time. The averaged value — the Extended Speech Intelligibility Index (ESII) — was a good predictor for speech intelligibility in fluctuating noise. These model adaptations are all based on audibility (AI and SII) or modulation reduction (STI), and therefore do not account for higher order phenomena such as MM and IM.

Various other models have been proposed to predict speech intelligibility in fluctuating noise. Examples are a binaural version of the SII and ESII (Beutelmann and Brand, 2006; Beutelmann *et al.*, 2010), the ESII with speech input instead of stationary noise (Meyer and Brand, 2013), the multi-resolution speech based envelope power spectrum model or mr-sEPSM (Jørgensen *et al.*, 2013), the general power spectrum model or GPSM (Biberger and Ewert, 2016; 2017), the extended short-time objective intelligibility measure or ESTOI (Taal *et al.*, 2011; Jensen and Taal, 2016), and a model based on automatic speech recognition (Schädler *et al.*, 2015).

Several of these intelligibility models also incorporate the effects of reverberation. The model by Beutelmann *et al.* (2010) accounts for reverberated noise, but not for the detrimental effect of reverberation on the speech signal. The mr-sEPSM (Jørgensen *et al.*, 2013) and GPSM (Biberger and Ewert, 2016; 2017) model the separate effects of fluctuating noise and reverberation on intelligibility, but not the combined effect of both types of distortion. George *et al.* (2008) combined the STI- and ESII-models to predict intelligibility in fluctuating noise and reverberation. They used the STI to estimate the SNR at 50% intelligibility (critical signal-to-noise ratio or cSNR) in stationary noise at a certain reverberation time. Next, the ESII-value at the cSNR was calculated and used to estimate the cSNR in fluctuating noise. To our knowledge, predictions of the combined effects of fluctuating noise and reverberation on speech intelligibility by a single model have not been reported.

3.2.3 Purpose of current study

The classic STI-method can be used in room acoustics to determine speech transmission quality and predict speech intelligibility. However, predictions of intelligibility in fluctuating noise are beyond the scope of the current standard (IEC60268-16, 2011). This limitation of the classic STI gives rise to problems during measurements in realistic environments (e.g., classrooms or office floors) since many background noises are non-stationary.

In the current study the Extended Speech Transmission Index (ESTI) is presented. The proposed approach is similar to the ESII method as introduced by Rhebergen and Versfeld (2005) and Rhebergen *et al.* (2006), and therefore primarily deals with the FMB. The ESTI aims to increase the applicability of the STI. However, the effects of IM, MM, and context are still not fully covered. This will be addressed in the discussion section. Since the focus of this study is on extension of the original STI-method, no comparisons were made between the ESTI and other existing models.

In the current study, speech intelligibility in stationary and fluctuating maskers was measured in five different reverberant conditions. An 8 Hz interrupted noise (IN8) and a speech-like noise were used as fluctuating maskers. A stationary masker served as the reference condition. The model parameters were adjusted to accurately model intelligibility in these conditions. Finally, data from other studies were used to test the validity of the model more thoroughly.

3.3 Materials and methods

Two datasets are used in the current paper. To optimize parameters of the ESTI-model, speech intelligibility measurements were done (see section 3.3.1). Throughout the paper, $\boldsymbol{dataset01}$ will be used to refer to these data. To test the model, data from the literature were used and will be referred to as $\boldsymbol{dataset02}$.

3.3.1 Speech intelligibility measurements

3.3.1.1 Subjects

Nine normally hearing subjects were recruited (three males and six females) with mean age 22.1 yr. (range 18 – 32 yr.). All were native Dutch speakers and no hearing or language problems were reported. All subjects had pure tone thresholds of 20 dB HL (hearing level) or better at the octave frequencies between 250 and 4000 Hz.

Subjects were recruited via posters. They gave written informed consent and received compensation for participating. Approval for the project (NL48348.018.14) was given by the Ethical Review Board (Medisch Ethische Toetsingscommisie, Amsterdam Medical Centre).

3.3.1.2 Stimuli

The target speech consisted of Dutch sentences, uttered by a female speaker (Versfeld *et al.*, 2000). The speech material consisted of 39 lists of 13 sentences each. Each sentence contained between four and nine words, and between seven and ten syllables. For the practice trials, sentences from Plomp and Mimpen (1979) were used. Three types of maskers were used. Stationary speech-shaped noise (SSN) was used as the reference condition. The IN8 was created by modulating the SSN with an 8 Hz square wave with modulation depth of 100% and a duty cycle of 50%. The third masker was the International Speech Test Signal [ISTS; (Holube *et al.*, 2010)]. The ISTS is an unintelligible speech-like signal, created by segmenting and mixing female speech in six different languages. As opposed to the other noises, the frequency spectrum of the ISTS is different from that of the target speech (see Fig. 3-1).

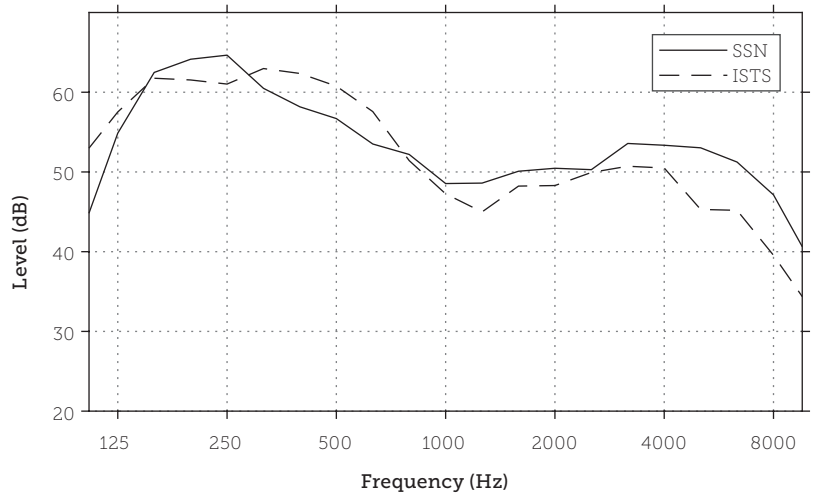


Fig. 3-1: Frequency spectrum in 1/3 octave bands of SSN, based on the Dutch female sentences (Versfeld et al., 2000), and ISTS (Holube et al., 2010) at 65 dB (A)].

3.3.1.3 Reverberation

Four degrees of reverberation were used: T_{60} = 0.1, 0.4, 0.8, and 1.2 s. Four artificial impulse responses were created by multiplying exponential decay envelopes with white noise[iv] (George *et al.*, 2008). The slope of the envelope

iv All impulse responses with T_{60} > 0.1 s had a white spectrum between 125 and 8000 Hz. The impulse response with T_{60} = 0.1 s had a white spectrum above 500 Hz, but had ~3 dB more energy at 125 and 250 Hz.

corresponded to the desired reverberation times as mentioned above. Speech and noise signals were convolved separately with one of these four impulse responses, depending on the test condition. Fig. 3-2 shows the effect of reverberation on the two fluctuating noises in the temporal domain. Fig. 3-3 shows the effect in the modulation domain. Although modulations appear to have disappeared for high reverberation times using IN8 in Fig. 3-2 (left panel), the 8 Hz modulations are still present.

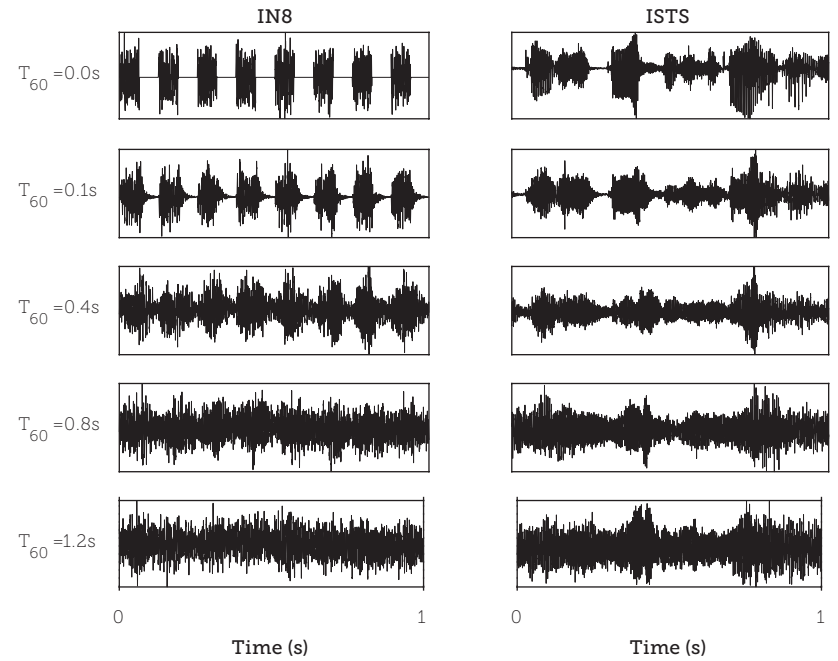


Fig. 3-2: One second depiction of IN8 in the left panel and the ISTS in the right panel (Holube et al., 2010) with different degrees of reverberation.

3.3.1.4 Procedure

The SNR that was required for 50% intelligibility of complete sentences was measured using an adaptive procedure as described by Plomp and Mimpen (1979). This will be referred to as the cSNR. The first sentence of each list was presented at an SNR of –10 dB (for SSN) or –20 dB (for IN8 and ISTS) with the noise level fixed at 65 dB (A)[v]. This sentence was repeatedly presented, each

v The original publication by Van Schoonhoven et al. (2019) erroneously stated that the speech level was fixed. All calculations were done correctly using a fixed noise level. This error is corrected in the current text.

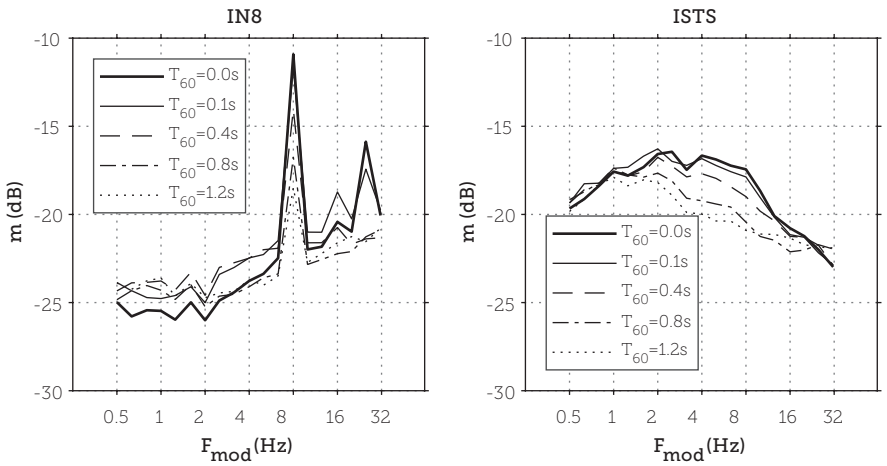


Fig. 3-3: Modulation spectra of IN8 in the left panel and the ISTS in the right panel (Holube et al., 2010) with different degrees of reverberation.

time with a 4 dB higher SNR than the previous presentation. After the first correct answer, the step size was changed to 2 dB. From this point onward, all remaining sentences of the list were presented only once. The SNR of the next sentence was increased by 2 dB after an incorrect response and decreased by 2 dB after a correct response. Each sentence was used once per subject and the whole sentence had to be repeated correctly for the answer to be correct. The cSNR was calculated by averaging the SNRs of sentences 5 – 14 of one list during one trial (the 14th sentence was not actually presented, but its SNR was calculated based on the previous answer). All individual data points were an average between test and retest for the SSN and IN8 conditions. No retest was done for the ISTS condition.

The total experiment was preceded by one practice trial with reverberation and one practice trial without reverberation. All reverberation conditions using one noise type (e.g., SSN) were presented within one block. The sequence of these three main blocks (blocks A, B, and C in Fig. 3-4) was randomized across subjects. Each main block was preceded by one practice trial without reverberation and one practice trial with reverberation. The sequence of the test- (and retest-) trials was balanced and pseudo-randomized across blocks and subjects using Latin squares (Wagenaar, 1969). Subjects were allowed a 5-min break each 20 min. The total visit time was between 2 and 2.5 h. See Fig. 3-4 for a visual representation of the test conditions.

The long-term rms of all noises was based on the SNR required for that presentation. This implicates that the peaks of the IN8 masker were 3 dB higher

than the long-term rms of the SSN masker at the same SNR. Speech and noise were scaled after reverberation. Signals were presented monaurally to the right ear through a TDH39P headphone (Telephonics Corp., Farmingdale, NY) via a 24 bit/192 kHz Fireface 800 audio interface (RME, Haimhausen, Germany). Subjects were seated in a sound treated booth. MATLAB (version 2017b, MathWorks Inc., Natick, MA) was used for presentation of the sounds and analysis of the results. A sampling frequency of 44.1 kHz and a bit depth of 16 bits/sample were used for all signals.

3.3.2 ESTI-model

The ESTI was calculated analogous to the ESII (Rhebergen *et al.*, 2006). To account for reverberation, both the noise and speech signal were convolved using the appropriate impulse response. The speech, noise, and impulse response were then filtered using an octave filter bank [sixth-order class 1, (ANSI-S1.11, 2004)] between 125 and 8000 Hz. Analyses were done per octave band q (with q = 1, ..., 7).

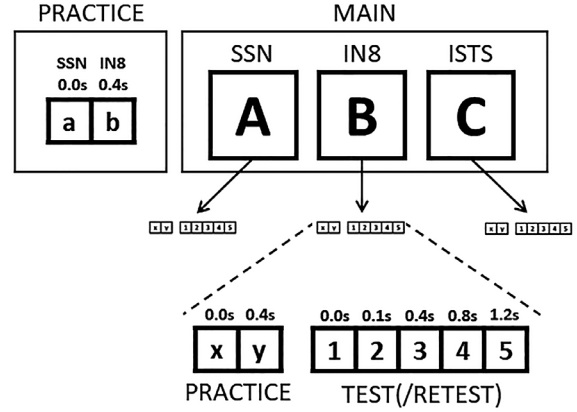


Fig. 3-4: Visual representation of the measurement conditions. The practice trials a and b preceded the total experiment. The main section was divided into three blocks: A, B, and C, based on noise type. Within each block a practice trial without reverberation (x) and a practice trial with reverberation (y) were presented. Conditions 1 – 5 were presented twice in blocks A and B (test and retest) and once in block C (test only). The sequence of blocks A, B, and C was randomized across subjects. The sequence of conditions 1 – 5 (including retest conditions 1' – 5' for A and B) was pseudo-randomized across blocks and across subjects using Latin squares (Wagenaar, 1969).

3.3.2.1 Time window

The indirect measurement method according to Schroeder (1981) makes it possible to calculate the contribution of the SNR and reverberation to the STI separately. In the classic STI, the temporal characteristics of the speech and noise are discarded since only the long-term averaged spectra are used in the calculations. The current ESTI-model was developed to account for noise fluctuations. Therefore, the noise level is calculated as a function of time. To limit deviations from the classic STI, a stationary test signal was currently used as speech input for the model. For an octave band filtered signal $\boldsymbol{s(t, q)}$ of length $\boldsymbol{T}$ this results in:

$$S(q) = 10\log_{10} s^2(q)_{rms} = 10\log_{10}\sqrt{\frac{1}{T}\int_0^T s^2(t,q)\,dt} \quad (3\text{-}1)$$

Using a sliding rectangular time window, the rms-value of the fluctuating noise $\boldsymbol{n(t, q)}$ with length $\boldsymbol{T}$ was calculated as a function of time:

$$N(p,q) = 10\log_{10}\sqrt{\frac{1}{T_w}\int_{p\tau}^{p\tau+T_w} n^2(t,q)\,dt} \quad (3\text{-}2)$$

with $\boldsymbol{T_w}$ as the time window length, $\boldsymbol{s}$ as the step size, and $\boldsymbol{p}$ as indices for the time windows. Both a fixed window length and a frequency dependent window length were tested. A fixed window with a length of 2.0 ms resulted in the best fit of $\boldsymbol{dataset01}$ (see section 3.5 for more information regarding this choice). However, since this window length is shorter than the period in the octave bands with lower center frequencies, adjustments were made to ensure a minimum of one period per time window for all frequencies within each band (see Table 3-1). The step size $\boldsymbol{s}$ was 2.0 ms for all octave bands.

Table 3-1: Time window lengths per octave frequency band, tested in the optimization phase. The step size corresponded to the window length of the highest octave band. The minimum window length always ensured a minimum of one period being present for the lowest frequency in that octave band [indicated by *; according to Van Schijndel et al. (1999)]. ‡ indicates time windows based on Rhebergen and Versfeld (2005). § indicates time windows based on Shailer and Moore (1983). A fixed time window length of 2.0 ms was eventually chosen (indicated by †).

Frequency (Hz)	125	250	500	1000	2000	4000	8000
Frequency dependent‡	37.4	22.2	17.2	15.0	14.7	10.0	9.4
Frequency dependent§	27.2	20.5	14.4	8.0	6.0	4.5	3.0
Fixed (1.0 ms)	11.3*	5.6*	2.8*	1.4*	1.0	1.0	1.0
Fixed (2.0 ms)†	11.3*	5.6*	2.8*	2.0	2.0	2.0	2.0
Fixed (4.0 ms)	11.3*	5.6*	4.0	4.0	4.0	4.0	4.0

3.3.2.2 Forward masking

Various studies described the masking of a target signal by a preceding masker (e.g., Duifhuis, 1973; Moore and Glasberg, 1983; Gifford *et al.*, 2007; Fogerty *et al.*, 2017). When a fluctuating masker is used, forward masking plays a role, especially when the offsets are abrupt (Rhebergen *et al.*, 2006). The model of Ludvigsen (1985) was used to incorporate forward masking in the masker signal [see Eq. (3-3) for the general relationship]. It describes how the masked threshold ($\boldsymbol{MT}$) decreases exponentially as a function of the time after the masker is interrupted (post-masker duration or $\boldsymbol{t_{pm}}$). $\boldsymbol{MT}$ is also a function of post-masker time ($\boldsymbol{T_0}$), recovery time ($\boldsymbol{T_f}$), and hearing threshold ($\boldsymbol{HTL}$). Values of 1 ms and 150 ms were used for $\boldsymbol{T_0}$ and $\boldsymbol{T_f}$, respectively (see section 3.5 for more information regarding this choice).

$$MT\left(p, \frac{t_{pm}}{\tau}\right) = N(p) - \frac{\log\left(t_{pm}/T_0\right)}{\log\left(T_f/T_0\right)} \times \left[N(p) - HTL\right] \quad (3\text{-}3)$$

with $T_0 \le t_{pm} \le T_f$

The value of $\boldsymbol{MT}$ between $\boldsymbol{t_{pm}}$ = 0 and $\boldsymbol{t_{pm}} = \boldsymbol{T_0}$ is equal to $\boldsymbol{N(p)}$. For each time window $\boldsymbol{p}$ the $\boldsymbol{MT}$ as a function of the post-masker duration was determined. $\boldsymbol{MT}$ was then compared to $\boldsymbol{N}$ between $\boldsymbol{p}$ and $p + \frac{Tf}{\tau}$. This has the following implications for the effective masking noise:

$$N_{eff}\left(p + \frac{t_{pm}}{\tau}\right) = \max\left\{N\left(p + \frac{t_{pm}}{\tau}\right), MT\left(p, \frac{t_{pm}}{\tau}\right)\right\} \quad (3\text{-}4)$$

This calculation was done for each time window $\boldsymbol{p}$ and each octave band $\boldsymbol{q}$ ($\boldsymbol{q}$ was omitted in the above equations for clarity purposes). The effect of the noise is constant for all modulation frequency bands $\boldsymbol{r}$.

3.3.2.3 MTF

The modulation reduction is a result of both reverberation and noise and is expressed in the MTF. Based on the relation described by Schroeder (1981) the Fourier transform of the impulse response was used to calculate the reverberation component of the $\boldsymbol{MTF}$: $\boldsymbol{MTF_{rev}}$. The acoustic conditions do not change over time, and therefore $\boldsymbol{MTF_{rev}}$ is only dependent on the modulation frequency band $\boldsymbol{r}$ and octave frequency band $\boldsymbol{q}$. Van Wijngaarden and Houtgast (2004) suggested that calculation of the STI was more accurate for conversational speech when modulation frequencies up to 31.5 Hz were incorporated. This resulted in 18 modulation frequency bands ranging between 0.63 and 31.5 Hz

with 1/3 octave intervals. MTF_{rev} is calculated according to Eq. (3-5), where h is the impulse response, BW_r is the bandwidth of the modulation frequency band r, and $F_{r,l}$ and $F_{r,u}$ are the lower and upper boundaries, respectively, of the modulation frequency band r. The nested fraction was taken from Schroeder (1981) and represents the MTF. Summation over the modulation frequencies yields the MTF as a function of the modulation frequency band. This rectangular filter shape was chosen to limit the number of adjustable model parameters during optimization.

$$MTF_{rev}(q,r) = \frac{1}{BW_r} \int_{F_{r,l}}^{F_{r,u}} \left[\frac{\left| \int_0^\infty h^2(t,q) e^{-i2\pi F t} \mathrm{d}t \right|}{\int_0^\infty h^2(t,q) \mathrm{d}t} \right] dF \tag{3-5}$$

The noise component of the MTF is MTF_{SNR}. In the classic STI, this part is independent of time. However, in the ESTI the SNR is calculated per time window, and therefore the following relation applies [note that $S(p)$ is a stationary signal and can be considered as a constant along the q-axis]:

$$MTF_{SNR}(p,q) = \left[1 + 10^{\frac{1}{10}\left(N_{eff}(p,q) - S(p)\right)} \right]^{-1} \tag{3-6}$$

The total MTF is the product of the two components:

$$MTF(p,q,r) = \prod_{p,q,r} \left[MTF_{SNR}(p,q)\, MTF_{rev}(q,r) \right] \tag{3-7}$$

Like the original STI, the current model assumes that the effect of the noise on modulation reduction is uniform across all modulation frequencies. However, this only applies within each time window, which is a relatively short time scale compared to the modulation frequencies of interest. Therefore, all important masker modulations are preserved in the MTF.
Note that, as a consequence of the aforementioned assumption, MTF_{SNR} can be considered as a constant along the r-axis [Eq. (3-6)]. Similarly, MTF_{rev} is independent of time [Eq. (3-5)].

3.3.2.4 ESTI calculation

The resulting MTF is a function of time window number (p), octave frequency band (q), and modulation frequency band (r). The subsequent calculation of the STI-value per time window is now equivalent to the classic STI calculation (Houtgast *et al.*, 1980). The MTF is corrected for upward spread of masking and the hearing threshold. Upward spread of masking of octave band q by octave

band q-1 is modeled with a level dependent slope of masking. See Table A.1 in IEC60268-16 (2011) for the values used here. Next, the apparent SNR is calculated:

$$SNR_{app}(p,q,r) = 10 \log_{10} \frac{MTF(p,q,r)}{1 - MTF(p,q,r)} \tag{3-8}$$

The apparent SNR is clipped between –15 and +15 dB, and used to calculate the modulation transfer index (MTI):

$$MTI(p,q) = \frac{1}{R} \sum_r \frac{SNR_{app}(p,q,r) + 15}{30} \tag{3-9}$$

A weighted sum of the modulation indices for all octave bands then results in the STI per time window:

$$\begin{aligned} STI(p) = {} & \left(\alpha_1 MTI(p,1) - \beta_1 \sqrt{MTI(p,1) MTI(p,2)} \right) \\ & + \left(\alpha_2 MTI(p,2) - \beta_2 \sqrt{MTI(p,2) MTI(p,3)} \right) \\ & + \cdots + \alpha_7 MTI(p,7) \end{aligned} \tag{3-10}$$

with α_q and β_q as the frequency dependent octave-weighting factor and redundancy correction factor per octave band, respectively. The values for α_q and β_q as suggested by the IEC standard, IEC60268-16 (2011), in Table A.3 were used in the current study. The final step is calculation of the ESTI:

$$ESTI = \frac{1}{P} \sum_p STI(p) \tag{3-11}$$

3.3.3 Optimization and validation of the model

The SSN masker without reverberation was chosen as the reference condition. The ESTI in the reference condition was calculated at the cSNR, and was used to predict the cSNR in the other noise and reverberation conditions. These model predictions were compared to the observed cSNR-values.
$Dataset01$ was used to optimize the new model parameters regarding time averaging and forward masking. $Dataset02$ was used to compare predictions of the optimized model to intelligibility data from the existing literature.

3.3.3.1 Experimental data measured in the current study (dataset01)

With the extension of the STI, two model parameters were added: time window length (T_w) and forward masking time (T_f). We varied these parameters and calculated the model predictions for each possible combination, and compared these to the observed cSNR-values. The optimal linear fit was calculated to

determine the accuracy of the predicted cSNR-values for each noise type. The combination of parameter values that resulted in the highest coefficient of determination (R^2) was eventually chosen. Forward masking times between 0 and 400 ms with 50 ms steps were used during the optimization. Both fixed and frequency dependent time windows were used (see Table 3-1).

3.3.3.2 Experimental data derived from literature (dataset02)

Next, speech reception data using Dutch speech from existing literature in combination with the current data were used as input for the ESTI-model. This is referred to as $dataset02$. As in section 3.3.3.1, the predictions were also based on the reference condition (SSN masker without reverberation). The forward masking time and time window length that resulted in the best predictions in section 3.3.3.1 were used. The goal of this last step was to test the optimized model using more conditions than were used in the current speech intelligibility experiments. All data that were included can be found in Appendix B.

3.4 Results

3.4.1 Speech intelligibility measurements

The SNRs that were required for 50% intelligibility of complete sentences (cSNR) are depicted in Table 3-2. A paired t-test (using Bonferroni correction) was performed to test for differences between the fluctuating maskers (IN8 and ISTS) and the SSN. The cSNR in SSN without reverberation is –3.4 dB, which is relatively high compared to Rhebergen *et al.* (2006), who reported a value of –5.5 dB, but similar to Versfeld *et al.* (2000) and Francart *et al.* (2011), who reported cSNR-values of –4.11 dB and –3.6 dB, respectively.

As expected, the introduction of gaps using the IN8 masker led to release of masking, resulting in better performance. The average improvement is 11.6 dB, which is similar to the benefit of 12.1 dB reported by Rhebergen *et al.* (2006). The release of masking when using the ISTS masker is limited to an improvement of 3.6 dB and is not significant. Francart *et al.* (2011) reported an improvement of 1.9 dB using the ISTS in comparison to SSN (not tested for significance). The standard deviation for fluctuating maskers is larger than for the stationary masker, especially for the ISTS.

Intelligibility deteriorates when reverberation is introduced. Deterioration is largest using the IN8 masker since reverberant energy fills the gaps between the noise blocks most effectively. Although 8 Hz modulations appear in the IN8 modulation spectrum until T_{60} reaches values as high as 5 s (see Fig. 3-3), the benefit of listening in the gaps has disappeared at T_{60} = 0.4 s. Remarkably, when

T_{60} = 1.2 s, the SSN masker leads to better intelligibility than both fluctuating maskers ($p < 0.05$). See Table 3-2 and Fig. 3-5.

Table 3-2: Mean cSNR of sentences of nine subjects. The standard deviation is in brackets. (†) indicates the conditions that were tested with eight subjects. Significant differences per column between the fluctuating maskers and SSN are indicated with *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

cSNR (dB)	T_{60} = 0.0 s	T_{60} = 0.1 s	T_{60} = 0.4 s	T_{60} = 0.8 s	T_{60} = 1.2 s
SSN	-3.4 (1.0)	-2.5 (0.9)	1.2 (0.9)	5.3 (2.4)	7.2 (1.0)
IN8	-15.0*** (1.8)	-6.2*** (1.6)	1.5 (1.1)	5.8 (2.4)	10.1* (1.9)
ISTS	-7.0 (3.3)	-3.1 (1.9)	0.8 (1.4)	8.6 (3.0)†	12.0* (3.2)†

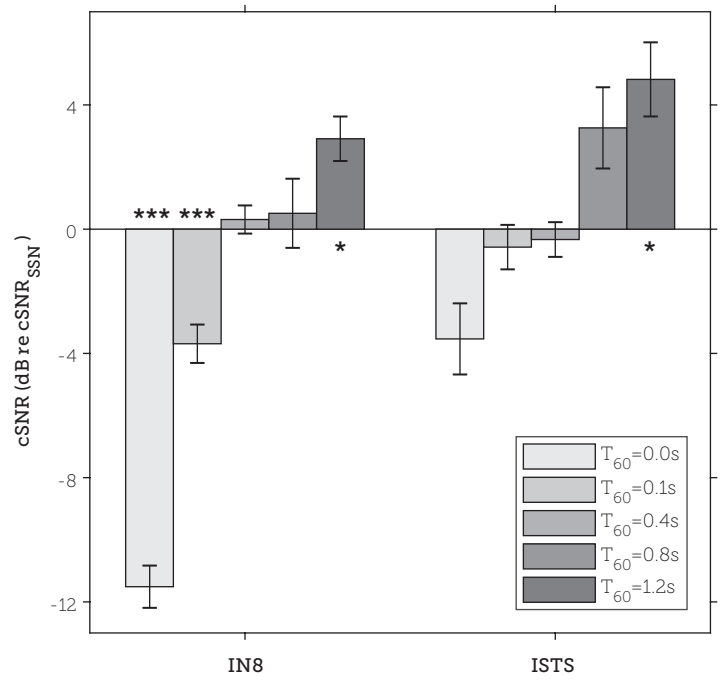


Fig. 3-5: Difference in cSNR for the SSN relative to the fluctuating maskers (IN8 and the ISTS) for different reverberation times. Vertical bars represent the standard deviation of the difference. Significant differences between the fluctuating maskers and SSN are indicated with ***($p < 0.001$), **($p < 0.01$), and *($p < 0.05$).

3.4.2 ESTI-model

The underlying assumption of many speech intelligibility models is that a transfer function exists between model outcome and performance. Examples are the AI (Fletcher and Galt, 1950; Pavlovic, 1984; Studebaker *et al.*, 1993), the ESII (Rhebergen and Versfeld, 2005), and the STI (Houtgast *et al.*, 1980; Steeneken and Houtgast, 2002; IEC60268-16, 2011). In the current study, speech intelligibility was measured using the cSNR so only one point of the transfer function between ESTI and performance was known. According to the model, a certain ESTI-value is needed for 50% intelligibility, independent of reverberation or noise type. Therefore, once the ESTI-value for the cSNR in the reference condition is known, the cSNR for other conditions can be predicted. However, this approach does not yield information about the complete transfer function between ESTI and intelligibility. The SSN masker without reverberation was chosen as the reference condition since this is the most basic test condition and was available in all studies.

This approach is demonstrated in Fig. 3-6 by the iso-ESTI contours for the three maskers SSN, IN8, and ISTS (Houtgast and Steeneken, 1985). The cSNR for SSN without reverberation is –3.4 dB, which corresponds to an ESTI-value of 0.411. Each point on all three curves in the plot corresponds to the same ESTI-value. As reverberation increases, a higher SNR is required to reach the same ESTI-value. At low reverberation times the FMB leads to a lower cSNR for the fluctuating maskers than for the SSN masker. At high reverberation times the

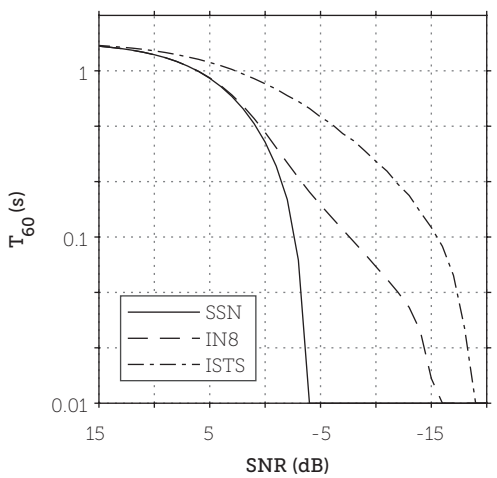


Fig. 3-6: iso-ESTI contour for an ESTI-value of 0.411, which is the ESTI at cSNR for the SSN masker in this example. All points on all three curves correspond to this ESTI-value.

gaps in the fluctuating noises are smeared, and the ESTI is dominated by the reverberation. This causes the three curves to converge.

For $dataset01$, model calculations were done per individual subject. These calculations served to optimize the model. Since no individual data from the existing literature were available, model verification using $dataset02$ was done with group average cSNRs.

3.4.2.1 Experimental data measured in the current study

In Table 3-3 and Fig. 3-7 both the classic STI-values and ESTI-values are depicted for $dataset01$. These are the (E)STI-values at cSNR (see Table 3-2). A forward masking time (T_f) of 150 ms and a fixed time window (T_w) of 2.0 ms (see Table 3-1) resulted in the best model predictions and were used to calculate all ESTI-values.

Table 3-3: ESTI- and classic STI-values for data from $dataset01$ at cSNR.

	T_{60} = 0.0 s	T_{60} = 0.1 s	T_{60} = 0.4 s	T_{60} = 0.8 s	T_{60} = 1.2 s
SSN (classic STI)	0.387	0.393	0.424	0.431	0.399
SSN (ESTI)	0.411	0.416	0.439	0.437	0.402
IN8 (ESTI)	0.419	0.440	0.459	0.446	0.427
ISTS (ESTI)	0.675	0.651	0.558	0.506	0.444

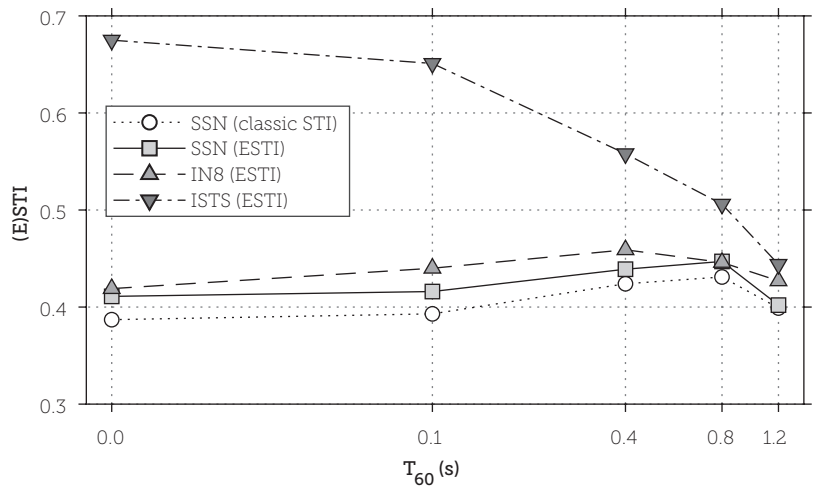


Fig. 3-7: (E)STI-values at cSNR of the data of the current study as a function of reverberation time (T_{60})

Ideally, the (E)STI-values would be independent of noise type and reverberation time. The (E)STI-values for SSN and IN8 show a similar pattern, but the values are relatively high compared to other studies [e.g., the value of 0.37 was used by George *et al.* (2008)]. This topic will be addressed in section 3.5. ESTI-values for the ISTS masker are even higher, especially for lower reverberation times. This might be the effect of MM, context effects (CE), and/or IM. This will also be addressed in section 3.5.

Fig. 3-8 shows individual cSNR predictions. As mentioned earlier, the SSN masker without reverberation served as the reference condition. The ESTI-value found for this condition was used to predict the cSNRs for the other conditions.

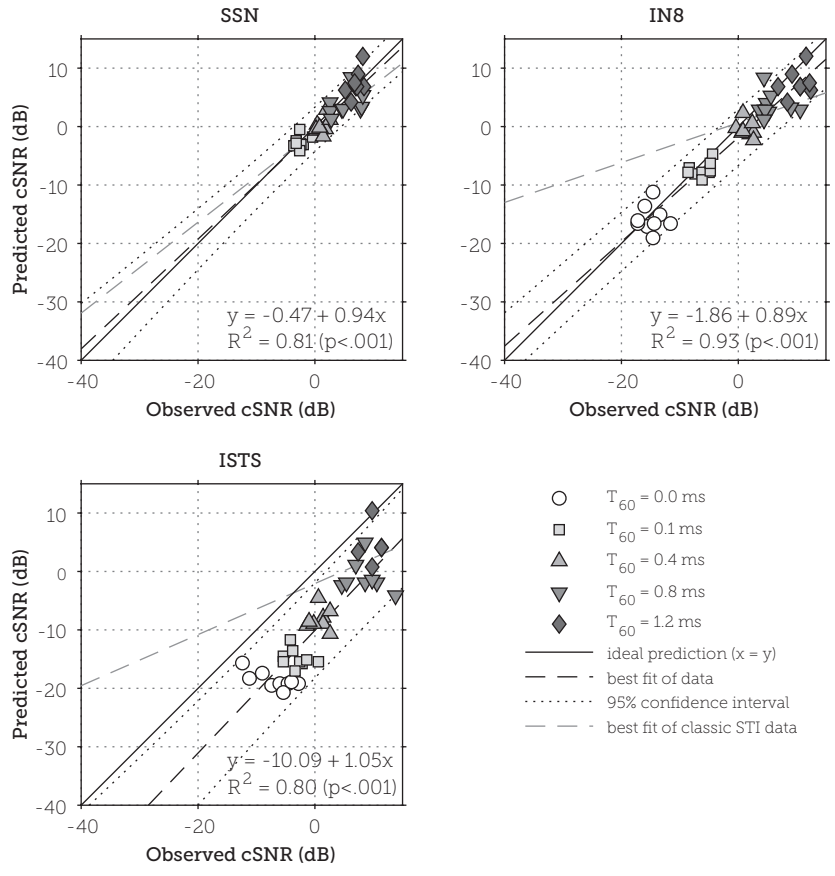


Fig. 3-8: All individual cSNR predictions of ***dataset01***, separated based on masker. The coefficient of determination (R^2) was based on the best linear fit of the data per panel. Note that no values are plotted for SSN, T_{60} = 0.0 s, since this condition served as reference condition.

The explained variance by the best linear fit is 81%, 93%, and 80% for SSN, IN8, and ISTS, respectively. In Fig. 3-8 it can be seen that the best linear fit for SSN and IN8 is close to the main diagonal ($y = x$). The slope of the best linear fit for the ISTS masker is close to 1, but a systematic overestimation of about 10 dB is seen. This might be attributable to MM, CE, and/or IM and will be addressed in section 3.5. Also, the best linear fit using the classic STI-model is shown (gray, dashed line). This fit is similar to the ESTI fit for SSN, but clearly deviates for the IN8 masker and to a lesser extent for the ISTS masker. This deviation reflects the fact that listening in the gaps is not accounted for in the classic STI-model.

3.4.2.2 Experimental data derived from literature

Fig. 3-9 and Table B-1 (see Appendix B) show ***dataset02***, including cSNR calculations based on the ESTI-model. The ESTI at cSNR in the reference condition (SSN masker without reverberation) was calculated first. This ESTI-value was then used to

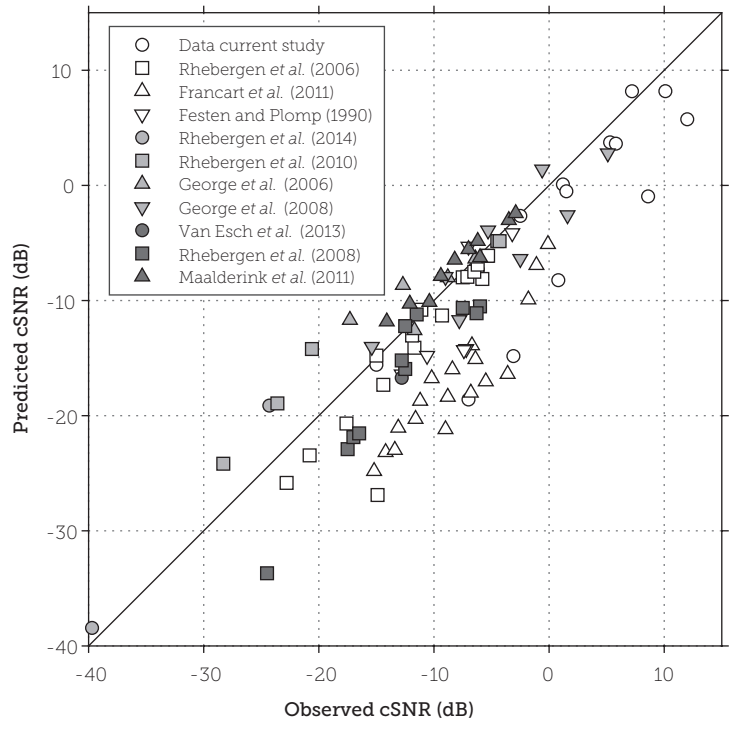


Fig. 3-9: cSNR predictions based on the ESTI of ***dataset02*** depicted in Table B-1 (see Appendix B). Predictions were based on the ESTI-value in stationary noise without reverberation for that specific study.

predict the cSNR for the other noise and reverberation conditions of that specific study. In Fig. 3-10, these data are separated based on the type of fluctuations (stationary noises, artificial fluctuations, and speech-like fluctuations) and type of fine structure (artificial fine structure and speech-like fine structure).

Fig. 3-10 also shows the R^2-values based on the least squares linear fit. For stationary noises with artificial fine structure, 95% of the observed variance is explained by the best linear fit. For noises with artificial fluctuations and artificial fine structure, this value is 87%. When maskers have speech-like fluctuations, the values of R^2 are similar, but the best fit clearly deviates from the optimal model prediction. The slope of the best fit is in these cases close to unity, but the larger intercept points toward a systematic error of the model. This error is close to 3 dB for artificial fine structure and little under 10 dB for speech-like fine structure. When the slope of the best fit is forced to unity, the explained variance remains similar (76% and 92%) as well as the intercept (4.0 dB and 9.5 dB). Again, in all panels the best linear fit of the classic STI is also depicted. The deviations from the optimal fit for artificial and speech-like fluctuations are similar as in Fig. 3-8.

The bottom left panel in Fig. 3-10 shows the data points of the multitalker babble maskers that were reported by Francart *et al.* (2011) and Rhebergen *et al.* (2008). For this condition there are few data points with little variance, which poses difficulties in the calculation of the optimal fit. This explains the large confidence interval.

The best linear fit was also calculated for seven datapoints using real life non-speech maskers (Rhebergen *et al.*, 2008). Examples are music and construction noise. These data points are not shown in Fig. 3-10. The R^2 for the best linear fit was 0.90 ($p < 0.01$). When forcing the slope to unity, the explained variance was 77% with an intercept of 1.2 dB.

3.5 General discussion

The current study introduced the ESTI in order to deal with one of the primary limitations of the classic STI: prediction of speech intelligibility in fluctuating noise. Monaural speech intelligibility experiments using fluctuating and stationary maskers, with and without reverberation were done on normally hearing subjects in order to improve the model. Additionally, existing data from the literature were used to test the model.

It was not the intention of the authors to provide the reader with an elaborate comparison of models, a theoretical framework on central auditory processing,

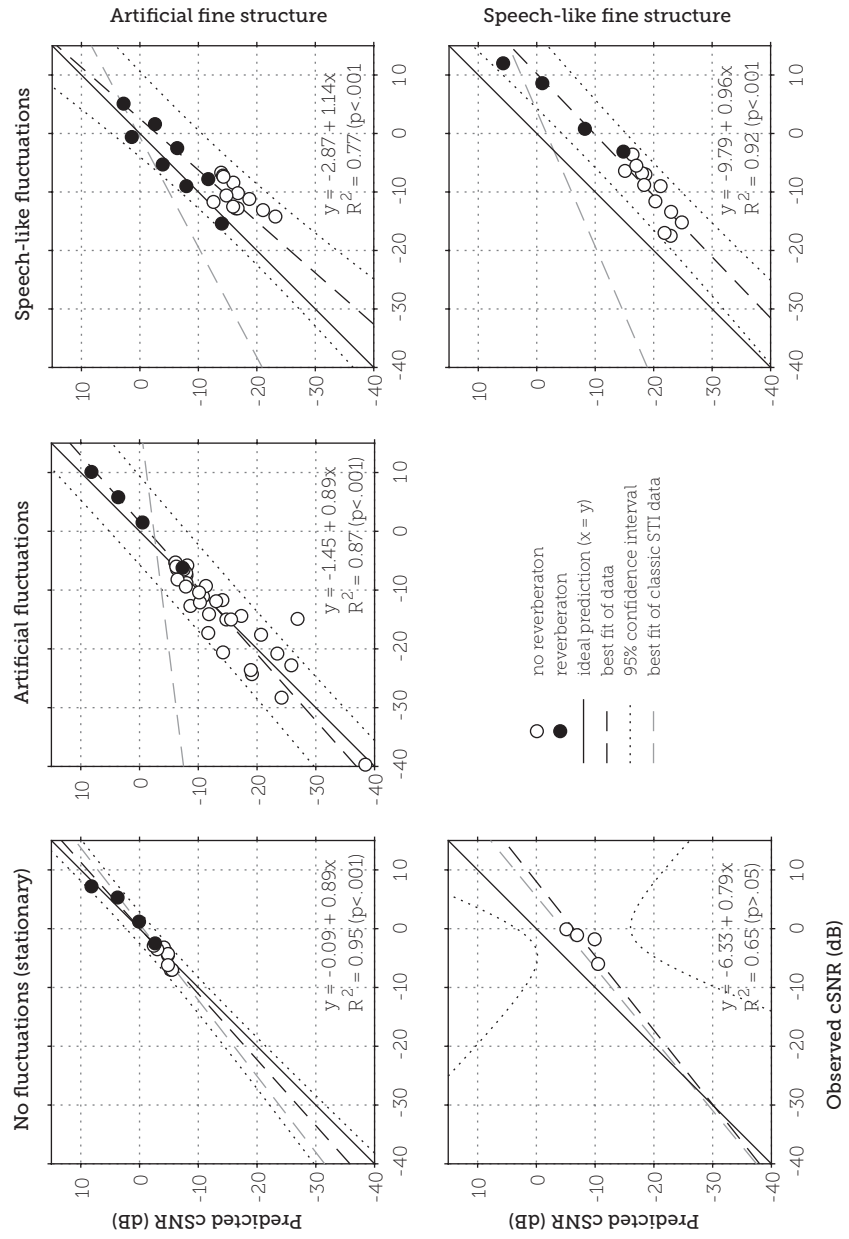


Fig. 3-10: All data from ***dataset02*** separated based on fluctuation type (columns) and fine structure (rows) as follows: artificial fine structure (top row), speech-like fine structure (bottom row), stationary masker (left column), maskers with artificial fluctuations (middle column), and maskers with speech-like fluctuations (right column). Filled symbols represent conditions with reverberation. The coefficient of determination (R^2) was based on the best linear fit of the data per panel.

or an extensive review on different forms of masking. It was the goal to improve the classic STI-model based on the data measured in the current study and validate this approach using literature data.

3.5.1 Speech intelligibility measurements

As expected, intelligibility improved when gaps were introduced in the noise. However, with high reverberation times (T_{60} = 1.2 s) there is a significant disadvantage when listening in fluctuating noise in comparison to the stationary masker. This is true for both IN8 and ISTS.

It was reported earlier that, although the ISTS is unintelligible, some form of IM is introduced by the masker because of the speech-like characteristics (Holube *et al.*, 2010; Francart *et al.*, 2011). They tested without reverberation and reported that the masker can be distracting and might draw unwanted attention from the listener. The speech-like characteristics of the ISTS masker lead to higher salience than, for instance, a stationary masker, and might therefore lead to difficulties in object selection (Shinn-Cunningham, 2008). IM is not directly obvious, since the listener also benefits from the spectro-temporal gaps in the masker, leading to a net improvement relative to the SSN masker. Francart *et al.* (2011) broke down the factors that contributed to the difference in cSNRs between SSN and the ISTS. They suggested a 7.5 dB advantage due to dip listening, a 2.1 dB disadvantage due to spectral differences between masker and target, and another 4.6 dB disadvantage due to IM. Since the modulation spectra of the ISTS and speech signal are similar, MM might also play a role.

In the current study, the detrimental effect of reverberation (T_{60} = 1.2 s) on the cSNR using SSN is 10.6 dB (see Table 3-2). The effects of reverberation on the masker itself are negligible, so this difference is primarily the result of reverberating the speech signal. On the contrary, the detrimental effect of reverberation with the ISTS masker is 19.0 dB. Under the assumption that reverberation has little effect on the amount of IM and/or MM caused by the ISTS, the extra 8.4 dB disadvantage as a result of reverberation is caused by smearing of the spectro-temporal gaps. This value is of similar magnitude as the 7.5 dB found by Francart *et al.* (2011). However, as can be seen in Fig. 3-3 and Fig. 3-11, modulations are still present in the ISTS after reverberation. It is therefore possible that complete smearing of the spectro-temporal gaps would lead to an even larger disadvantage.

Since IN8 does not contain any speech-like modulations or fine structures, it is not likely that IM plays a role when using this masker. For T_{60} = 1.2 s subjects perform 2.9 dB worse with the IN8 masker relative to the SSN masker ($p < 0.05$). One possibility is that the modulations that are still present in the noise (see Fig. 3-3) do not provide the listener with any temporal gaps that are deep enough for

dip listening but do cause MM (e.g., Kwon and Turner, 2001; Fogerty *et al.*, 2016). In Fig. 3-11 it can be seen that the distribution of local SNRs in SSN and IN8 is similar for T_{60} = 1.2 s. The possibility of MM will be addressed in the discussion of the ESTI-model below.

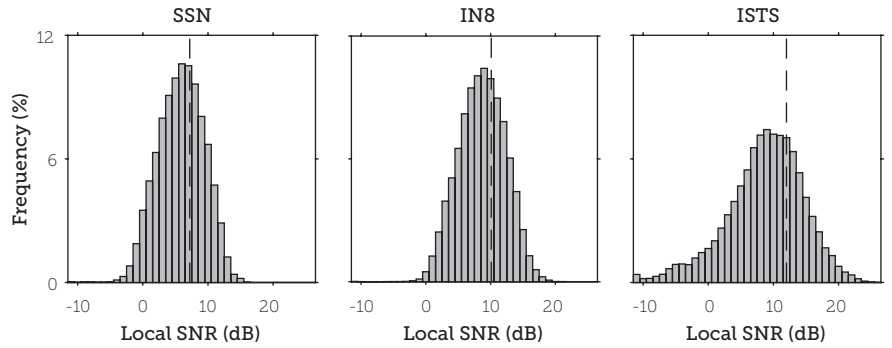


Fig. 3-11: Distribution of local SNRs at the (long-term) cSNR for the different maskers at T_{60} = 1.2 s. The histograms were calculated using a sliding 2 ms time window. A noise segment of 10 s was used and a constant speech level was assumed. Bins were 1 dB wide. The dashed vertical line represents the long-term cSNR.

A limitation of the current experiments is the monaural presentation of speech and noise. Binaural effects play an important role in speech reception in noise and reverberation, but were not incorporated in the current model. Van Wijngaarden and Drullman (2008) proposed a binaural version of the classic STI, which might serve as a basis for a binaural version of the ESTI for fluctuating maskers. They used an interaural cross-correlogram for three octave bands as a front-end for their STI calculations. Another option to incorporate binaural effects is the equalization and cancellation front-end as proposed by Beutelmann and Brand (2006) and Beutelmann *et al.* (2009; 2010).

3.5.2 The ESTI-model

The ESTI-values for SSN in Table 3-3 and Fig. 3-7 are higher than the classic STI-values. This is caused by random fluctuations in the SSN. The long-term SNR of the speech and noise signal forms the basis of the classic STI. On the contrary, the proposed ESTI is defined as the average value of all instantaneous STI-values, which are based on the short-term SNR-values. Each short-term SNR-value is based on the long-term rms of the speech signal [Eq. (3-1)] and the short-term rms of the noise signal [Eq. (3-2)].

The order of calculations has an effect on the eventual (E)STI-value, especially when short time windows (< 50 ms) are used, since random fluctuations are not averaged out. Using the ESTI, the logarithm of the instantaneous sound energy is taken before averaging [see Eqs. (3-7) – (3-11)]. As a consequence, peaks in the noise will be less dominant than when the logarithm is taken after averaging, as is the case for the classic STI. When peaks in the noise are less dominant, the instantaneous SNRs will tend to be higher than the long-term SNR and therefore lead to a higher ESTI. However, the primary goal of the ESTI is not to reach similarity to the classic STI per se. The objective is rather to predict the cSNR for fluctuating and stationary noises.

In the current analyses, the different maskers were characterized based on the type of fine structure and types of fluctuations. It should be noted that signal types cannot be described fully by these characteristics. Other factors also play a role, like the similarities between F0 of masker and target signals, the modulation spectrum of the masker, and the intelligibility of the masker.

3.5.2.1 Parameter estimation

In the ESTI-model, the original STI parameters were left intact [according to IEC60268-16 (2011)]. As mentioned in section 3.2, the time window length (T_w) and forward masking time (T_f) were introduced to account for the calculations over time. The influence of these parameters on the accuracy of the ESTI-model was studied.

The effect of T_f on the explained variance of the model (R^2) using the SSN and ISTS maskers was small. These maskers contain no or few abrupt offsets, which limits the effect of forward masking (Schlauch *et al.*, 2001; Rhebergen *et al.*, 2005). An effect of T_f on the IN8 masker is seen in Fig. 3-12, where the current value (T_f = 150 ms) was compared to the value of 200 ms as proposed by Ludvigsen (1985).The model results were more accurate using the value of 150 ms than for lower or higher values. Omitting forward masking (T_f = 0 ms) or using higher values for T_f (e.g., 400 ms) results in a drop in explained variance. This is shown in Fig. 3-12 for the default time window length of 2 ms. When analyzing the interrupted maskers in the literature data ($dataset02$) with frequencies between 8 and 128 Hz, the R^2-values are 0.80 and 0.86 for T_f = 200 ms and T_f = 150 ms, respectively. So, the value of 150 ms also leads to better model performance for other interruption rates than 8 Hz.

The effect of T_w on IN8 is also clear from Fig. 3-12. The explained variance drops dramatically from 16 ms onward. Again, this is related to the interruption frequency. When performing the same analysis for $dataset02$ (not shown), this drop is present from 8 ms onward due to the inclusion of higher interruption frequencies. The R^2-value for shorter time windows (1, 2, and 4 ms) is relatively

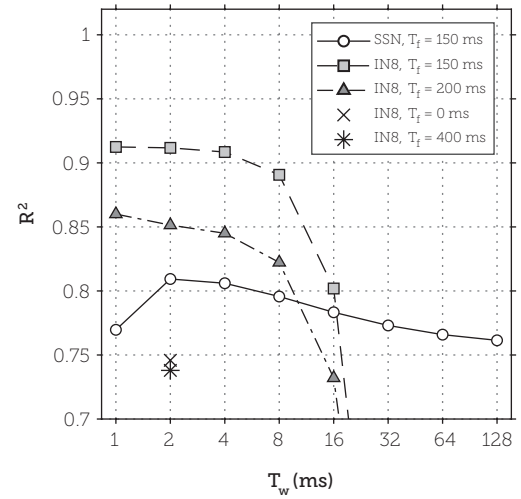


Fig. 3-12: Explained variance (R^2) of the ESTI-model as a function of the time window length (T_w) when the slope of the linear fit is forced to unity ($dataset01$). Values are depicted for various forward masking times and two maskers (SSN and IN8). For clarity purposes, the R^2-values for T_f = 0 ms and T_f = 400 ms were only shown for T_w = 2 ms. The effects are similar for other values of T_w. There was no effect of T_f on the SSN masker, so these data are omitted. R^2 drops to 0.6 when T_f is increased to 800 ms (not shown). Note that not all values of T_w were used in the optimization phase of the model, but are shown here for clarity purposes.

constant. In Fig. 3-12 the R^2-values for SSN as a function of T_w are also depicted. A gradual decrease with increasing window length is seen here. The lower R^2-value for T_w = 1 ms is related to the post-masker time T_0 of the forward masking model.

T_0 was defined by Ludvigsen (1985)as the time after interruption of the masker that the MT remains constant. After T_0, the exponentially decaying forward masking function starts [see Eq. (3-3)]. When longer time windows are used (T_w = 2 ms), fast fluctuations in the SSN are less prominent due to averaging. However, when T_w = 1 ms the short peaks in the masker are artificially kept at a high level during T_0.This decreases model accuracy for increasing values of T_0 at T_w = 1 ms.

In the mr-sEPSM by Jørgensen *et al.* (2013) the signal is segmented using rectangular time windows after modulation filtering. They chose a modulation frequency dependent window length equal to the inverse of the center frequency of the corresponding modulation filter. This means that the window lengths ranged between 3.9 and 1000 ms. In the current ESTI-model, time

windows of these lengths negatively affected model accuracy for the interrupted noises. In Table B-1 data from maskers with interruption frequencies of 32 and 64 Hz are presented (Rhebergen *et al.*, 2006). The masking release observed in these conditions is 5.6 dB and 2.0 dB. Data from Miller and Licklider (1950) also show that listeners still obtain benefit from 100 Hz interrupted noise as compared to stationary noise. In these examples, the gaps in the noise are equal to 15.6 ms (32 Hz), 7.8 ms (64 Hz), and 5 ms (100 Hz). In the current model, long time window lengths as used by Jørgensen *et al.* (2013) would smear out these fast interruptions, which does not allow accurate modeling of the observed benefit using interrupted maskers.

3.5.2.2 MM

For some fluctuating maskers that were investigated, envelope similarities between speech and masker exist, which might have contributed to MM. According to Fogerty *et al.* (2016), similarities in modulation rates in target and masker might limit the effect of masking release. Shinn-Cunningham (2008) argues that the more unique and distinct the target signal is, the better the masker signal is suppressed due to more robust object selection. Based on this statement, maskers with speech-like fluctuations and artificial fine structure will be less prone to IM than actual speech maskers like ISTS since no fine structure is present. Model discrepancies using modulated maskers without fine structure might be mostly related to MM. However, since IM can occur when target and masker have similar higher order features (Shinn-Cunningham, 2008), IM cannot be completely ruled out due to the envelope similarities. An average 3 – 4 dB error of our model was seen for these maskers (top right panel in Fig. 3-10). As a first-order approach, this value could be used as a correction factor in our model to account for MM. This is in line with the 3 – 4 dB as suggested by Schubotz *et al.* (2016), based on the ESII predictions. This also corresponds to the 2.9 dB disadvantage of the IN8 masker as compared to the SSN masker using high reverberation times, which was mentioned earlier (based on $dataset01$). However, the effect of differences and similarities between target and masker envelopes must be studied more thoroughly.

3.5.2.3 CE

The current model shows a discrepancy when masker fluctuations are speech-like. As previously mentioned, this might be related to MM, but can also be due to the effect of context. The dominant modulation frequency in speech of 3 – 4 Hz is related to the number of syllables per second (Houtgast *et al.*, 1980). Listening in the gaps at higher masker modulation rates (> 6 – 8 Hz) therefore gives the listener multiple "looks" per syllable, which increases the likelihood of

correctly repeating the entire target signal. This is the result of perceptual restoration (Warren, 1970; Saija *et al.*, 2014). When the dominant modulation frequency of the masker decreases, the listener might only get one look per syllable, or none at all, decreasing the possibilities of perceptual restoration. The dominant modulation frequency of speech-like maskers is around 3 – 4 Hz, which limits the number of looks per syllable and increases the likelihood of meaningful parts in the speech being masked.

There is an outlier in the central panel in Fig. 3-10 where intelligibility is overestimated by more than 10 dB. Although this concerns noise with artificial fluctuations (interrupted noise), the interruption frequency of 4 Hz is close to the dominant modulation frequency of running speech (Houtgast *et al.*, 1980). How well the listener can use perceptual restoration depends on the context of the speech material used (Miller *et al.*, 1951; Boothroyd and Nittrouer, 1988; Bronkhorst *et al.*, 1993). Boothroyd and Nittrouer (1988) compared intelligibility of high predictable sentences (semantic and syntactic context), low predictable sentences (syntactic context), and unpredictable sentences (no context, apart from coarticulatory cues). Predictable sentences resulted in better intelligibility. Several models exist to account for this effect of context (Boothroyd and Nittrouer, 1988; Bronkhorst *et al.*, 1993). Incorporation of these models in the ESTI-model might lead to increased model accuracy.

CE and MM are fundamentally different phenomena but play a role under the same circumstances: envelope similarities between target and masker. In CE, top-down processes play an important role, since the listener can combine fragments of speech with contextual information to infer the complete utterance. This is related to vocabulary, linguistic skills, and knowledge of the subject (Boothroyd and Nittrouer, 1988; Benoit, 1990; Bronkhorst *et al.*, 1993). On the other hand, MM might be related to bottom-up processes like masking of specific modulation rate channels as described by Dau *et al.* (1997a). Based on the current results, it is not possible to distinguish between the two phenomena.

3.5.2.4 IM

Rosen *et al.* (2013) tested speech reception using different numbers of talkers in a multitalker babble masker. They also used the noise vocoded version of the babbles, and the envelope of the babbles to modulate Gaussian noise. For all number of talkers, the babble noise was the most effective masker. They found that the benefit of the gaps disappeared for four or more talkers, but also stated that this strongly depends on the speech material used. Since true babble noise remained the most effective masker for 8 and 16 talkers, speech-like characteristics like the fine structure probably affected intelligibility. This is in line with the effect of reverberation using the ISTS masker in the current study since

spectro-temporal gaps decrease while the fine structure is largely maintained. The bottom right panel of Fig. 3-10 shows that the intercept of the best linear fit of the data deviates 9 – 10 dB from the optimal model prediction for maskers with a speech-like fine structure. Envelope similarities between target and masker also play a part when using these maskers, so it is likely that MM is involved. A 3 – 4 dB correction factor as a first-order approach to deal with MM was suggested earlier. The temporal fine structure in ISTS and other speech-like maskers therefore lead to an extra systematic error of 5 – 7 dB. Competing speech signals produce IM (Durlach *et al.*, 2003; Durlach, 2006; Holube *et al.*, 2010), so this error can be used as a first-order correction of the ESTI-model for IM. The 5 – 7 dB suggested here is similar to the 6 – 7 dB that was mentioned by Rhebergen *et al.* (2005) and Schubotz *et al.* (2016), based on ESII predictions. Again, the differences and similarities between signals must be studied more thoroughly in order to confirm or refute this approach.

3.5.2.5 General limitations

Although MM, CE, and IM play an important part in speech perception in fluctuating noise, they are currently not part of the ESTI-model. And, as already mentioned, there is overlap between the factors discussed above. IM is dominant when the masker signal is competing speech, but the envelope similarities between masker and target suggest that MM is also involved. These effects can be corrected using empirical correction factors of 3 – 4 dB for MM and 5 – 7 dB for IM, but more complex interactions may be involved. Furthermore, meaningful entities in the speech are more likely to be masked when the envelopes of masker and target are similar. Therefore, both IM and MM might also interact with CE. How these phenomena influence each other and are affected by bottom-up and top-down processes remains open to debate.
The correction factors for IM and MM are currently only valid for 50% speech reception since the full psychometric curve is not known. Furthermore, it is a first-order approach to deal with maskers with speech-like characteristics. In practice, the experimenter who conducts the ESTI measurement can use the noise recording to analyze the noise. When the modulation spectrum is speech-like, the correction for MM can be applied. If the fine structure is speech-like, the correction for IM can be applied. However, this is difficult to judge, especially when background noises are not actual speech but only have some speech-like characteristics. In these cases, it is not clear which correction value needs to be applied. When the interactions between different aspects of speech and maskers are better understood, as well as their influence on the model, the above corrections can be implemented in an algorithm with a solid evidence-based foundation.

Like the classic STI-model, the current ESTI-model only uses modulation reduction due to noise and reverberation in order to predict speech reception. It is therefore a relatively simple approach for the complex problem of modeling speech intelligibility. When encountering higher order problems in fluctuating noises like MM, CE, and IM, our model [and other existing models; see Schubotz *et al.* (2016)] fails to predict intelligibility accurately. This is due to the complexity of higher auditory and cognitive processes that play a part in speech perception and our incomplete understanding thereof. Also, interindividual differences that are independent of the hearing threshold get increasingly important in complex listening environments. Examples are linguistic skills (Brouwer *et al.*, 2012), cognitive capabilities (Koelewijn *et al.*, 2012), and working memory (Zekveld *et al.*, 2013).
STI measurements are classically done using intensity modulated test signals. One of the drawbacks of this method is the reduced measurement accuracy in fluctuating background noise. The indirect measurement method [based on Schroeder (1981)] overcomes this problem. An impulse response measurement is needed to estimate the contribution of reverberation to the MTF. A separate recording of the background noise is used to calculate the modulation reduction due to noise. Van Schoonhoven *et al.* (2017) described the conditions under which the impulse response can be measured reliably in fluctuating background noise in order to calculate the STI. This indirect measurement method can also be applied when using the ESTI.

3.6 Conclusions

We presented the ESTI-model as an extension of the classic STI in order to predict speech intelligibility in fluctuating noises and reverberant environments. The validation data presented in the current paper in combination with the indirect measurement method led to a broader applicability of the ESTI in room acoustics. Intelligibility in noises with speech-like modulations and, to a lesser extent, speech-like fine structure, is still systematically underestimated. This is probably caused by a combination of MM, CE, and IM. The next step is to investigate the role of these aspects and how to incorporate this into the ESTI-model.

3.7 Acknowledgements

The authors would like to thank Jan Verhave of Embedded Acoustics for providing the MATLAB scripts to calculate the STI. They would also like to thank the two anonymous reviewers for their detailed and elaborate comments.

CHAPTER 4

A CONTEXT-BASED APPROACH TO PREDICT SPEECH INTELLIGIBILITY IN INTERRUPTED NOISE: MODEL DESIGN

Van Schoonhoven, J., Rhebergen, K.S., Dreschler, W.A. (2022)
Journal of the Acoustical Society of America 151(2): 1404-1415

4.1 Abstract

The Extended Speech Transmission Index (ESTI) by Van Schoonhoven *et al.* (2019) was used successfully to predict intelligibility of sentences in fluctuating background noise. However, prediction accuracy was poor when the modulation frequency of the masker was low (< 8 Hz). In the current paper, the ESTI was calculated per phoneme to estimate phoneme intelligibility. In the next step, the ESTI-model was combined with one of two context models (Boothroyd and Nittrouer, 1988; Bronkhorst *et al.*, 1993) in order to improve model predictions. This approach was validated using interrupted speech data, after which it was used to predict speech intelligibility of words in interrupted noise. Model predictions improved using this new method, especially for maskers with interruption rates below 5 Hz. Calculating the ESTI at phoneme level combined with a context model is therefore a viable option to improve prediction accuracy.

4.2 Introduction

4.2.1 Modelling speech intelligibility

Numerous methods exist to model speech intelligibility. The Articulation Index (AI) by Fletcher and Galt (1950) was the first detailed analysis of speech that made prediction of intelligibility of nonsense words possible. French and Steinberg (1947) related the signal-to-noise ratio (SNR) to the AI using a linear relation. The AI was adjusted over the years and is now known as the Speech Intelligibility Index (SII) (ANSI-S3.5, 1997). It is a model which is primarily based on audibility. Several other models which are based on the SII were developed in the following years (e.g., Rhebergen and Versfeld, 2005; Beutelmann and Brand, 2006; Meyer and Brand, 2013).

Another group of models is primarily based on the detection and processing of speech modulations. The first model that used this approach was the Speech Transmission Index (STI) (e.g., Houtgast and Steeneken, 1973; Steeneken and Houtgast, 2002; IEC60268-16, 2011). It can be used to analyze modulation reduction in the speech due to noise and/or reverberation to obtain an index between 0 and 1. If a transfer function (TF) for a certain speech corpus is known, this STI-value can be used to predict intelligibility. Several other models have been constructed that use speech modulations to predict intelligibility (e.g., Jørgensen *et al.*, 2013; Biberger and Ewert, 2016; Jensen and Taal, 2016).

The original STI is only valid when the background noise is stationary. For applications in fluctuating background noise, Van Schoonhoven *et al.* (2019) introduced the Extended Speech Transmission Index (ESTI). They calculated the STI for short time windows as opposed to one long-term STI-value. In order to do so, the noise signal was filtered in octave bands, after which the root-mean-square (rms) values per time window were calculated. These sliding rectangular time windows had a length of 11.3 ms for the lowest octave band and 2 ms for the highest. Especially when fluctuations in the noise are sudden, forward masking plays a role, which was introduced in the ESTI-model based on Ludvigsen (1985). The Modulation Transfer Function (MTF) as a result of the noise was then calculated for each time window separately. The MTF based on the reverberation was derived separately based on Schroeder (1981). The product of both MTFs was then used to calculate a STI-value per time window according to the original STI method described in IEC60268-16 (2011). Basically, the original STI was calculated for each point in time, as if the noise characteristics at that moment had a continuous character. These STI-values were averaged to calculate the ESTI. The only addition to the original STI, besides the temporal approach, was the introduction of forward masking. A detailed description can be found in Van Schoonhoven *et al.* (2019).

This extension of the STI-model provided the opportunity to calculate local STI-values for speech in fluctuating maskers. The long-term average of these local STI-values served as an accurate predictor of the intelligibility of Dutch sentences using fluctuating maskers with higher (≥ 8 Hz) modulation frequencies. However, at lower modulation frequencies, the probability increased that complete meaningful elements were masked during the noise peaks. Under these conditions, the ESTI-model failed to accurately predict sentence intelligibility. In other words, a prerequisite of the current ESTI-model is that noise modulations must be fast enough to guarantee glimpses at all meaningful elements. When this condition is not met, the ESTI overestimates intelligibility.

To deal with this limitation, the current study focused on calculating the ESTI-value for each individual element (e.g., each syllable in a sentence, or each phoneme in a word), instead of for the whole speech token. A TF was needed to convert the ESTI per element to the intelligibility per element. The next step was to combine the intelligibility of all elements, at which point the model needed to account for elements that were completely masked. To achieve this goal, the effect of context was added to the model. This approach was based on the method suggested by Bronkhorst *et al.* (1993) in their Sec. III C, where their context model was used to predict intelligibility of interrupted speech, based on the perceived fraction of the speech signal. The current study used a comparable approach, only now applied to speech in fluctuating noise.

4.2.2 Context effects

Miller *et al.* (1951) studied speech intelligibility in noise as a function of context. The intelligibility of words in meaningful sentences was higher than of words in isolation. When one word in a sentence is missed, the number of possible answers is restricted as a result of context. Therefore, the chance of correctly guessing this missing word increases. They stated that this effect is similar to limiting the size of the text vocabulary since a smaller size also restricts the number of possible answers. In other words: the entropy of the speech decreases with decreasing vocabulary size.

Boothroyd (1968) stated that the probability of recognizing a speech sound within a word depends on intrinsic and contextual factors. Intrinsic factors are the acoustical properties of the speech sound and the frequency of occurrence of the phoneme class. Contextual factors are related to acoustical influences of adjacent speech sounds, second-order phoneme probability, and first-order word probability. The author derived the different factors for the recognition of consonant-vowel-consonant (CVC) words and related them to experimental data. This work has inspired various models regarding context effects in speech. The two models that were used in the current study are those by Boothroyd and Nittrouer (1988) and by Bronkhorst *et al.* (1993). Both models combine the intelligibility of elements in isolation with the effect of context to estimate the intelligibility of the complete speech token.

Boothroyd and Nittrouer (1988) introduced two constants (***k*** and ***j***) to quantify the degree of context. The ***k***-factor represents the increase in channels of statistically independent information due to context. Consider presenting a phoneme in isolation, or as part of a CVC-word. A CVC-word provides the listener with context, which increases the probability of recognizing the phoneme. This increased recognition probability is represented by ***k***. The ***j***-factor reflects the number of independent channels of information in a whole speech token. When no context is available (e.g., in nonsense CVC-words) the listener needs sensory information about each phoneme to recognize the complete word. In this case, ***j*** is approximately equal to the number of elements. When context is introduced, the value of ***j*** decreases.

Another approach to modelling context was introduced by Bronkhorst *et al.* (1993) and was partly based on Boothroyd (1968). They presented a two-stage model for recognition of speech. In the first stage, identification is based on sensory information alone. The second stage adds the effects of context to account for the missed elements in stage one, represented by the context factor ***c***. Consider a CVC-word with the final phoneme completely missing. The recognition probability of the whole word is the product of the recognition probabilities of the first two phonemes, multiplied by the probability that the final phoneme can be inferred on the basis of context. The latter probability is represented by ***c***.

The models by Bronkhorst *et al.* (1993) and Boothroyd and Nittrouer (1988) are fundamentally different. The advantage of the approach by Boothroyd and Nittrouer is that it is relatively simple and intuitive. Also, this model can be applied to existing data relatively easily. One drawback is that the model assumes equal recognition probabilities of the individual elements, which decreases its applicability in fluctuating background conditions. Also, in CVC-words, the vowels are generally more easily recognized than the consonants. Furthermore, the ***k***- and ***j***-factor both represent the effect of context, but it is not clear how they are interrelated. In contrast, the model of Bronkhorst can deal with these aspects, but it is more complex and therefore less intuitive.

4.2.3 Purpose of the current study

The goal of the current study was to revise the ESTI-model in order to improve its prediction accuracy for speech intelligibility in fluctuating noises at low (< 8 Hz) modulation rates. To achieve this, the ESTI was used to estimate intelligibility for each speech element instead of for each whole speech token. Next, the

intelligibility of all elements was combined using the context model of Boothroyd and Nittrouer (1988) or Bronkhorst *et al.* (1993). This combined approach assured that the amount of context determined the probability that a masked element was recognized. The existing data by Miller and Licklider (1950) were used to evaluate this method, which consists of monosyllabic words presented in stationary noise, in interrupted noise and with silent interruptions.
The current approach was based on the approach by Bronkhorst *et al.* (1993) who used the fraction of perceived speech in interrupted words combined with their own context model to predict intelligibility. To validate our approach, this procedure was replicated. Additionally, the context model by Boothroyd and Nittrouer was used in place of the Bronkhorst model to compare the two models. The final step was to carry out the main purpose of the current study by predicting intelligibility in interrupted noise using the ESTI in combination with each of the two context models.

4.3 Materials and methods

4.3.1 Intelligibility data

The data by Miller and Licklider (1950) were used to evaluate the modelling approach that was proposed in this study. They used phonetically balanced, monosyllabic word lists published by Egan (1948)[vi]. The speech was distorted by regularly spaced silent periods at duty cycles (DCs) of 12.5%, 25%, 50%, and 75% and regularly spaced interrupted white noise at a DC of 50%. The long-term SNRs when using interrupted noise were –15, –6, +3, and +12 dB[vii]. Scores were obtained using normal hearing listeners. All data were read from Figs. 4 and 8 from Miller and Licklider (1950). See Appendix D for the numerical data that were used in the current paper.

4.3.2 Model overview

Bronkhorst *et al.* (1993) modelled recognition of interrupted speech based on the speech time fraction (STF) using their own context model. In the current study, this method was replicated. In addition, their STF method was combined

vi Note that in the original study, the speech material by Egan (1948) was used, in which not all words had the same structure. The majority of words were CVC-words, but also CV, VC, CCVC, CVCC, and CCVCC words were used during testing (where C represents consonant, and V represents vowel). However, we simplified the model by assuming that all words have a consonant-vowel-consonant structure.

vii Note that the long-term SNRs are used in the current paper, where Miller and Licklider (1950) reported the SNR during the noise bursts. The SNRs reported here are therefore 3 dB higher than those reported in the original study.

with the context model by Boothroyd and Nittrouer (1988). To predict intelligibility of speech in interrupted noise, the ESTI (Van Schoonhoven *et al.*, 2019) replaced the STF, but the rest of the approach remained the same. This led to several different approaches for the predictions of intelligibility of interrupted speech and speech in interrupted noise using both context models. The nomenclature used for these approaches is depicted in Table 4-1.

Table 4-1: Nomenclature for the various methods used in this paper.

Condition	Approach	Context model
Interrupted speech	$cSTF_1$	Bronkhorst *et al.* (1993)
	$cSTF_2$	Boothroyd and Nittrouer (1988)
Interrupted noise	$cESTI_1$	Bronkhorst *et al.* (1993)
	$cESTI_2$	Boothroyd and Nittrouer (1988)
	ESTI	None

4.3.2.1 Interrupted speech

Bronkhorst *et al.* (1993) calculated the perceived fraction per phoneme (the STF) and linked this to the recognition probability of the isolated phoneme (q_e). Once this TF was known, the perceived fractions of all elements for various phase shifts, DCs, and interruption rates were calculated to predict q_e, which was then used as input for the context model to predict the word score p_w. In the current study, when predicting intelligibility of interrupted speech using the context model by Bronkhorst et al. ($cSTF_1$), the approach was identical to the method they described in Sec. III C of Bronkhorst *et al.* (1993). Fig. 4-1 shows a schematic overview of the necessary steps.

4.3.2.2 Interrupted noise

The main focus of the current study was to predict speech intelligibility in interrupted noise. In Fig. 4-2, data from Miller and Licklider (1950) together with model predictions based on Van Schoonhoven *et al.* (2019) are shown. The various symbols represent the observed word scores as shown in Miller and Licklider (1950). The dashed-dotted lines in the left panel represent the model predictions using the ESTI-model. It is clear that the original ESTI fails at rates lower than 5 Hz.
Predicting speech intelligibility in interrupted noise was done using an approach similar to that described for interrupted speech. Instead of calculating the STF per phoneme, the ESTI per phoneme was obtained. The major difference between the approaches was that interrupted speech is either on or off,

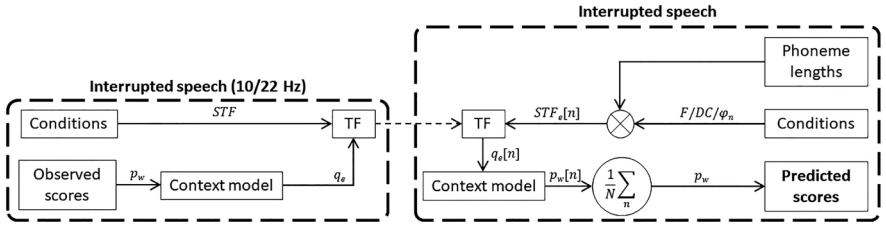


Fig. 4-1: Schematic overview of the steps to predict scores in interrupted speech as described by Bronkhorst et al. (1993). The left box shows the steps to estimate the TF based on scores in interrupted speech. The right box shows the steps to use this TF to estimate word scores in all N conditions, based on phoneme length, interruption rate (F), DC, and timing of the interruptions (φ). Subscript e refers to element (phonemes in the case of CVC-words) and subscript w refers to whole (words in the case of CVC-words). The context model either refers to Bronkhorst et al. (1993) ($cSTF_1$) or Boothroyd and Nittrouer (1988) ($cSTF_2$).

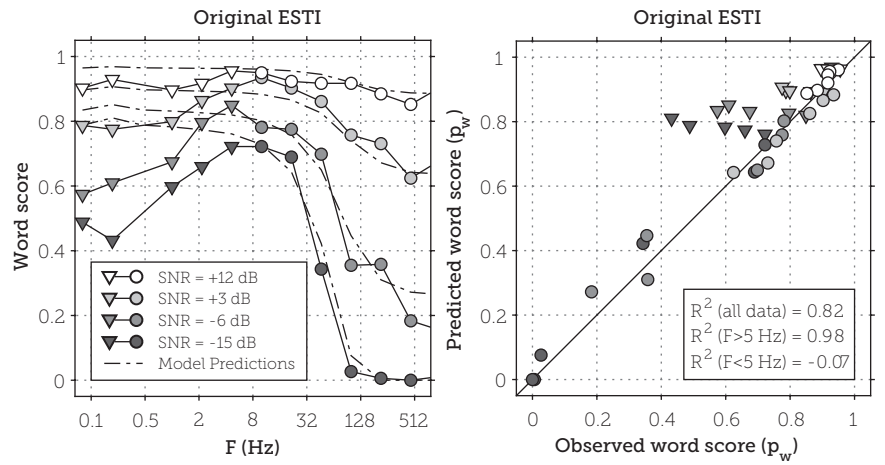


Fig. 4-2: Interrupted noise data by Miller and Licklider (1950) (symbols), together with the model predictions by the original ESTI (Van Schoonhoven et al., 2019). Triangles represent interruption rates < 5 Hz and circles represent interruption rates > 5 Hz. The TF was based on the data in stationary noise.

corresponding to a local value of 1 or 0, respectively. The average of these values during one phoneme corresponds to the STF per phoneme. In contrast, the local ESTI-value can take any value between 0 and 1, depending on the local SNR (and potentially on the reverberation). The ESTI per phoneme, therefore,

reflects information about the on-/off-time of the noise, as well as information about the detrimental effect of the noise itself (and of the reverberation). In Fig. 4-3, the steps of this approach are shown.

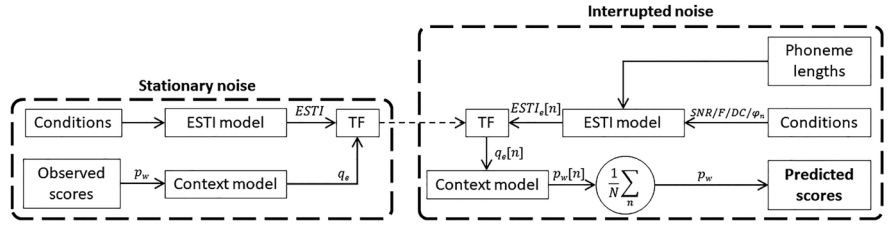


Fig. 4-3: Schematic overview of the steps to predict scores in interrupted noise, based on the method described by Bronkhorst et al. (1993). The left box shows the steps to estimate the TF based on scores in stationary noise. The right box shows the steps to use this TF to estimate word scores in all N conditions, based on phoneme length, interruption rate (F), DC, SNR, and timing of the interruptions (φ). Subscript e refers to element (phonemes in the case of CVC-words) and subscript w refers to whole (words in the case of CVC-words). The context model either refers to Bronkhorst et al. (1993) ($cESTI_1$) or Boothroyd and Nittrouer (1988) ($cESTI_2$).

4.3.3 Context models

4.3.3.1 Boothroyd and Nittrouer model

The context parameters in the Boothroyd and Nittrouer (1988) context model are k and j. Their first assumption was that context adds channels of independent data, equivalent to those already available from the speech itself. This means that the logarithms of error probabilities of contextual and sensory channels are additive. In other words, either the sensory or the contextual channel is sufficient to recognize the complete speech element:

$$\log(1 - p_e) = \log(1 - q_e) + \log(1 - c_e) \quad (4\text{-}1)$$

with p_e as the probability of recognizing a speech element in context, q_e as the probability of recognizing the same element from sensory information alone, and c_e as the probability of speech recognition from context alone. An element can be a phoneme as a part of a word, but also a syllable or word as a part of a sentence.

The authors further assumed that, since both the target speech and the context must be perceived under the same conditions (e.g., masking noise), $\log(1 - q_e)$ is proportional to $\log(1 - c_e)$. Hence, the following relation applies:

$$p_e = 1 - (1 - q_e)^k \quad (4\text{-}2)$$

with k as the proportionality constant which reflects the degree of context. Absence of context is represented by k = 1 (p_e = q_e and c_e = 0) and a higher k signifies an increase in context.
The second relation by Boothroyd and Nittrouer (1988) described the recognition of a whole (e.g., a CVC-word) with respect to the recognition of its elements (e.g., the phonemes):

$$p_w = {p_e}^j \tag{4-3}$$

with p_w as recognition of a whole and p_e as the recognition of the elements. For nonsense words, the context factor j is equal to the number of elements (when coarticulation cues are disregarded). When context is added, j decreases. When a listener only needs one element to recognize the whole speech token, this means that j = 1.

4.3.3.2 Bronkhorst model

The model by Bronkhorst *et al.* (1993) is a two-stage model. In the first stage, intelligibility depends solely on sensory information. Contextual information is added during the second stage. Let q_e be the probability of correctly identifying an element based on sensory information alone, with equal probabilities for each element. Then, the probability of correctly identifying a complete speech token of n elements based on sensory information alone is the product of the individual probabilities: $Q_0 = (q_e)^n$, where subscript 0 refers to the number of errors made in the sensory stage.
The second stage introduces c as the probability of identifying an element based on context alone, where this element was missed in the sensory stage. Note that these context probabilities are not coupled to a specific element. For example, the probability of missing one element in the sensory stage equals $n(q_e)^{n-1}(1-q_e)$. The probability of correctly identifying this missed element in the second stage using context equals c_1. Consequently, correct identification of the whole speech token in this example using both sensory and contextual information equals

$$p_w = nc_1Q_1 \tag{4-4}$$

with

$$Q_1 = (q_e)^{n-1}(1-q_e) \tag{4-5}$$

These equations can be generalized for any missing number of elements in both stages, which leads to Eqs. (4-6) and (4-7) for recognition probability of the whole:

$$p_w = Q_0 + \sum_{i=1}^{n}\left(Q_i \prod_{m=1}^{i} c_{i-m+1}\right) \tag{4-6}$$

with

$$Q_i = \binom{n}{i}(q_e)^{n-i}(1-q_e)^i \tag{4-7}$$

However, this expression is only valid for equal recognition probabilities of each element (q_e). It is also possible to define different recognition probabilities for the separate elements. This can be useful in the case of CVC-words, where recognition probability of the vowel is generally higher than the consonants under similar conditions (Bosman, 1989; Fogerty, 2014). Besides this, it is also possible to define the probability of recognizing n-i elements (with i = 0...n), leading to a set of equations for $p_{w,n-i}$. This can be useful when fitting the model since more datapoints are available. The complete set of equations for CVC-words is displayed in Appendix C.
To limit the number of parameters during the fitting of the model using CVC-words, the vowel and consonant recognition in isolation are related by κ using the following mathematical relation:

$$q_v = 1-(1-q_c)^\kappa \tag{4-8}$$

4.3.3.3 Context factors

The context factors were needed in two steps during the current modelling approach. First, since the isolated phoneme scores were not reported, q_e had to be estimated based on the reported word scores p_w. This step was necessary to estimate the TF. Next, after the TF was used to calculate a value for q_e based on the STF (for interrupted speech) or on the ESTI (for interrupted noise), the final word score p_w was estimated using these context factors.
No details were available about the monosyllabic words in the study by Miller and Licklider (1950) to estimate the context factors for this speech material. Therefore, values reported by Bronkhorst *et al.* (1993) about the Dutch CVC-words by Bosman and Smoorenburg (1995) were used. They estimated the c-factors for their own model, and the j-factor for the context model by Boothroyd and Nittrouer (1988). These values are depicted in Table 4-2. Also, the values for κ, which links the vowel and consonant scores via Eq. (4-8), are depicted. The k-factor was not reported by Bronkhorst.
According to Eq. (4-2), k relates the phoneme score in isolation (q_e) to the phoneme score in context (p_e). Therefore, to estimate k, information was needed about these scores. Since q_e was not available, the assumption was made that q_e and p_e are equal in nonsense words. This assumption states that it is equally likely for a listener to recognize a phoneme in isolation as it is in a nonsense

Table 4-2: Context factors for meaningful words as estimated by Bronkhorst et al. (1993). † c_3 is set to 0 since guessing was not allowed. * k-values were not reported, but were estimated based on the data available in Bosman and Smoorenburg (1995).

		Quiet	Noise
Bronkhorst model	c_1	0.25	0.47
	c_2	0.11	0.20
	c_3†	0.00	0.00
	κ	3.9	2.6
Boothroyd and Nittrouer model	j	2.7	2.2
	k	1.5*	1.3*

word. However, in reality, this is not the case, due to coarticulation, and durational and linguistic cues. Due to a lack of a better alternative, p_e in nonsense words was used as a proxy for q_e in meaningful words to estimate k. Using the nonsense words described by Bosman and Smoorenburg (1995), this resulted in k = 1.5 for intelligibility of meaningful words in quiet and k = 1.3 for intelligibility of meaningful words in stationary noise. Note that Boothroyd and Nittrouer (1988) found a value of 1.32 in stationary noise.

4.3.4 Estimation of TF

The goal of a TF is to translate the amount of speech information available to an estimated isolated phoneme score (q_e). An exponential function was fitted to the data [see Eq. (4-9)] using a linear least squares approach:

$$q_e = \gamma + \alpha e^{\beta x} \tag{4-9}$$

Here, x can be the STF or the ESTI. Note that this relation is different from the relation used by Bronkhorst *et al.* (1993), where only one model parameter was fitted (γ = 1 and α = –1). This led to q_e = 1 at STF = 1 and q_e > 0 for STF > 0 in their paper. Since this is not necessarily the case, the two additional parameters were introduced here.

4.3.4.1 Interrupted speech

Miller and Licklider (1950) stated that for interrupted speech at a rate of approximately 10 Hz, the word score was solely dependent on the fraction of speech that was available to the listener (speech-time fraction or STF). This is visible in Fig. 6 in their paper. This rate ensured a glimpse at each phoneme, even for low DCs. At lower interruption rates, complete phonemes were often missed, leading to a decrease in speech recognition. At higher rates (especially between 100 and 3000 Hz), the spectral splatter created by the sidebands due to the abrupt interruptions interfered with speech recognition.

The average phoneme length was 100–250 ms and the average word length was 600 ms (see Miller and Licklider (1950), their Fig. 3). Therefore, an interruption rate of 10 Hz led to at least one glimpse per phoneme. Assuming that the audible distortion as a result of spectral splatter is negligible up to ~20 Hz, the average scores at rates of 10 and 22 Hz for each DC were used to estimate the TF between STF and q_e, leading to four datapoints.

The isolated phoneme scores were calculated using the context factors depicted in Table 4-2. There are theoretical arguments to either use the context parameters obtained in quiet or those obtained in noise for the modelling of the interrupted speech data. The main difference is the fact that listeners have less access to coarticulation cues when speech is presented near threshold in quiet, leading to a smaller effect of context (Bronkhorst *et al.*, 1993). However, interrupted speech was presented at levels well above threshold, giving the listener sufficient access to these cues. It is unknown how important these cues are for correct identification of words under these conditions, in comparison to stationary noise. We chose to model interrupted speech perception based on the context parameters in quiet.

4.3.4.2 Interrupted noise

The original ESTI-model assumes that equal ESTI-values are needed to ensure equal intelligibility in stationary and in fluctuating noise. Therefore, the TF between the ESTI and q_e that was needed to predict speech intelligibility in *interrupted noise*, was based on intelligibility in *stationary noise*. For the calculation of the ESTI, properties of the speech and noise had to be known. White noise was used as a masker [since this was originally used by Miller and Licklider (1950)], multiplied by a square wave in the case of interrupted noise. SNRs were based on the long-term rms-values in the frequency range between 100 and 7000 Hz. For the speech signal, a stationary noise signal was used with the long-term average spectrum based on the International Speech Test Signal or ISTS (Holube *et al.*, 2010). Note that this is not the true spectrum of the speech that was used in the study by Miller and Licklider (1950).

4.3.5 Prediction of word scores

To calculate the STF and ESTI per phoneme, information was needed about word and phoneme lengths, and about the duration and timing of the peaks and gaps of the interruptions. Word lengths were uniformly distributed between 480 and 720 ms with relative phoneme durations of 22%, 45%, and 33% for the initial consonant, vowel, and final consonant, respectively. This choice was based on Fig. 3 from Miller and Licklider (1950).

For each word length, the start of the first interruption was varied, so that various phase shifts were addressed. For each phase shift, the STF or ESTI of each phoneme was calculated. Next, the TF was used to estimate q_e, which was then fed to either of the two context models to calculate the predicted word scores. All predicted word scores were averaged to calculate the mean predicted value p_w.

4.3.6 Comparison of models

The observed and predicted scores were compared to analyze model accuracy, which was quantified by the coefficient of determination (R^2). This value was calculated by subtracting the ratio of the total sum of squares and the residual sum of squares from 1.

4.4 Results

4.4.1 Transfer function

The TF related the available speech information to the isolated phoneme score (q_e). The available speech information was represented by the STF for interrupted speech, or by the ESTI for interrupted noise. The first step was to estimate the TFs for both conditions and both context models, resulting in four different functions. The parameters in Eq. (4-9) were fitted to the data and the resulting values are depicted in Table 4-3. The corresponding TFs are shown in Fig. 4-4.

Table 4-3: Parameters used in the fit of the TF in Fig. 4-4 using Eq. (4-9). The variable x can be replaced by STF for interrupted speech, and by ESTI for interrupted noise. †The values of α and γ by Bronkhorst et al. (1993) were not fitted, but were set to the depicted values.

		α	β	γ
Interrupted speech	$cSTF_1$	-1.3	-8.9	0.97
	$cSTF_2$	-1.1	-6.9	0.92
	Bronkhorst *et al.* (1993)	-1†	-6.5	1†
Interrupted noise	$cESTI_1$	-1.6	-4.8	0.99
	$cESTI_2$	-1.5	-5.4	0.92

The values of q_e used to fit the TFs were derived from the word scores from Miller and Licklider (1950) using both context models. This was done using Eqs. (4-2) and (4-3) for the Boothroyd and Nittrouer model, and using Eqs. (4-6) and (4-7) for the Bronkhorst model. The q_e-values in interrupted speech were

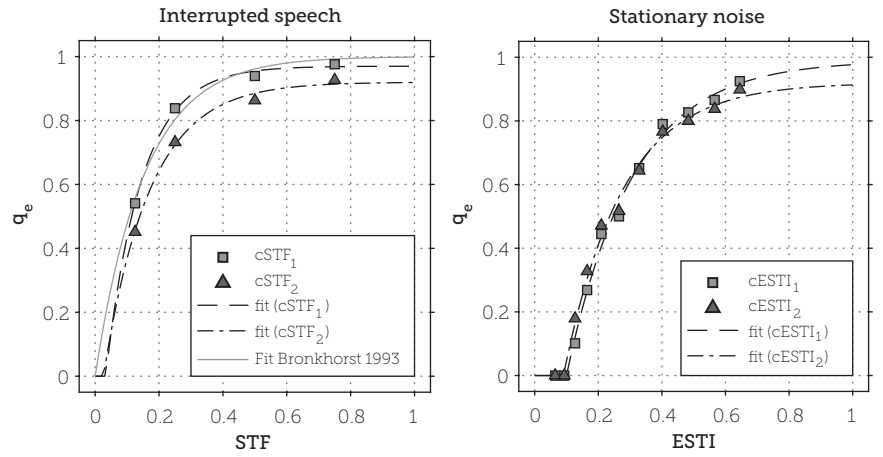


Fig. 4-4: TF between STF and q_e for interrupted speech (left panel) and ESTI and q_e for stationary noise (right panel). The q_e-scores in the left panel were derived from the word scores at interruption rates of 10 and 22 Hz using context factors in quiet. The depicted datapoints are the averaged values over these rates for each of the four duty cycles. The q_e-scores in the right panel were derived from word scores in stationary noise at various SNRs using context factors in noise. In the left panel, the fit by Bronkhorst et al. (1993) is also shown.

derived from the word scores at interruption rates of 10 and 22 Hz, depicted in Fig. 4 of Miller and Licklider (1950). The averaged values per STF at these rates were used here. The q_e-values in interrupted noise were derived from the word scores in stationary noise as presented by Fig. 9 of Miller and Licklider (1950). Since both models apply context differently, the values of q_e depend on the context model that is used.

The STF in the left panel of Fig. 4-4 is equal to the DC in the corresponding measurement condition. The ESTI in the right panel was calculated based on the properties of the speech and the stationary noise.

4.4.2 Model predictions

4.4.2.1 Interrupted speech

In Fig. 4-5, the model predictions for interrupted speech are shown, together with the original data from Miller and Licklider (1950). The roll-off at low rates is clearly visible in the model predictions using both context models. The data and predictions are shown up to an interruption rate of 512 Hz. Starting at 46 Hz at the lowest DC, spectral splatter due to the interruptions caused masking of the speech itself. This problem was a side effect related to the presentation mode. Modelling this aspect was beyond the scope of the current study.

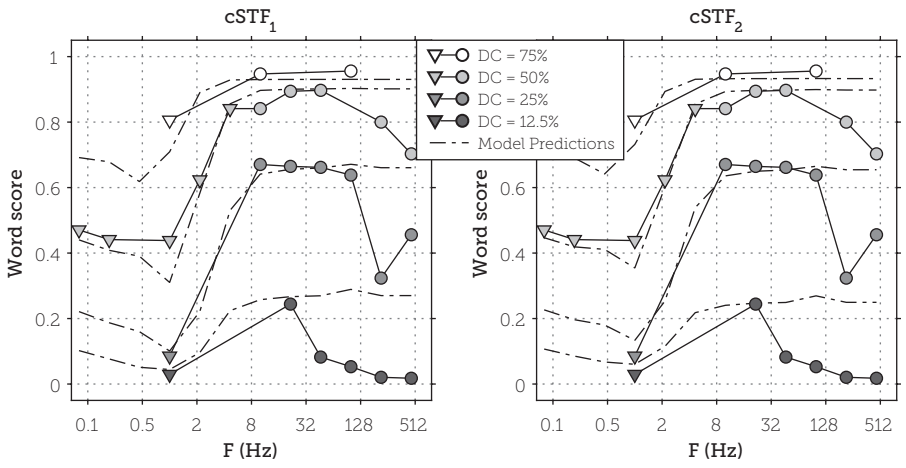


Fig. 4-5: Interrupted speech data by Miller and Licklider (1950) (symbols), together with the model predictions using $cSTF_1$ (STF + Bronkhorst context model) and $cSTF_2$ (STF + Boothroyd and Nittrouer context model). Triangles represent interruption rates < 5 Hz and circles represent interruption rates > 5 Hz.

4.4.2.2 Interrupted noise

In Fig. 4-6, the model predictions of speech intelligibility in interrupted noise are depicted. When comparing these predictions to the original ESTI predictions in Fig. 4-2, it appears that predictions at rates > 5 Hz are comparable to those of the original ESTI-model. More importantly, as opposed to the original ESTI, the drop in scores below 5 Hz was captured by the new model. Where the original ESTI showed a gradually increasing predicted intelligibility for decreasing interruption rates (see Fig. 4-2), the current model predicted a roll-off with a minimum score around 1 Hz. This trend was also visible in the original data.

Model predictions appear reasonable for higher SNRs (+3 dB and +12 dB), especially using $cESTI_1$ (the two upper lines in the left panel of Fig. 4-6). However, $cESTI_2$ yielded lower intelligibility estimations at higher SNRs. Furthermore, both $cESTI_1$ and $cESTI_2$ underestimated intelligibility at lower SNRs (–15 and –6 dB), especially between modulation rates of 0.5 and 2 Hz. The drop in word scores at these rates was the result of complete masking of whole phonemes by the noise peaks. The drop in model predictions was more dramatic than the drop in actual scores, especially around 1 Hz at lower SNRs. Apparently, both models overestimated the detrimental effect of masking complete meaningful elements.

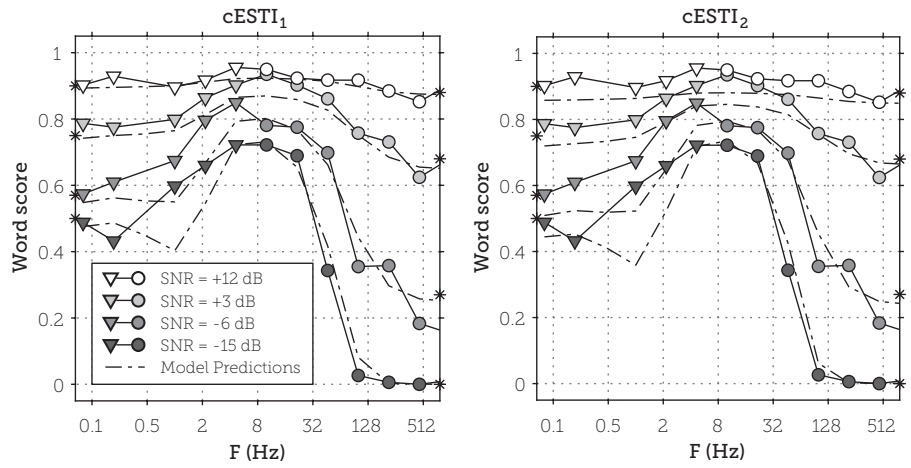


Fig. 4-6: Interrupted noise data by Miller and Licklider (1950) (symbols), together with the model predictions using $cESTI_1$ (ESTI + Bronkhorst context model) and $cESTI_2$ (ESTI + Boothroyd and Nittrouer context model). Triangles represent interruption rates < 5 Hz and circles represent interruption rates > 5 Hz.

In Fig. 4-7, the observed and predicted word scores are depicted for interrupted noise. Prediction accuracy seems reasonably good. The accuracy increased in comparison to the original ESTI (see Fig. 4-2). R^2 = 82% for the original ESTI, whereas the explained variance increased to 95% and 92% for $cESTI_1$ and $cESTI_2$, respectively. Accuracy especially increased for interruption rates lower than 5 Hz: R^2 = –0.07 for the original ESTI[viii] and is now 0.78 ($cESTI_1$) and 0.62 ($cESTI_2$).

viii The used linear model to fit the observed and predicted data is always represented by $y = x$, and not necessarily by the best linear fit $y = ax + b$. When the mean of the data is a better predictor than $y = x$, a (counterintuitive) negative value of R^2 is obtained.

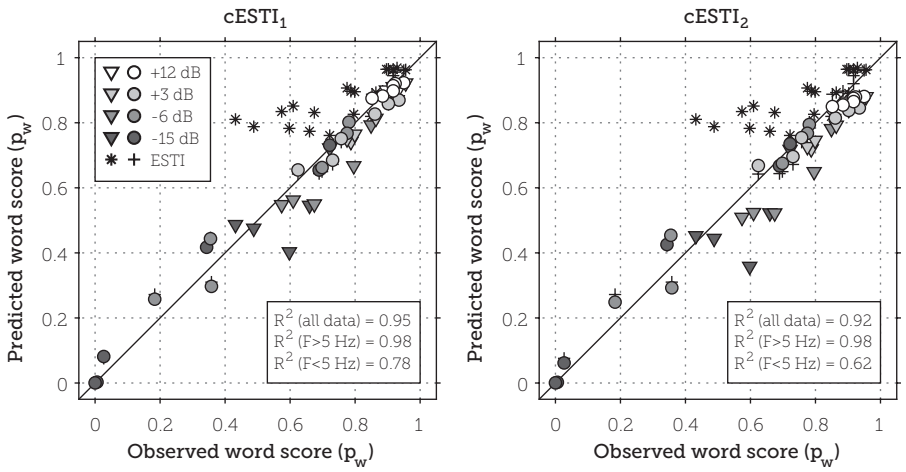


Fig. 4-7: Interrupted noise. Predicted versus observed word scores using $cESTI_1$ (ESTI + Bronkhorst mode) and $cESTI_2$ (ESTI + Boothroyd and Nittrouer model). Triangles represent interruption rates < 5 Hz and circles represent interruption rates > 5 Hz. Stars (< 5 Hz) and plus-signs (> 5 Hz) represent predictions by the original ESTI (see also Fig. 4-2, right panel).

4.5 Discussion

The main goal of the current study was to combine the ESTI (Van Schoonhoven *et al.*, 2019) with one of two existing context models (Boothroyd and Nittrouer, 1988; Bronkhorst *et al.*, 1993) to evaluate the prediction of speech intelligibility in interrupted noise. This was done by predicting intelligibility at phoneme level using the ESTI, followed by the application of the context models to obtain the word scores. By combining the ESTI with a context model, bottom-up processes were separated from top-down processes. The ESTI itself represented the sensory stage, and solely reflected the quality of the acoustical information that was presented to the listener. The resulting local ESTI-values were then used as an input for higher order processes, represented by the context model. Although still a simplified approach, the assessment of speech quality at phoneme-level and the introduction of context better reflected the processing of speech information than the original STI and ESTI.

4.5.1 Model performance

The first, general observation is that the addition of both context models leads to a more accurate prediction of the data of Miller and Licklider (1950), especially

for interruption rates lower than 5 Hz (compare Fig. 4-7 and the right panel of Fig. 4-2). This observation is especially true for low SNRs, where the discrepancy between measured data and predictions of the original ESTI-model is largest. The explained variance (R^2) for F < 5 Hz and all SNRs increases from −0.07 to 0.78 and 0.62 for the $cESTI_1$ and $cESTI_2$, respectively.

At a broadband SNR of –15 dB, the SNR at the noise peaks is –18 dB. Since intelligibility in stationary noise under these conditions is not possible (Miller and Licklider, 1950), we can assume that almost no speech information is available to the listener during the on-cycles of the noise. As a consequence, the pattern for interrupted noise at low rates is similar to the observed and predicted pattern of interrupted speech. After all, during the off-cycles of interrupted speech, there is also no speech information available. Furthermore, at low interruption rates, the effects of forward masking are negligible. The reason the model predictions for interrupted noise and speech at low interruption rates are not identical is that the local STI-value at an SNR of –18 dB is still 0.12. Therefore, the model still assumes there is some useable speech information available under these conditions.

At higher interruption rates (> 5 Hz), intelligibility in interrupted noise starts to deviate drastically from interrupted speech. At these rates, the gaps between the noise get shorter and the relative contribution of forward masking starts playing a more important role. Eventually, the character of the noise gets more and more continuous, with an asymptote at approximately 100–200 Hz (Miller and Licklider, 1950; Rhebergen *et al.*, 2006). Above this rate, intelligibility of speech masked by stationary and interrupted noise is similar. Since forward masking is accounted for by the model, this pattern is also seen in the model predictions.

At SNRs higher than –18 dB, the mechanisms become more complex, since the listener now also has access to speech sounds during the on-cycles of the noise. The current model assigns a local STI-value at each point in time, which would be 1 during the gaps in the noise (optimal intelligibility). For interrupted noise at low SNRs and interrupted speech, the local STI-values at the other timepoints are close to 0 (no intelligibility possible). However, at higher SNRs, the local STI-values during the noise on-cycles will be significantly larger than 0, leading to a contribution in predicted intelligibility.

When perceiving speech in interrupted noise at higher SNRs, traditional glimpsing is augmented by the perception of higher level speech sounds during the on-cycles of the noise. The cESTI-model deals with this combination in a straightforward manner. As an example, when a complete phoneme is masked by noise at 0 dB, the resulting local STI-value would be roughly 0.5. This leads to an isolated phoneme score of around 0.85 [according to Eq. (4-9)].

This predicted intelligibility is the same when 50% of a phoneme is completely masked by noise at a very unfavorable SNR of, for instance, –30 dB. The *first* half of the phoneme is then entirely unavailable to the listener, whereas the listener has complete access to the *second* half of the same phoneme, resulting again in a local ESTI-value of 0.5 and a predicted intelligibility of 0.85. Consequently, the cESTI approach does not distinguish between glimpsing speech or energetically masked speech. It is a relatively simple approach, which is not an accurate representation of the true mechanisms of speech perception. For example, the model assumes that the temporal distribution of information per phoneme is uniform, which is not the case (Smits, 2000; Smits *et al.*, 2003). However, despite these shortcomings, the cESTI-model predicts the pattern of speech intelligibility at all interruption rates and SNRs rather well.

The model predictions for both interrupted noise and speech show a pronounced V-shape around 1 Hz. For interrupted speech, this coincides with a dip in the intelligibility scores of the measured data. However, the interrupted noise data do not show this V-shape. Shafiro *et al.* (2018) discussed that (for interrupted speech and text) the number of words between two interruptions determines performance at low rates, and the number of interruptions per word determines performance at high rates. Assuming (for simplicity) equal phoneme lengths of 200 ms, there is a guaranteed glimpse at each phoneme above 2.5 Hz and a DC of 50% due to an increasing number of interruptions per word. Below 1.25 Hz the maximum number of interruptions per word is two, and the duration of the silent periods makes it impossible to get a glimpse of each phoneme. For lower interruption rates, the probability of perceiving all three phonemes increases again, since the number of words per interruption increases. Therefore, the total probability of perceiving all three phonemes shows a dip at 1.25 Hz, which corresponds to the observed V-shape. This appears to be a transition region between the dominance of the number of words during each speech fragment and the number of interruptions per word as described by Shafiro and colleagues.

The reason for the inaccuracy of the model around this V-shape in interrupted noise might be related to the model choices that were made. Intelligibility of interrupted speech was modelled based on context factors in quiet (see Table 4-2) and a TF based on interrupted speech (left panel of Fig. 4-4). Intelligibility in interrupted noise was modelled using data based on stationary noise. For higher interruption rates (> 5 Hz), the character of interrupted noise gets more and more continuous. However, at lower interruption rates the noise is perceived as separate blocks. For example, at a rate of 0.1 Hz the on- and off-cycles of the noise are 5 s. It is more logical to model intelligibility under these conditions separately for quiet and noise. After all, it is not a question of

glimpsing speech anymore, but rather an alternation between intelligibility in quiet and in noise.

This explanation was tested by using the context factors in quiet for the model predictions of speech in interrupted noise (not shown in section 4.4). For $cESTI_1$ and $cEST_2$, the R^2-values decreased from 0.98 to 0.62 and from 0.98 to 0.65 respectively for F_{int} > 5 Hz, when being compared to context factors in noise. This is expected behavior since interrupted noise at higher rates behaves more like stationary noise. Looking at lower rates (F_{int} < 5 Hz), R^2 increased from 0.62 to 0.81 for $cESTI_2$ and remained similar for $cESTI_1$ at ~0.78. A combined approach with different degrees of context at lower modulation rates might be more accurate, but inevitably increases complexity of the model.

4.5.2 Model comparison

The main difference between performance of $cESTI_1$ and $cESTI_2$ is that the latter functions poorly at higher SNRs at all interruption rates. This becomes clear in Fig. 4-6, where $cESTI_2$ underestimates the word scores at SNRs of +3 and +12 dB. The main cause lies in the TFs, which were based on the data in stationary noise. As is visible in the left and right panel of Fig. 4-4, the TFs of $cESTI_1$ and $cESTI_2$ are different, especially for higher ESTI-values. The TFs plateau at $\sim\gamma$ at maximum ESTI, which is 0.99 for $cESTI_1$ and 0.92 for $cESTI_2$. This is probably caused by the lack of data at higher ESTI-values (> 0.7). When forcing γ to unity when fitting the TFs, the explained variance of $cESTI_2$ increased to 95% (not shown in section 4.4).

The primary advantage of $cESTI_2$ is its applicability. By knowing p_e and p_w, a simple fitting procedure leads to the j-factor. When also testing nonsense words, a similar procedure leads to k. A prerequisite is that the characteristics and circumstances (e.g., speaker or SNR) are the same for nonsense and meaningful words.

The theoretical disadvantage of $cESTI_2$ is that one of the primary assumptions was violated. Boothroyd and Nittrouer (1988) stated that the target speech and the context must be perceived under the same conditions (e.g., continuous masking noise). This assumption formed the basis for the introduction of the context factor k [see Eqs. (4-1) and (4-2)]. For example, when a listener only misses the first consonant in a CVC-word, the correct phoneme can be guessed based on the perceived vowel and final consonant. In stationary noise, all three phonemes are presented under same conditions. However, for speech in interrupted noise at lower rates, this is not the case. It is possible that the target speech (vowel and final consonant) coincides with a gap in the noise, and the context (first consonant) coincides with a noise peak. However, the consequences of this violation are limited, since the outcome of both models are similar,

probably because the results were averaged over all possible phase shifts of the interruptions. $cESTI_1$ deals with this issue by using a more elaborate approach, since the recognition probabilities of the individual phonemes are treated separately. However, with the current speech material, the benefit of this approach appears to be limited.

The chosen approach of the current study also has applications to other speech materials. When sentences are considered as whole speech tokens instead of words, the syllables can serve as elements. Bronkhorst *et al.* (1993), Sec. III D, already applied their model to sentence recognition. In the same fashion, the current approach could theoretically be applied to sentence intelligibility in interrupted noise. Due to the heterogeneity of sentence material and the larger contribution of context, the current study focused solely on monosyllabic words. Besides extrapolation to other speech materials, a next step would be the application to more realistic non-stationary noises, like speech-modulated noise or babble noise.

4.6 Conclusions

This study evaluated an improved implementation of the Extended Speech Transmission Index (ESTI) to estimate intelligibility of monosyllabic words. The new model included two key factors. First, the ESTI was calculated per phoneme instead of per word. Second, the ESTI was combined with either the context model by Boothroyd and Nittrouer (1988) or by Bronkhorst *et al.* (1993) to predict word scores in interrupted noise. Compared to the original ESTI, this cESTI-model better predicted the roll-off in speech scores at interruption frequencies below 5 Hz.

For both context models, the performance was similar and the prediction accuracy was good. The ESTI combined with the Bronkhorst model was more elaborate and theoretically more suitable for the application to interrupted noise since the specific behavior of the individual phonemes could be taken into account. However, this theoretical benefit was barely reflected in better performance. Therefore, we regard the ESTI combined with the simpler Boothroyd and Nittrouer model to be a more suitable candidate to model intelligibility in non-stationary conditions. A next step would be a more thorough validation of the cESTI-model using speech materials with different degrees of context.

4.7 Acknowledgements

This work was financially supported by the Heinsius-Houbolt foundation. The authors thank the two reviewers and the associate editor for their valuable comments.

CHAPTER 5

A CONTEXT-BASED APPROACH TO PREDICT SPEECH INTELLIGIBILITY IN INTERRUPTED NOISE: MODEL EVALUATION

Van Schoonhoven, J., Rhebergen, K.S., Dreschler, W.A. (2023)
Manuscript submitted for publication in the Journal of the Acoustical Society of America

5.1 Abstract

The context-based Extended Speech Transmission Index (cESTI) by Van Schoonhoven *et al.* (2022) was successfully used to predict the intelligibility of meaningful, monosyllabic words in interrupted noise. However, it is not clear how the model behaves when using different degrees of context. In the current paper, intelligibility of meaningful and nonsense CVC-words in stationary and interrupted noise was measured in fourteen normally hearing adults. Intelligibility of nonsense words in interrupted noise at –18 dB SNR was relatively poor, possibly because listeners did not profit from coarticulatory cues as they did in stationary noise. With 75% of the total variance explained, the cESTI-model performed better than the original ESTI-model (R^2 = 27%), especially due to better predictions at low interruption rates. However, predictions for meaningful word scores were relatively poor, mainly due to remaining inaccuracies at low interruption rates and a large effect of forward masking. Adjusting parameters of the forward masking function improved the accuracy of the model to a total explained variance of 83%, while the predicted power of previous published (c) ESTI data remained similar.

5.2 Introduction

5.2.1 Context effects

When elements of speech are missed, the listener is often able to fill in these missed items on the basis of context. The listener might have a priori information about the stimulus set or knowledge of the sentence topic. Also, phonological and lexical constraints, and syntactic and semantic rules may play an important part in 'guessing' missed elements (e.g., Boothroyd, 1968; Boothroyd and Nittrouer, 1988). The contextual information that is available reduces the number of options and therefore increases the probability of correct identification of a missed element.

Several authors have attempted to model the effect of context. The models by Boothroyd and Nittrouer (1988) and by Bronkhorst *et al.* (1993) are probably best known. Bronkhorst *et al.* (1993) developed a two-stage model, with recognition of speech elements based solely on sensory information in the first stage. In stage two, context was introduced to account for the missed elements in stage one. Using a different approach, Boothroyd and Nittrouer (1988) coupled the element scores in isolation to the element scores in context. This relation represented the increase of channels of statistically independent information due to context. Besides this, they coupled the element scores in context to the recognition scores of the entire speech token. This relation reflected the number of independent channels of information in a whole speech token. Van Schoonhoven *et al.* (2022) used the context model by Boothroyd and Nittrouer (1988) in combination with the Extended Speech Transmission Index (Van Schoonhoven *et al.*, 2019) to predict word recognition in non-stationary background noise.

5.2.2 cESTI

The original Speech Transmission Index (STI) was based on the relation between modulation reduction due to noise and/or reverberation on the one hand, and the decrease in speech intelligibility on the (e.g., Houtgast and Steeneken, 1978; Steeneken and Houtgast, 2002; IEC60268-16, 2011). The extended STI or ESTI (Van Schoonhoven *et al.*, 2019) was based on the STI per timeframe, and incorporated forward masking and averaging of all local STI-values. This version better dealt with fluctuating background noises, but still had difficulties in noises with low (< 5 Hz) modulation frequencies.

To deal with this shortcoming, Van Schoonhoven *et al.* (2022) developed a context-based version of the ESTI: the cESTI. In this model, a transfer function was fitted to relate the ESTI to the isolated phoneme score instead of the entire word. This approach made the model more robust in case of low-frequency

masker fluctuations that occurred during the speech token. For example, when only two of the three phonemes in a CVC (Consonant-Vowel-Consonant) word are audible, the cESTI-model treats these phonemes separately. The probability of correctly perceiving these two phonemes will be high, whereas the probability of perceiving the third phoneme will be low. When the isolated phoneme scores were estimated using the ESTI, the context models of Boothroyd and Nittrouer (1988) and of Bronkhorst *et al.* (1993) were applied to calculate word scores in interrupted noise. The ESTI in combination with both context models yielded similar results. Therefore, the authors chose the simpler model by Boothroyd and Nittrouer as the best addition to the ESTI. This cESTI-model successfully predicted the dip for lower interruption frequencies in the data by Miller and Licklider (1950) with meaningful monosyllabic words. However, it is unknown how this model behaves for speech materials with different degrees of context.

5.2.3 Purpose of the current study

The purpose of the current study was to evaluate the cESTI-model described by Van Schoonhoven *et al.* (2022). To achieve this, speech intelligibility measurements were conducted in normal hearing subjects in stationary and interrupted noise, using both meaningful and nonsense CVC-words. The results were analyzed and compared to the cESTI predictions.

5.3 Materials and methods

5.3.1 Word intelligibility measurements

5.3.1.1 Subjects

Fourteen normally hearing subjects were recruited (six males and eight females) with mean age 25.9 years (range 18–44 years). All were native Dutch speakers and no hearing or language problems were reported. All subjects had pure tone thresholds of 20 dB HL or better at the octave frequencies between 250 and 4000 Hz.

Subjects were recruited via posters. They gave written informed consent and received compensation for participating. Approval for the project (NL48348.018.14) was given by the Ethical Review Board (METC AMC).

5.3.1.2 Stimuli

The target speech consisted of Dutch meaningful and nonsense CVC-words, uttered by a female speaker and presented in separate lists of 12 words. The corpus of the meaningful words consisted of 180 unique words, distributed over 45 lists created for adults [see appendix B.4 of Bosman (1989)]. The corpus

of the nonsense words consisted of 187 unique words, distributed over 48 lists, based on the 16 lists that showed the least variation in syllable score. See appendix A.2 of Bosman (1989) for detailed information. The cumulative distributions of phoneme and word lengths of meaningful and nonsense words are shown in Fig. 5-1.

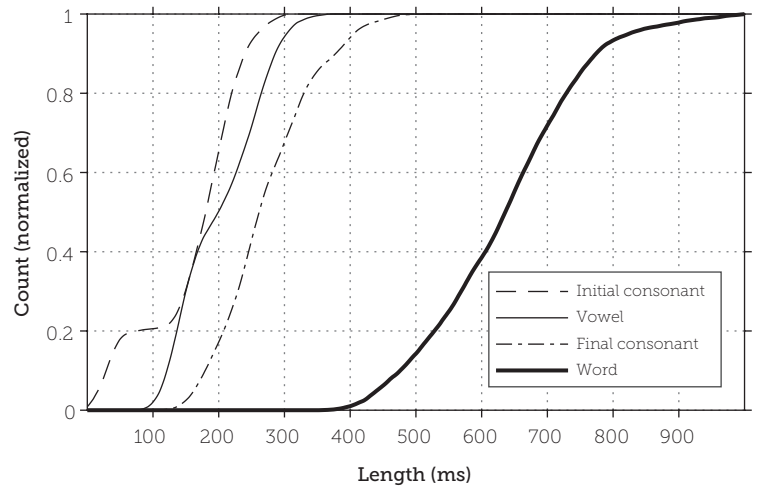


Fig. 5-1: Cumulative distribution of phoneme and word lengths of meaningful and nonsense words combined

Noise was always presented at 65 dB (A). Two types of distortion conditions were used. Stationary speech-shaped noise (SSN) was used as the reference condition at fixed SNRs of –12, –9, –6, –3, 0 and +3 dB. As a second condition, speech was masked by interrupted noise (IN). For this purpose, SSN was interrupted by silent periods using a square wave at octave frequencies between 0.5 Hz and 16 Hz, at a duty cycle of 50%. The long-term SNR was either –18 dB or –9 dB. A 4 ms raised-cosine function was applied at the on- and offset of each interruption in order to minimize spectral splatter. The timing of the interruptions was randomly altered per presentation. See Table 5-1 for a summary of the conditions.

5.3.1.3 Procedure

Signals were presented monaurally to the right ear through TDH39P headphones via a 24 bit/192 kHz Fireface 800 audio interface (RME, Haimhausen, Germany). Subjects were seated in a sound treated booth. Matlab (version 2017b, Mathworks

Table 5-1: All conditions, including number of subjects per condition (N). These conditions apply to the presentation of both meaningful and nonsense words. Interruptions of speech and noise were spaced regularly at octave frequencies (so at 0.5, 1, 2, 4, 8 and 16 Hz). See main text for further details.

	SNR	N		*F* (Hz)	-18 dB SNR	-9 dB SNR
SSN	-12 dB	7	IN	0.5	14	14
	-9 dB	14		1	14	14
	-6 dB	14		2	14	14
	-3 dB	14		4	14	14
	0 dB	14		8	14	14
	+3 dB	7		16	14	14

Inc., USA) was used for presentation of the sounds and for analysis of the results. A sampling frequency of 44.1 kHz and a bit depth of 16 bits/sample were used for all signals.

The conditions with distorting noise (SSN and IN) were presented in separate blocks. The average total word length was 633 ms (see Fig. 5-1). The 4, 8 and 16 Hz interruptions have a relatively continuous character, since the listener had access to parts of all three phonemes (assuming equal phoneme lengths as a first order approximation). The probability of perceiving at least part of all phonemes for the lower interruption rates was approximately 40%, 30% and 70% for 0.5 Hz, 1 Hz and 2 Hz respectively. Due to this difference in continuity, the lower and higher interruption rates were presented in different blocks, leading to three main blocks: SSN, IN_{low} and IN_{high}. The order of these blocks was randomly varied between subjects.

Within each block, the speech type (meaningful or nonsense) was presented in subblocks with randomly varying order between blocks. The presentation modes (SNR and/or interruption properties) were randomly varied within each subblock. The order of the word lists was pseudorandomized (not completely randomized, since words occurred in more than one list). Each word list was presented once per subject.

The total experiment was preceded by one list of meaningful words and one list of nonsense words in a random condition. Each of the three main blocks (SSN, IN_{low} and IN_{high}) was preceded by a practice list of either nonsense or meaningful words. The first word of a list was always used as practice. Subjects had a priori knowledge about the type of words and about the condition.

To reduce the test time, not all conditions were presented to all subjects. The SNRs of –12 dB and +3 in stationary noise were presented to seven of the fourteen subjects. These conditions were pseudo randomly distributed over all subjects,

making sure that the same number of conditions was presented to each subject, leading to a total number of test conditions of 35. Subjects were allowed a break each 20 minutes. Subjects had to repeat the words they heard. Answering with partial words was encouraged. All data was recorded and scoring was done post hoc. The conditions are depicted in Table 5-1.

5.3.2 Application of cESTI to CVC-words

5.3.2.1 Model Overview

In Van Schoonhoven *et al.* (2019) the calculation of the ESTI was described elaborately. The basis for this approach was the original STI as described in IEC60268-16 (2011) and the ESII-model (Rhebergen *et al.*, 2006). To account for forward masking, the model by Ludvigsen (1985) was used. To calculate the ESTI, the impulse response and local SNR were used to determine the apparent SNR for short, sliding time windows (2 – 11.3 ms). Per time window, a local STI-value was calculated according to IEC60268-16 (2011). Eventually, all local STI-values were averaged to obtain one ESTI-value.

Van Schoonhoven *et al.* (2022) proposed to calculate the ESTI per isolated phoneme instead of per complete monosyllabic word. A transfer function was applied to estimate the intelligibility of each isolated phoneme using the ESTI, after which the estimated phoneme scores were combined to predict the word scores. In this step, context was added to the model to account for the 'guessing' of elements that were missed. At this point, the ESTI-model was combined with the context model by Boothroyd and Nittrouer (1988).

The context model by Boothroyd and Nittrouer (1988) uses parameters k and j. The k-factor relates the phoneme score in isolation (q_e) to the phoneme score in context (p_e). See also Eq. (5-1). When q_e and p_e are equal (no context), k = 1. The j-factor relates the phoneme score in context to the word score (p_w) and is shown in Eq. (5-2). In CVC-words, when j = 3, information is needed about each phoneme to correctly identify the complete word. No context is available in this case. In summary, the cESTI-model calculates the ESTI per phoneme to estimate q_e, and uses k and j to estimate the word scores p_w.

5.3.2.2 Estimation of Context Factors

For the current CVC speech material, phoneme scores (p_e) were available for meaningful and nonsense words. This means that j and k were estimated using the following relations:

$$(1 - p_e) = (1 - q_e)^k \Longrightarrow k = \frac{\log(1 - p_e)}{\log(1 - q_e)} \quad (5\text{-}1)$$

$$p_w = {p_e}^j \Longrightarrow j = \frac{\log p_w}{\log p_e} \tag{5-2}$$

Because q_e was not available, the value of p_e in nonsense words was used as a proxy. Note that coarticulation cues were disregarded, possibly overestimating the value of q_e and underestimating k. Also, by definition this choice led to $k = 1$ for nonsense words.

5.3.2.3 Estimation of Transfer Function

The transfer function related the ESTI to the isolated phoneme score. To apply the model to non-stationary maskers, the transfer function was based on the results in stationary noise (SSN). For CVC-words in SSN, the isolated phoneme score q_e was estimated based on the values of p_e when k was known. Because the ESTI for each condition was also known, a transfer function between ESTI and q_e was calculated. As in Van Schoonhoven *et al.* (2022), the transfer function was of the form:

$$q_e = \gamma + \alpha e^{\beta ESTI} \tag{5-3}$$

5.3.2.4 Prediction of Intelligibility

To predict word intelligibility in interrupted noise, the ESTI was calculated for each phoneme in the utterance. An average word length of 633 ms (+/– 116) was used and an average duration of 161 ms (+/– 74 ms), 201 ms (+/– 63) and 271 ms (+/– 72 ms) for the initial consonant, vowel and final consonant, respectively (see Fig. 5-1).

The start of the speech token relative to the noise was varied, resulting in different phase shifts. For each phase shift, the ESTI was calculated for all phonemes in the speech token. Using the transfer function, values of q_e were calculated based on the ESTI-value per phoneme. The context model was then applied to estimate intelligibility of the whole speech token based on the values of q_e. The calculation is depicted in Eq. (5-4), where M represents the number of phonemes per word (3 in this case).

$$p_w = \left(\frac{1}{M} \sum_{m=1}^{M} \left(1 - \left(1 - q_{e,m} \right)^k \right) \right)^j \tag{5-4}$$

5.4 Results

5.4.1 Speech intelligibility measurements

5.4.1.1 Stationary noise

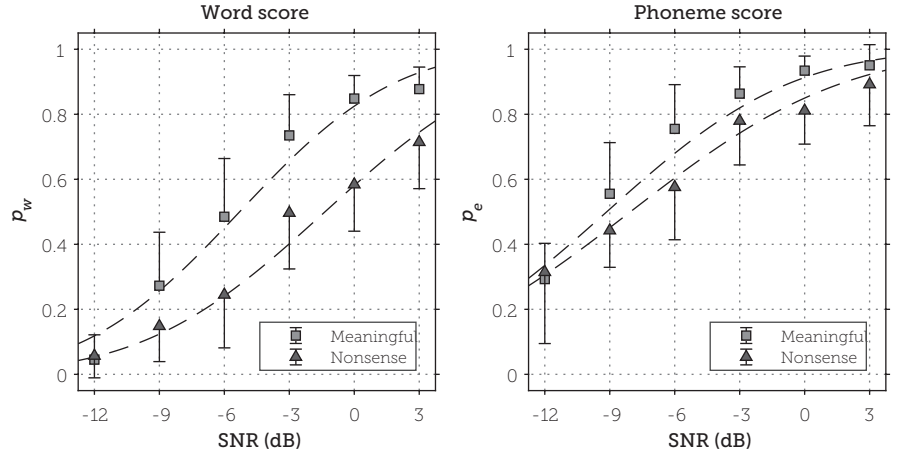


Fig. 5-2: Word and phoneme scores of meaningful and nonsense CVCs in stationary noise. Vertical bars represent standard deviations. Dashed lines represent the best fit of a linear regression model using a binomial distribution.

In Fig. 5-2 the results for meaningful and nonsense words are presented with SSN used as a masker. Word scores (p_w) are depicted in the left panel and phoneme scores (p_e) are depicted in the right panel. At –12 dB, there is no difference between scores in nonsense and meaningful words. In poor conditions, the access to contextual information is very limited, leading to no advantage due to context.

5.4.1.2 Interrupted noise

In Fig. 5-3 the results are presented for meaningful and nonsense words in interrupted noise. When performing multivariate ANOVA with type (meaningful or nonsense words), F_{int} and SNR as grouping variables, the scores at 8 Hz significantly differ from those at 1 Hz ($p < 0.001$) and at 0.5 Hz ($p < 0.05$). Furthermore, the scores at 4 Hz are significantly higher than the scores at 1 Hz ($p < 0.05$). No other significant effects were found.

In Fig. 5-3, a typical dip around 1 Hz can be seen in nonsense word scores at –9 dB SNR. This pattern is also visible in the data by Miller and Licklider (1950) for

monosyllabic words in interrupted speech and noise, and Shafiro *et al.* (2018) for interrupted speech. However, in the other conditions of the current study this pattern is not so obvious. In the left panel of Fig. 5-3 a plateau around 1 Hz is visible at –18 dB, but not at –9 dB. Nonsense word scores at –18 dB SNR also do not show this pattern. In the latter condition, intelligibility is more or less the same for interruption frequencies between 0.5 and 16 Hz.

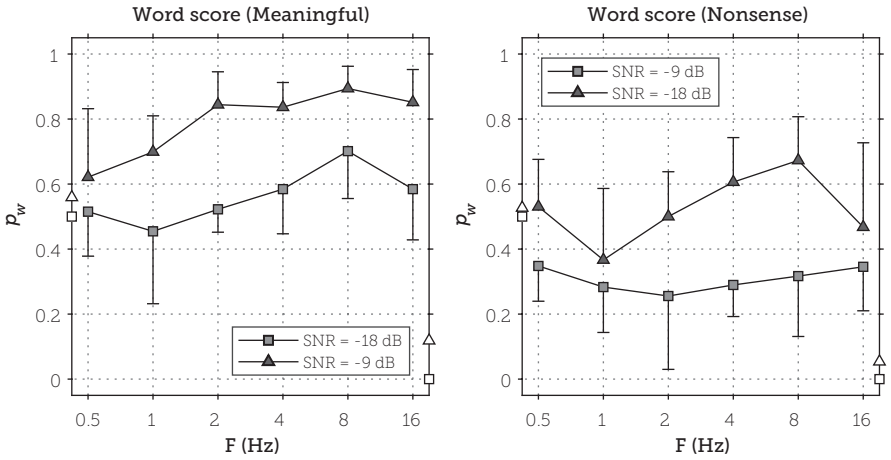


Fig. 5-3: Word scores of meaningful and nonsense CVCs in interrupted noise at rates between 0.5 and 16 Hz, long-term SNRs of –9 and –18 dB and a duty cycle of 50%. Vertical bars represent standard deviations. The open symbols represent the theoretical scores at infinitely low and high interruption frequencies, based on the (estimated) scores in stationary noise at –21 dB and –12 dB

5.4.2 Application of cESTI to CVC-words

5.4.2.1 Context factors

Using the results in SSN, the context factors for CVC-words were calculated based on Eqs. (5-1) and (5-2). This resulted in values of k = 1.4 and j = 2.2 in meaningful words, and k = 1 (by definition) and j = 2.7 in nonsense words. Bronkhorst *et al.* (1993) reported j-values for the same speech material of 2.2 and 2.8 for meaningful and nonsense words respectively.

Table 5-2: Context-values found in the current study and those reported by Bronkhorst et al. (1993) [based on the data from Bosman and Smoorenburg (1995)]

		Current study	Bronkhorst *et al.* (1993)
Meaningful words	k	1.4	1.3
	j	2.2	2.2
Nonsense words	k	1.0	1.0
	j	2.7	2.8

5.4.2.2 Transfer function

Based on the phoneme score p_e in stationary noise and the values for k, the isolated phoneme score q_e was calculated. Also, per phoneme the local ESTI-value was determined. These results were used to fit the transfer function according to Eq. (5-3). The data and the fit are depicted in Fig. 5-4. Here, the fit from van Van Schoonhoven *et al.* (2022) is also shown. In Table 5-3 the corresponding parameter values are displayed.

Table 5-3: Parameters according to Eq (5-3). The data is visualized in Fig. 5-4.

	α	β	γ
cESTI (current)	-1.3	-2.6	1.17
cESTI Van Schoonhoven *et al.* (2022)	-1.5	-5.4	0.92

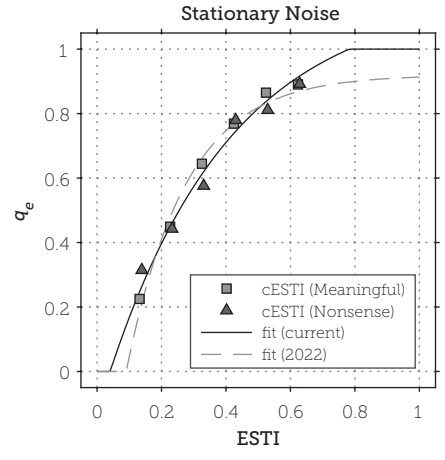


Fig. 5-4: Transfer function which relates the local ESTI-value to the isolated phoneme score (q_e). Both data for meaningful and nonsense words are depicted. See Table 5-3 for details of the fit. The relation found by Van Schoonhoven et al. (2022) is also shown.

The fit shows some similarities to the fit found by Van Schoonhoven *et al.* (2022). However, the main difference is the behavior at higher (> 0.6) ESTI-values. Where the 2022 TF plateaus around 0.9, the current TF reaches its maximum of 1 at an ESTI below 0.8. Unfortunately, both transfer functions were based on ESTI-values lower than 0.7. Therefore, the data at lower ESTI-values determined the curvature of the fit, influencing the behavior at high ESTI-values. Intuitively, it might make sense that optimal conditions (ESTI = 1) lead to maximum intelligibility of 100%.

5.4.2.3 Model predictions in interrupted noise

After the context values and transfer function were estimated, the results in interrupted noise were modelled. Results are depicted in Fig. 5-5. It appears that predictions for nonsense words show the best correspondence to the observed scores. When looking at all data combined, an R^2 (based on $y = x$) of 75% was found (see Fig. 5-6), compared to R^2 = 27% for the original ESTI-model (Van Schoonhoven *et al.*, 2022).

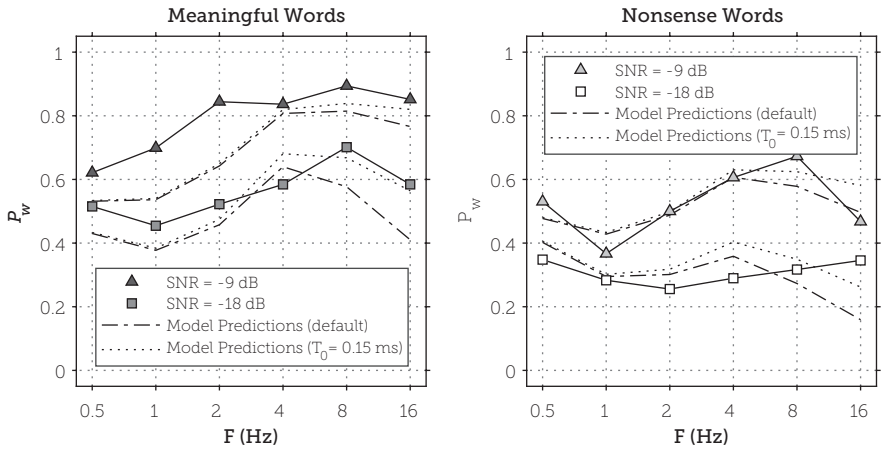


Fig. 5-5: Word scores in interrupted noise as a function of interrupted rates for meaningful words (left panel) and nonsense words (right panel). The different symbols represent the different SNRs. The dash-dot lines represent the cESTI-model predictions with default parameters. The thin dotted lines represent model predictions with an adjusted forward masking parameter T_0, set to 0.15 ms instead of 1 ms. See section 5.5 for more details.

The ellipses in Fig. 5-6 show the confidence intervals of meaningful and nonsense words separately. The longitudinal axis of both ellipses runs parallel to the main diagonal (deviation of the angle < 1°). The distance between the diagonal corresponding to the nonsense words and the main diagonal is < 1%. On the contrary, meaningful word scores are underestimated by 6.6% on average.

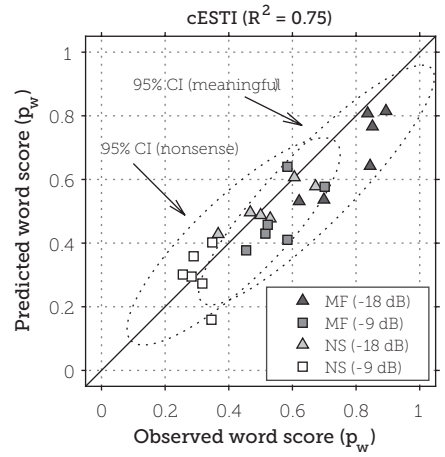


Fig. 5-6: Relation between observed and predicted word scores using the cESTI mode (with default parameter values). The total R^2 is 75%. R^2 for meaningful words equals 38%, R^2 for nonsense words equals 69%. Ellipses show the 95% confidence intervals of the data concerning meaningful (MF) and nonsense (NS) words.

5.5 Discussion

The main goal of the current study was to evaluate the cESTI-model (Van Schoonhoven *et al.*, 2022) using meaningful and nonsense CVC-words in interrupted noise. Word intelligibility in stationary and interrupted noise was measured using normally hearing subjects. After estimating the context factors and a single transfer function based on the word scores in stationary noise, the cESTI-model was applied to the results in interrupted noise at two SNRs and six interruption rates. The explained variance of the model for meaningful and nonsense data combined was 75% (based on the line $y = x$). This prediction accuracy can largely be attributed to the high precision when modelling nonsense words. The values for R^2 for meaningful and nonsense words are 38% and 69% respectively.

5.5.1 Speech intelligibility measurements

5.5.1.1 General

During the noise peaks of the –18 and –9 dB SNR conditions in interrupted noise, the local SNRs are –21 dB and –12 dB respectively. In the current study, word scores in stationary noise were not measured at –21 dB. However, Bosman and Smoorenburg (1995) did test at this SNR using the same speech material (see their Fig. 3a) and found a word score of 0% and a phoneme score of 1% for both meaningful and nonsense words. The current assumption is therefore that intelligibility in interrupted noise is virtually impossible during the noise peaks at a long-term SNR of –18 dB. Note that the STI-value at an SNR of –21 dB is 0, which is in line with this statement.

Therefore, in interrupted noise at –18 dB, the listener only had access to speech information during the gaps in the noise, which is similar for interrupted speech. At infinitely low interruption rates, there is an asymptote at 50% for both word and phoneme recognition. At higher rates, the phoneme score gradually increases, since the probability of perceiving one or more (partial) phonemes also increases. However, the word score first typically decreases at higher rates in both interrupted noise and speech, with a dip at 1 Hz for meaningful words (e.g., Miller and Licklider, 1950; Shafiro *et al.*, 2018). Looking at the current results, the dip is most significant for nonsense words at –9 dB. In the meaningful word data, the dip is not obvious, but there is a plateau visible around 1 Hz at –18 dB. At these rates, the probability of perceiving all three (partial) phonemes is at its lowest point. Around 1 Hz the length of each noise block (and of the glimpses between the noise blocks) is approximately equal to the length of two phonemes, which means that the listener has access to all three phonemes only for about 25% of the time (assuming equal phoneme lengths). At lower rates, the probability that three (partial) phonemes fall between two noise blocks increases. At higher rates, especially when the noise blocks are shorter than one phoneme, the probability to get access to all phonemes through multiple glimpses also increases, leading to a higher word score.

So, increasing the rate beyond ~1 Hz results in more glimpses per word, but at the cost of less speech information per glimpse. Normally, the benefit of more glimpses outweighs the shorter duration, hence the increased intelligibility. The listener profits from more frequent glimpses to reconstruct the utterance (Miller and Licklider, 1950; Cooke, 2006). Eventually, this effect is counteracted by forward masking, directly after the rapid offset of a noise block. Based on the data of Miller and Licklider (1950) the effect of forward masking increases at rates above 2 Hz, since the scores for interrupted speech and interrupted noise scores start to deviate. Optimal intelligibility in interrupted noise lies around a rate of 8 Hz (e.g., Rhebergen *et al.*, 2006), also depending on the nature of the speech.

5.5.1.2 Nonsense words

Both the typical plateau around 1 Hz and the optimum around 8 Hz are not clear when observing the nonsense word scores at –18 dB SNR (see Fig. 5-3 and Fig. 5-5). The nonsense word scores are relatively constant between 0.5 and 16 Hz at 25% to 35%. So, why do the listeners barely benefit from more frequent gaps in the noise like they do in meaningful words? One explanation might be the tendency to give up quicker in difficult circumstances, especially when listening to nonsense words. However, when analyzing this tendency to give up, it does not occur at rates above 1 Hz. At these rates, listeners have access to multiple phonemes and always guess when phonemes are missed, both when listening to nonsense and meaningful words. Another explanation might be the bias towards answering with meaningful words when listening to nonsense words (Bosman, 1989). Sense bias can be defined as the number of meaningful answers relative to the total number of incorrect answers when presenting nonsense words. In stationary noise, this amounts to 46%. As a comparison, the nonsense bias when presenting meaningful words is 10%, which mostly occurs at low rates, since often one or two phonemes are completely masked and the listener answers only what was intelligible. The sense bias in interrupted noise is 48% for both –18 dB SNR and –9 dB SNR and does therefore not differ from the value in stationary noise. Also, there is no relation between sense bias and interruption rate. Note that Bosman (1989) reported a sense bias of 48% for young normally hearing subjects in stationary noise.

Another possible explanation of the pattern in nonsense words at –18 dB might be the relation between phoneme score and word score. As shown in Eq. (5-2), j can be used to express this relation. However, note that in this case j is not a measure of context per se, but more a way to express how high the phoneme score should be to reach a certain word score. A higher value of j means that a higher phoneme score is needed and vice versa. At infinitely low interruption rates, both phoneme score and word score are 50% at –18 dB, represented by j = 1. This immediately emphasizes the statement that in this case j does not represent context, since j = 1 normally represents maximum context.

The relation between phoneme and word score as a function of interruption rate is depicted in Fig. 5-7. As expected, the value of j is low for low rates, after which it increases until a rate of 2 Hz. Apparently, although a dip in word score is typically seen around 1 Hz, listeners need the most phonemic information around 2 Hz in order to reach a certain word score. Above 2 Hz, j drops again. Obviously, j is lower for meaningful words than for nonsense words due to the difference in context of the speech material. At 4 Hz j reaches a stable value for meaningful words at around 95% of the stationary noise value. On the contrary, the j-value for nonsense words never drops below 110% of the stationary noise

value. In other words, at higher rates (≥ 4 Hz) listeners need about the same phoneme score when listening to meaningful words in interrupted noise as they would in stationary noise to reach the same word score. However, to reach the same nonsense word score, they need a higher phoneme score in interrupted noise than they would in stationary noise. An explanation might be that glimpsing speech in interrupted noise is relatively hard in the absence of context. Listeners might not be able to make use of coarticulatory cues as efficiently as they do in stationary noise.

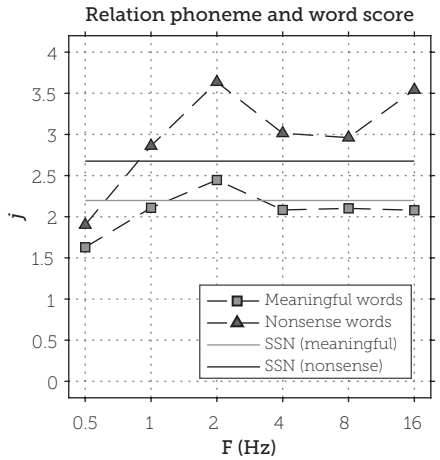


Fig. 5-7: Relation between phoneme and word scores in interrupted noise as a function of interruption rate, expressed by ***j***. Note that ***j*** does not reflect context per se, but only the relation between phoneme and word score. The values for meaningful and nonsense words are depicted. As a reference, the values of ***j*** for SSN are also shown

Glimpsing speech in interrupted noise heavily depends on top-down restoration of the missed elements using syntactic, semantic and lexical constraints, expectations, and context (e.g., Warren, 1970; Verschuure and Brocaar, 1983; Baskent *et al.*, 2010). In nonsense words, the listener need to rely more on coarticulatory, allophonic and durational cues to extract the missing phoneme (Bronkhorst *et al.*, 1993), since other contextual information is limited. Various studies state that top-down processes fail to improve speech intelligibility when the bottom-up signal is degraded beyond a certain point (Baskent, 2012; Patro and Mendel, 2016). The degradations in these studies are mostly in the spectral domain. A possible explanation of the current results is that, in interrupted noise, degradations to the signal cause the listeners to fail to use these coarticulatory

cues to restore the speech. Semantic cues appear to be more robust under these conditions, causing intelligibility of meaningful words to be in line with stationary noise.

Note that in the situation above, the value of ***j*** approaches infinity when one of the three phonemes is always masked, the other two phonemes are fully intelligible and context is low. After all, the word score will be close to 0%, but the phoneme score will be around 67%. According to Eq. (5-2), the logarithm of a very small value in the numerator leads to a large value for ***j***. This might explain the peak at 2 Hz. But at higher rates there is always at least part of a phoneme accessible to the listener, which makes it less likely that this causes the higher value of ***j***. Furthermore, the ***j***-values for nonsense words are generally based on lower word scores than those for meaningful words. This might introduce a bias in the data described above. However, according to Boothroyd and Nittrouer (1988) and Smits and Zekveld (2021) ***j*** tends to increase with increasing word score, which is opposite from the pattern seen in Fig. 5-7.

5.5.1.3 Meaningful words

In meaningful words, a plateau around 1 Hz is visible at –18 dB SNR, but not at –9 dB SNR (see Fig. 5-3 and Fig. 5-5). In interrupted noise, the listener must combine information during the glimpses and during the noise fragments. The ability to reconstruct the complete word depends on the amount of context in the speech material and/or on speech information present during the noise peaks. At –9 dB there appears to be just enough information available during the noise blocks that can be combined with the speech information during the glimpses in order to make use of context effectively.

5.5.2 Model predictions

Prediction accuracy is higher for nonsense words than for meaningful words. Two observations regarding the model predictions are 1) the underestimation of meaningful word scores and 2) the rapid drop off at high interruption rates (8 and 16 Hz) at –18 dB.

5.5.2.1 Context factors

The underestimation of meaningful word scores is especially true for 0.5, 1 and 2 Hz. When noise gets a more continuous character at higher rates, model predictions are more accurate. Van Schoonhoven *et al.* (2022) observed a similar pattern and successfully tested the hypothesis that modelling using the transfer function and context factors obtained in quiet was more appropriate for lower interruption rates. The underlying motivation was the fact that interrupted noise at low rates does not have a continuous character anymore.

For example, at 0.5 Hz, the listener perceives separate noise blocks of 1 second, separated by silent intervals of 1 second. This is more a matter of combining intelligibility in noise with intelligibility in quiet, and not so much of glimpsing speech.

When using the transfer function and the context factors from Van Schoonhoven *et al.* (2022) for interrupted speech, an improvement in model accuracy is seen for rates below 4 Hz. In the current study, when using the new context factors only at the rates of 4 Hz and higher, R^2 drops from 75% to 70% for meaningful and nonsense words combined. This decreased accuracy is expected, since interrupted noise at higher rates behaves more like continuous noise. However, applying the context factors only to rates below 4 Hz R^2 increases from 0.68 to 0.85 for nonsense and meaningful words combined. This increase can largely be attributed to the higher explained variance for meaningful words at low rates: 73% versus 11%. This confirms the earlier hypothesis that the transfer function and context factors derived in speech in quiet and in interrupted speech are more suitable at lower interruption rates.

5.5.2.2 Forward masking

Another aspect is the sharp drop off that the model predicts at 8 and 16 Hz at –18 dB in both meaningful and nonsense words. It appears that the peak word score of the model (at 4 Hz) does not correspond to the peak word score of the data. This pattern was not observed by Van Schoonhoven *et al.* (2022) where the data by Miller and Licklider (1950) was modelled. In their Fig. 6 it is clear that the model predictions closely follow the observed data and that the drop off occurs above 20 Hz. As mentioned earlier, forward masking counteracts the benefit of more glimpses per word at higher rates. So, either the model underestimates the benefit of more glimpses per word, or the effect of forward masking is stronger in the model predictions than in the observed data.

One important difference between the intelligibility data of Miller and Licklider (1950) and the current study is the presentation level. In their study, the speech level was held constant at 90 dB (SPL) and the noise level was varied based on the desired SNR. For a long-term SNR of –15 dB, this means that the noise peaks were as high as 108 dB (SPL). In the current study, the noise level was fixed at 65 dB (A), which means that the noise peaks were at 68 dB (A). This 40 dB difference influences the forward masking function.

The forward masking as incorporated in the ESTI-model by Van Schoonhoven *et al.* (2019) was defined by Ludvigsen (1985) as:

$$MT = N - \frac{\log(t_{pm}/T_0)}{\log(T_f/T_0)}\,[N - HTL] \tag{5-5}$$

for $T_0 < t_{pm} < T_f$

Here, N is the noise level, MT is the masked threshold, t_{pm} is the post-masker duration, T_0 is the intersection point (set at 1 ms), T_f is the recovery time (set at 150 ms) and HTL is the hearing threshold. After a noise peak, MT remains at N during T_0. Between T_0 and T_f, MT decreases exponentially to HTL, after which it remains at that level. The post-masker duration where the masked threshold decreases with 20 dB relative to the initial value more than doubles for the different noise values (3.0 ms for 108 dB versus 7.5 ms for 68 dB). Consequently, at a given SNR the predicted influence of forward masking on speech intelligibility is larger for lower noise levels.

Ludvigsen (1985) based his model on experiments on four normally hearing and 13 sensorineurally hearing-impaired subjects. Masked thresholds for 1.5 s long pure tones in octave band filtered white noise were compared to masked thresholds in interrupted octave band filtered white noise. The interruption rate was 14.3 Hz with a 50% duty cycle, resulting in noise bursts and gaps of 35 ms long. Two masker levels were used. In order to fit the data to Eq. (5-5), the author set T_f to 200 ms, based on earlier results by Plomp (1964) and Kidd and Feth (1981), T_0 to 3 ms and t_{pm} to 35 ms. The latter value corresponds to the length of gaps in the noise. Using these values, the experimental data was predicted reasonably well. Although not described in the original publication, the value for T_0 is probably taken from Plomp (1964), where the minimal detectable gap between two noise pulses was investigated as a function of sensation level. When the sensation level of both pulses was the same, the just noticeable gap was optimally between 2 and 3 ms.

Jesteadt *et al.* (1982) proposed another forward masking model, based on a different type of experiment. They presented a pure tone masker of 296 ms, followed by a pure tone probe signal of 24 ms of the same frequency with a varying delay. Thresholds of the probe tone were obtained in four normally hearing subjects. The delays between masker and probe ranged from 5 to 40 ms and were measured between the 0-voltage points of the masker offset and signal onset ramp. Masker levels depended on frequency and ranged between 20 and 90 dB (SPL). They fitted the parameters a, b and c separately for five octave frequencies to calculate the amount of masking (M):

$$M = a\big(b - \log\big(t_{pm}\big)\big)(N - HTL_{masker} - c) \quad (5\text{-}6)$$

The masked threshold $(\boldsymbol{MT})$ is then found by adding the signal threshold in quiet:

$$MT = M + HTL_{signal} \quad (5\text{-}7)$$

Where the function of Ludvigsen remains at $\boldsymbol{N}$ until $\boldsymbol{T_0}$ ms, the model by Jesteadt estimates $\boldsymbol{MT} = \boldsymbol{N}$ at < 1 ms for all octave frequencies and < 0.1 ms for 250, 500 and 1000 Hz.

Moore and Glasberg (1983) conducted a similar experiment, but now using noise as a masker. The duration of the steady state part of all signals was 20 ms. They varied the signal delay between 0 and 20 ms when calculated between the 0-voltage points of the masker offset and signal onset ramp. Note that fitting Eq. (5-6) for data points at $\boldsymbol{t_{pm}}$ = 0 ms would yield an infinite slope of the masking function. Therefore, the authors based their fit on the –6 dB points of the *offsets* of both the masker and the signal. Similar to the results of Jesteadt et al. (1982), they found a linear relation between $\log\big(t_{pm}\big)$ and the amount of masking between the shortest and longest signal delay they investigated. Furthermore [and also similar to the results by Jesteadt *et al.* (1982)], when plotting the amount of masking as a function of masker level (i.e., growth-of-masking function), the slope was smaller than unity for all signal delays that were tested. This slope is generally close to one for simultaneous masking and smaller than one for non-simultaneous masking Moore (2007). Note that the slope of the growth-of-masking function corresponding to Eq. (5-5) by Ludvigsen (1985) is unity when $t_{pm} = T_0$.

The derivative of the growth-of-masking function of Jesteadt *et al.* (1982) with respect to the masker level is:

$$\frac{dM}{dN} = a\big(b - \log\big(t_{pm}\big)\big) \overset{\frac{dM}{dN}=1}{\Longrightarrow} t_{pm} = 10^{b-1/a} \quad (5\text{-}8)$$

Based on the derived values for a and b by Jesteadt *et al.* (1982) in their Table III, the value for $\boldsymbol{t_{pm}}$ in Eq. (5-8) ranges between 2.4e-4 and 0.3 ms. Using the values by Moore and Glasberg (1983) in their Table III, $\boldsymbol{t_{pm}}$ ranges between 1.3 and 1.7 ms. The latter values are difficult to interpret, since the delay was defined as the duration between the masker offset and the signal offset with a signal duration of 20 ms (excluding the ramps). When correcting for the signal duration, the post-masker duration would be negative.

So, what is the most suitable forward masking function parameter to adjust, in order to increase the model accuracy? The chosen value of $\boldsymbol{T_f}$ = 150 ms was based on speech intelligibility data by Van Schoonhoven *et al.* (2019) and, although it slightly deviates from the value of 200 ms suggested by Ludvigsen (1985), it falls well within the normal range of 100 – 200 ms (Moore, 2007). When varying $\boldsymbol{T_f}$ to optimize the model predictions, values below 100 ms are found. Since this deviates too much from values normally found in the literature, the authors chose to keep $\boldsymbol{T_f}$ fixed, On the other hand, based on the above discussion, the original value of 1 ms for $\boldsymbol{T_0}$ might not be optimal.

In order to evaluate the influence of $\boldsymbol{T_0}$ on the model predictions, this parameter was varied between 0.01 and 3 ms. In Table 5-4, the explained variance is displayed as a function of $\boldsymbol{T_0}$. Both data from the current study and from Van Schoonhoven *et al.* (2022) are displayed here. Word scores from the current study appear to be predicted better by the cESTI-model when $\boldsymbol{T_0}$ is smaller, with an optimum between 0.05 and 0.3 ms. However, $\boldsymbol{R^2}$ for the 2022 data slightly decreases. At $\boldsymbol{T_0}$ = 0.15 ms, the current data appears to be optimally predicted. The thin, dotted lines in Fig. 5-5 show the model predictions with $\boldsymbol{T_0}$ = 0.15 ms. The increase in meaningful word score prediction accuracy from 38% to 60% is largest at this value. A slight decrease from $\boldsymbol{R^2}$ = 0.92 to 0.89 is seen for the 2022 data. Also, a decrease in model accuracy for the data by Van Schoonhoven *et al.* (2019) using Dutch sentence material is seen. This is particularly true for maskers with artificial fluctuations and artificial fine structure (mostly interrupted

Table 5-4: values of R^2 for different values of T_0. Values are provided for the current data and for the data by Van Schoonhoven et al. (2022)

	Meaningful and nonsense combined	Meaningful		Nonsense
	Current	Current	2022	Current
T_0 = 0.01 ms	0.81	0.60	0.89	0.68
T_0 = 0.05 ms	0.82	0.61	0.89	0.72
T_0 = 0.1 ms	0.82	0.61	0.89	0.74
T_0 = 0.15 ms	0.83	0.60	0.89	0.75
T_0 = 0.2 ms	0.82	0.60	0.89	0.76
T_0 = 0.3 ms	0.82	0.58	0.90	0.76
T_0 = 0.5 ms	0.81	0.54	0.91	0.75
T_0 = 1.0 ms	**0.75**	**0.38**	**0.92**	**0.69**
T_0 = 2.0 ms	0.49	-0.28	0.85	0.49
T_0 = 3.0 ms	0.30	-0.73	0.80	0.17

noises). Here, the model tends to overestimate intelligibility compared to the original value of T_0. The explained variance of this data based on $y = x$ drops from 0.88 to 0.81 when using T_0 = 0.15 ms.

A cause for the observed discrepancies might be that the current model parameters were fitted by Van Schoonhoven *et al.* (2019) on 50% intelligibility data, all measured using a fixed noise level of 65 dB (A). Firstly, no complete transfer function was available, so it is not clear how the model behaves at other points on the psychometric curve. Secondly, the forward masking function behaves different at different noise levels and this was not incorporated in the fitting of the parameters. However, despite these discrepancies, the model still functions well, with 75% explained variance for the standard forward masking parameters, and 83% for the adjusted value of T_0.

5.6 Conclusion

This study evaluated the cESTI-model (Van Schoonhoven *et al.*, 2022) using monosyllabic words with different degrees of context presented in interrupted noise. A single transfer function was used to estimate the isolated phoneme score based on the local ESTI-values. After that, context was introduced to predict meaningful and nonsense word scores.

Word scores were relatively low when nonsense words were presented at an SNR of –18 dB, especially at higher rates. Under these conditions, listeners barely profited from more frequent glimpses at the speech. Glimpsing speech might be more difficult in the absence of context in adverse listening conditions, possibly because listeners rely less on coarticulatory cues than in stationary noise.

The cESTI-model performed well and explained 75% of the total variance. However, predictions for meaningful word scores were relatively poor, mainly due to inaccuracies at low interruption rates and a large effect of forward masking. Adjusting parameters of the forward masking function improved the accuracy of the model to a total explained variance of 83%, while the predicted power for previously published cESTI data remained the same.

5.7 Acknowledgements

This work was financially supported by the Heinsius-Houbolt foundation. The authors would like to thank Arjan Bosman for providing the speech material.

CHAPTER 6

A CONTEXT-BASED MODEL TO PREDICT THE INTELLIGIBILITY OF SENTENCES IN NON-STATIONARY NOISES

Van Schoonhoven, J., Rhebergen, K.S., Dreschler, W.A. (2023)
Manuscript to be submitted for publication in the Journal of the Acoustical Society of America

6.1 Abstract

The context-based Extended Speech Transmission Index or cESTI (Van Schoonhoven *et al.*, 2022) was successfully applied to predict the intelligibility of monosyllabic words with different degrees of context in interrupted noise. The current study aimed to use the same model for the prediction of sentence intelligibility in different types of non-stationary noise. The necessary context factors and transfer functions were based on values found in existing literature. The cESTI performed similar to or better than the original ESTI when noise had speech-like characteristics. We hypothesize that the remaining inaccuracies in model predictions can be attributed to the limits of the modelling approach with regard to mechanisms like modulation masking and informational masking.

6.2 Introduction

The Extended Speech Transmission Index (ESTI) was introduced by Van Schoonhoven *et al.* (2019) and was an extension of the original Speech Transmission Index or STI (Houtgast and Steeneken, 1978; Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985; IEC60268-16, 2011). See chapter 3 for an elaborate description of the ESTI. The classic STI was based on the observation that a reduction of the modulation depth in speech leads to a decrease in intelligibility. Especially the effects of linear distortions in the field of room acoustics (like reverberation and stationary noise) were predicted accurately. However, the classic STI is relatively inaccurate when distortions are nonlinear and when maskers are non-stationary (IEC60268-16, 2011).

To deal with the latter shortcoming of non-stationary maskers, Van Schoonhoven *et al.* (2019) suggested calculating the STI per time frame (ESTI). This concept was based on the work by Rhebergen and Versfeld (2005), who proposed a similar extension of the Speech Intelligibility Index (SII). With the introduction of forward masking, the ESTI accurately predicted the point of 50% intelligibility (cSNR) of sentences in various types of non-stationary background noise. However, inaccuracies remained when noises had speech-like properties. It was hypothesized by the authors that this could be an effect of Modulation Masking (MM), Informational Masking (IM) and/or context effects. Van Schoonhoven *et al.* (2022) suggested to add context to the ESTI-model (cESTI) to check the latter hypothesis. To add context to the model, the ESTI-value was calculated per phoneme of monosyllabic words. Using a transfer function, the isolated phoneme score for each element was estimated based on the ESTI per phoneme. As a last step, one of two existing context models was applied to obtain the word scores (Boothroyd and Nittrouer, 1988; Bronkhorst *et al.*, 1993).

6.2.1 Context models

For an elaborate description of the Boothroyd and Nittrouer context model and the Bronkhorst context model the reader is referred to the original work or to Van Schoonhoven *et al.* (2022). See also chapter 4. Both context models relate the intelligibility of speech elements to the intelligibility of whole utterance. This can be the phoneme score to the word score, but also the word score to the sentence score. The manner in which this relation is modelled, is fundamentally different between both models.

The Boothroyd and Nittrouer model uses the k- and the j-factor. The k-factor [see Eq. (6-1)] is defined as the ratio between the log error probabilities of an element in context and in isolation. When no context is available, k is equal to 1. When the listener utilizes context, the recognition error of the element

decreases and k will be larger than 1. The j-factor [see Eq. (*6-2*)] is defined as the ratio between the log recognition probabilities of a whole (p_w) and of an element in context (p_e). When no context is available, the listener needs access to all the elements to perceive the whole utterance, which means that j is equal to the number of elements in the whole (N). When context is used, j decreases ($1 \leq j \leq N$).

$$k = \frac{\log(1 - p_e)}{\log(1 - q_e)} \tag{6-1}$$

$$j = \frac{\log(p_w)}{\log(p_e)} \tag{6-2}$$

The Bronkhorst model consists of a sensory stage and a context stage. In the first stage, the intelligibility of elements (q_e) is only based on the sensory information that is available to the listener. In this stage, the probability that a complete speech token of n elements is identified based on sensory information alone is the product of the individual probabilities: $Q_0 = (q_e)^n$. Here, subscript 0 refers to the number of errors made in the sensory stage.
In the second stage, the context parameters c_i are introduced. Here, c_i represents the probability of correctly identifying one of the i missing elements. For example, the probability of missing one of three elements in the sensory stage equals $3(q_e)^2(1 - q_e)$. The probability of correctly identifying the missed element in the second stage using context is equal to c_1. Consequently, correct identification of the whole speech token in this example using both sensory and contextual information equals

$$p_w = 3c_1(q_e)^2(1 - q_e) \tag{6-3}$$

Van Schoonhoven *et al.* (2022) used existing intelligibility data of monosyllabic words by Miller and Licklider (1950) to evaluate the cESTI-model. The authors observed that both the addition of the Bronkhorst, and the Boothroyd and Nittrouer model led to a clear improvement when speech was masked by interrupted noise at lower modulation rates (< 5 Hz). Since the addition of both context models led to similar results, it was suggested by the authors that the simpler Boothroyd and Nittrouer model was the more suitable choice as part of the cESTI-model. Van Schoonhoven *et al.* (2023) measured the intelligibility of meaningful and nonsense CVC-words in normally hearing subjects and compared the results to estimations using the cESTI. Model accuracy was higher in comparison with the original ESTI-model, although predictions of meaningful word intelligibility was low, particularly due to remaining inaccuracies at low interruption rates and a relatively large effect of forward

masking. After application of an alternative forward masking function the prediction accuracy of the cESTI model increased to 83%, while the predicted power of previously ESTI and cESTI data remained similar.

6.2.2 Goal current study

The main motivation to add context to the ESTI-model in the first place was the observation by Van Schoonhoven *et al.* (2019) that prediction of sentence intelligibility was inaccurate when the noise had speech-like characteristics. However, thus far only monosyllabic words were used to evaluate the cESTI-model. Therefore, the current work focused on the evaluation of the cESTI-model using the same sentence material that was used by Van Schoonhoven and colleagues in 2019.

6.3 Materials and methods

Generally, three steps were necessary to use the cESTI-model for the prediction of sentence intelligibility (see Fig. 6-1). Two of these steps were the same as described by Van Schoonhoven *et al.* (2022; 2023). See also chapters 4 and 5. First, the ESTI-value per phoneme was related to the isolated phoneme score (q_e) using an appropriate transfer function. Next, q_e was converted to the isolated sentence word score (Q_e) using a context-based transform (see Dingemanse and Goedegebure, 2019). Finally, Q_e was converted to a sentence score (P_w), again using a context model. Note that two types of word scores are used in this section. The isolated word score in a sentence is referred to as Q_e. Besides this, and consistent with the terminology in chapters 4 and 5, the CVC word score is referred to as p_w.

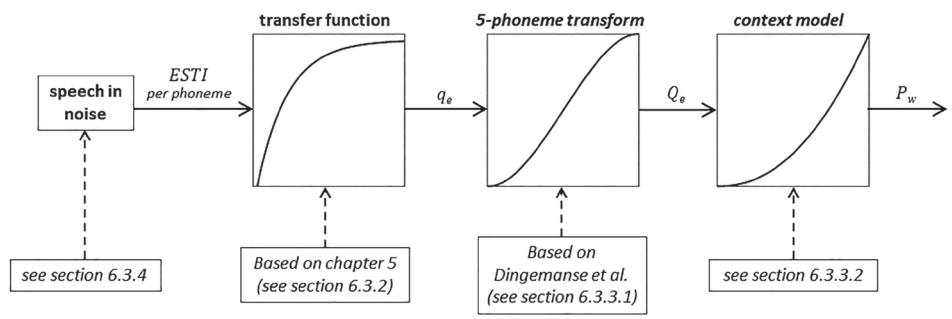


Fig. 6-1: Steps to go from the ESTI per phoneme to the sentence score P_w

The context models by Bronkhorst *et al.* (1993) or by Boothroyd and Nittrouer (1988) were both used as part of the cESTI-model. The terms $cESTI_1$ and $cESTI_2$ are used to refer to these models respectively, consistent with Van Schoonhoven *et al.* (2022). The Bronkhorst model was used as the basis of the current work. This choice was related to the availability of the values of the context parameters and will be explained below. The context values of the Boothroyd and Nittrouer model were derived in a later stage from those of the Bronkhorst model.

6.3.1 Sentence material

A prerequisite for the choice of sentence material was the ability to calculate the context factors necessary for the estimation of sentence scores. In the literature values of chapter 3, only the SNR at 50% sentence intelligibility (cSNR) was provided. This required the use of context factors estimated elsewhere. The work by Dingemanse and Goedegebure (2019) was chosen for this purpose. They calculated context factors based on 6-word sentences uttered by a female speaker (Versfeld *et al.*, 2000). Note that the sentences in the original corpus consist of four to nine words, so only a subset of these sentences was used. Furthermore, Bronkhorst *et al.* (1993) also calculated the context factors for the speech corpus developed by Plomp and Mimpen (1979), based on the intelligibility data by Bosman and Smoorenburg (1995), but only the data corresponding to the cSNR in quiet was used. It is not clear how these values relate to the context factors in noise and since the same study also showed that the difference between the use of context around threshold in quiet and in noise is relatively large, these context factors were not used here.

Since the context factors for female sentences from Versfeld *et al.* (2000) were available, only studies using this speech corpus were selected from Van Schoonhoven *et al.* (2019). Because Versfeld and colleagues used the same selection procedure for male and female sentences, the contextual information in both corpora was expected to be the same. The speaking style is of course different, but the sentences spoken by a male speaker were included in the current analyses. This was a pragmatic choice in order to increase the amount of data available. For more information on the used sentence material, the reader is referred to Appendices B and D. See Fig. 6-2 for the properties of the sentence material. In this figure, the similarities between the sentences spoken by the female and male speaker can also be observed.

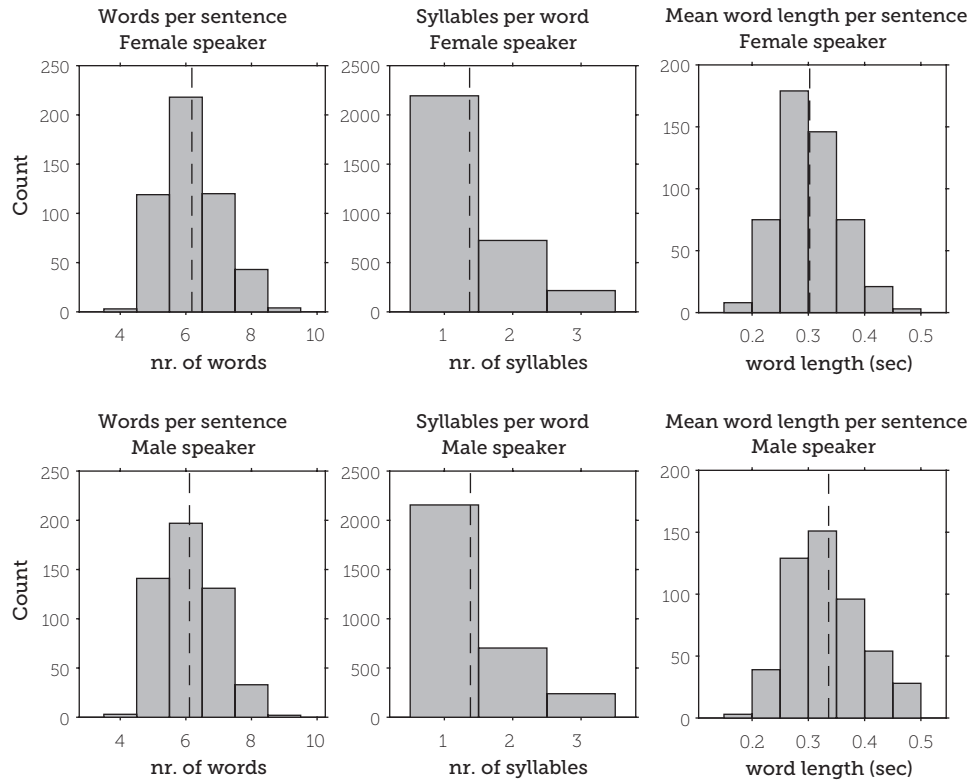


Fig. 6-2: Properties of the sentence material by Versfeld et al. (2000). The top row shows the data of the female speaker and the bottom row the data of the male speaker. The dashed lines represent the average values.

6.3.2 Transfer function

As with monosyllabic words, the transfer function for sentences aims to relate the local ESTI-value to the phoneme score in isolation (q_e) using Eq. (6-4).

$$q_e = \gamma + \alpha e^{\beta ESTI} \tag{6-4}$$

The best option was to use the transfer function that was obtained by Van Schoonhoven *et al.* (2023). They found values of –1.3, –2.6, and 1.17 for α, β, and γ respectively. A disadvantage of this approach is that words in sentences are more than two times shorter than the CVC-words presented in isolation [302 ms versus 670 ms for the female speaker of Versfeld *et al.* (2000)] and have on average 1.4 (+/– 0.2) syllables and 4.4 phonemes. As a result, the average

phoneme length in CVC-words is approximately three times longer than in sentences (223 ms versus 69 ms). How the phoneme length influences the isolated phoneme score will be discussed in section 6.5.

The estimated intelligibility using the ESTI was based on the assumption that listeners require the same ESTI-value in SSN as they do in non-stationary noise to reach 50% intelligibility (Van Schoonhoven *et al.*, 2019). To this end, the ESTI-value at cSNR in SSN was calculated per study and used as a reference for the conditions with non-stationary noise of the same study (i.e., the same group of listeners). To be able to use the same approach in the current chapter, the transfer function needed to be customized using the cSNR in SSN for each study separately. To achieve this, the estimated context factors (see section 6.3.3) were used to determine the value of q_e based on the point of 50% sentence intelligibility in SSN. In this way, per study, one point on the transfer function between ESTI and q_e was known. The transfer function found for CVC-words was then customized to include this point with minimal adjustment to the function itself, by minimizing the squared difference between the old and new values of α, β, and γ of Eq. (*6-4*). In this way, a family of transfer functions was derived from the transfer function found by Van Schoonhoven *et al.* (2023).

6.3.3 Context models

The goal of the context models was to relate the isolated phoneme score q_e to the sentence score P_w. The most challenging part was to relate q_e to Q_e. The transfer function to determine q_e was based on CVC-words, which have different characteristics (e.g., length, number of phonemes, and number of syllables) than words in sentences (Q_e). Therefore, a transformation was necessary.

6.3.3.1 Transform (q_e to Q_e)

Dingemanse and Goedegebure (2019) based this transformation on the average number of phonemes in the sentence words, which they set to five (the first integer larger than the true average of 4.4). To obtain the isolated phoneme score, they first transformed p_w in CVC-words back to q_e using the context parameters that were estimated for these CVC-words. Note that in the current work, q_e is already available using the ESTI and an appropriate transfer function, so p_w was not used. After q_e was known, the authors estimated Q_e based on P_w using the context parameters that were found for sentences. Finally, they estimated a new set of context factors for the transformation from q_e to Q_e, based on words of five phonemes. The same approach was used in the current study. However, Dingemanse and Goedegebure used this transformation model only with data obtained in cochlear implant recipients. Since the authors

found that these subjects tend to make more use of contextual information than normally hearing subjects, these values were not applicable for the current analyses. Therefore, the context values of the 5-phoneme transform were estimated based on the sentence data from the literature used in the current work. An important limitation was the number of datapoints available, since this estimation was only based on the averaged SNR at 50% intelligibility in SSN. In the optimization procedure to estimate the optimal context values for the 5-phoneme transform, c_5 was set to 0 (Bronkhorst *et al.*, 1993). Dingemanse and Goedegebure (2019) suggested that c_1 and c_2 take on a higher value, since it is relatively easy to guess a 5-phoneme word when only one or two phonemes are missed. On the other hand, c_3 and c_4 were believed to take on lower values, since it is relatively hard to guess the word when half of the phonemes are missed. To limit the number of free parameters in the current optimization procedure, the aim was to maximize c_1, while keeping $c_2 > 0.5$, $c_3 < 0.5$, $c_4 < 0.2$ and $c_5 = 0$, while satisfying the criterion $c_1 > c_2 > c_3 > c_4 > c_5$. The boundary conditions itself are rather arbitrary, but follow the above arguments, and follow the typically descending trend of context parameters (Bronkhorst *et al.*, 1993; Bronkhorst *et al.*, 2002).

6.3.3.2 Sentences (Q_e to P_w)

The context factors by Dingemanse and Goedegebure (2019) were based on 6-word sentences. In the literature data, all sentences of the original corpus were used and consisted of four to nine words. Bronkhorst *et al.* (2002) suggested that the value of the context parameters can be expressed as a function of the number of elements per sentence. They obtained Eq. (6-5) as the relation between the parameters from $c_{1,n}$ to $c_{n\text{-}1,n}$ (with n as the number of elements per sentence), and Eq. (6-6) as the relation between the values for $c_{1,n}$ for different values of n. Based on the available context factors for 6-word sentences, α_c and c_{min} were estimated using Eq. (6-5). The value for α_l was obtained using the suggested values by Bronkhorst *et al.* (2002) and Smits and Zekveld (2021).

$$c_{i+1,n} = c_{i,n}^{\alpha_c} + \left(1 - c_{i,n}^{\alpha_c}\right)\left(\frac{c_{min}}{1 + c_{min}^{\alpha_c - 1}}\right) \tag{6-5}$$

$$c_{1,n+1} = c_{1,n}^{\alpha_l} + \left(1 - c_{1,n}^{\alpha_l}\right)\left(\frac{c_{min}}{1 + c_{min}^{\alpha_l - 1}}\right) \tag{6-6}$$

6.3.3.3 Boothroyd and Nittrouer model

The Bronkhorst model and the Boothroyd and Nittrouer model both aim to estimate the effect of context using a different approach. Bronkhorst *et al.* (2002) expressed j and k as a function of c_i (and q_e) in their appendix. These mathematical relations were used in the current study to estimate the context factors j and k. Since these context factors depend on the element score (e.g., Boothroyd and Nittrouer, 1988; Dingemanse and Goedegebure, 2019), the values were calculated using the value of q_e corresponding to 50% sentence intelligibility.

6.3.4 Estimation of sentence score

In the previous paragraphs, the methods of obtaining the context factors and the transfer functions were described. Based on this information, the sentence score was estimated. The calculation scheme of estimating P_w was similar to the earlier works on the cESTI (Van Schoonhoven *et al.*, 2022; 2023). For each condition, the ESTI per phoneme in the sentence was calculated for all possible phase shifts of the sentence with respect to the noise (with steps of 2 ms). Note that no real sentences were used, but the location of a virtual phoneme with respect to the noise was used to estimate the ESTI of that phoneme. To this end, the timing of the phonemes and words in virtual sentences was based on the data of the actual sentences used in the speech intelligibility tests (see Table 6-1). A total of 100 virtual sentences were constructed to satisfy these conditions. When the ESTI per phoneme was known, the transfer function was used to calculate the values of q_e. The 5-phoneme transform was then applied to determine Q_e, after which P_w could be calculated. In Fig. 6-3 the steps are visualized to calculate P_w based on the ESTI.

Table 6-1: Average length of the speech elements used in the current study.

	Average length (+/– s.d.)
Sentences	1.8 s (+/– 0.2)
Words	0.3 s (+/– 0.05)
Phonemes	61 ms[ix] (+/– 11)

ix Note that this value is lower than the 69 ms reported earlier. This is caused by the difference in the actual number of phonemes per word (4.4) compared to the number of phonemes per word used for the model (5).

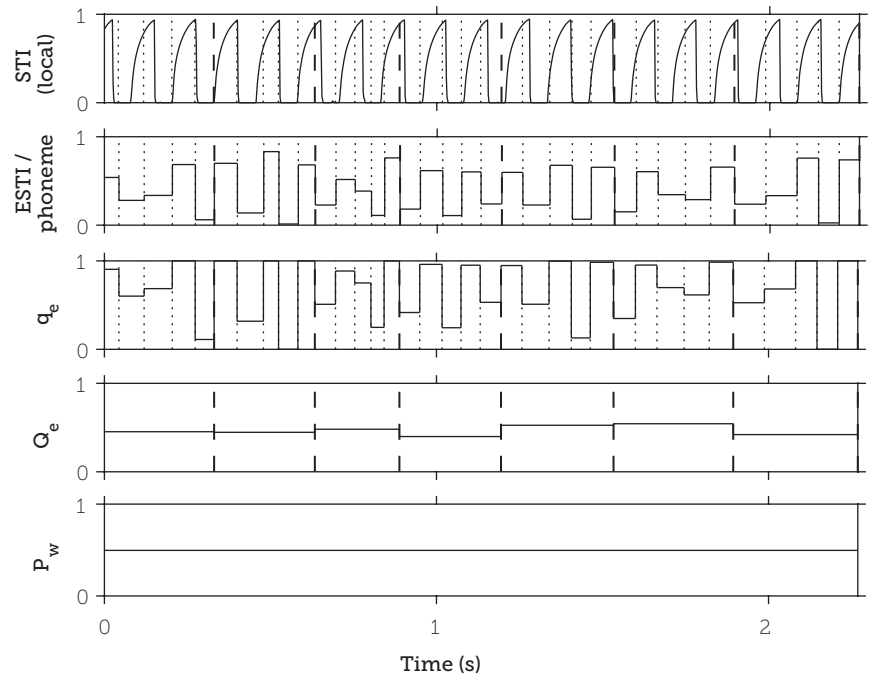


Fig. 6-3: Calculation steps to estimate P_w using the ESTI of an example sentence in 8 Hz interrupted noise at cSNR. The width of the plot represents one sentence. Thin dotted vertical lines represent phoneme boundaries. Thick dashed lines represent word boundaries. The top plot shows the variation in local STI based on the SNR. The grand average of these values would yield the ESTI. The second plot shows the ESTI-value per isolated phoneme. Using the transfer function, the isolated phoneme score per phoneme (q_e) is calculated (third plot). Based on the context factors of the 5-phoneme transform, the sentence word score in isolation (Q_e) is then estimated for each word (fourth plot). Finally, the sentence score (P_w) was calculated (fifth plot).

6.4 Results

6.4.1 Transfer function

The parameters from Eq. (6-4) were optimized to find the appropriate transfer function per study. This resulted in a family of transfer functions around the original version, caused by the variation in cSNRs over all the studies (see Fig. 6-4). The deviation of the fitting parameters ranged between 0 and 5% compared to the original values found by Van Schoonhoven *et al.* (2023).

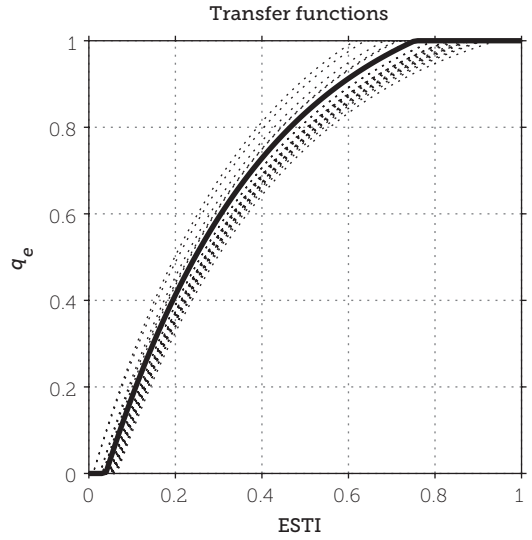


Fig. 6-4: Family of transfer functions used in the current study based on Eq. (6-4). The thick solid line represents the main function found by Van Schoonhoven et al. (2023). The other functions (thin dotted lines) were estimated based on the main transfer function, the cSNR per study of the data analyzed by Van Schoonhoven et al. (2019), and the context factors from Table 6-2.

6.4.2 Context factors

Context factors for the 5-phoneme transform and for the sentences were estimated using the Bronkhorst model. The context factors for the Boothroyd and Nittrouer model were derived from these values. All context factors are depicted in Table 6-2.

6.4.2.1 Transform (q_e to Q_e)

The optimization procedure in order to fit the context parameters for the transform (from q_e to Q_e) resulted in c_1 = 0.65, c_2 = 0.54, c_3 = 0.26 and c_4 = 0.05 (c_5 was set to 0). Especially c_1 and c_2 are lower than the values of 0.98 and 0.89 that were obtained by Dingemanse and Goedegebure (2019), calculated for cochlear implant recipients. This finding itself is expected since the authors already suggested that these subjects make more use of context than normally hearing subjects. However, especially the value for c_1 appears low, since the expectation is that a listener — when identifying four out of five phonemes correctly — has a higher probability than 65% of guessing the fifth element. This will be discussed more elaborately in section 6.5.

Table 6-2: All context values found for the different sentence lengths (first column) and for the 5-phoneme transform. Note that k and j from the Boothroyd and Nittrouer model were derived from c_i per sentence length at the point of 50% sentence intelligibility using the mathematical relations provided by Bronkhorst et al. (2002).

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	k	j
4-word	0.80	0.65	0.45	0	-	-	-	-	-	2.0	2.3
5-word	0.86	0.75	0.58	0.63	0	-	-	-	-	2.2	2.6
6-word	0.91	0.83	0.70	0.50	0.28	0	-	-	-	2.2	2.8
7-word	0.94	0.88	0.79	0.63	0.42	0.21	0	-	-	2.3	2.9
8-word	0.96	0.92	0.85	0.73	0.56	0.34	0.15	0	-	2.4	3.0
9-word	0.97	0.95	0.90	0.82	0.68	0.48	0.26	0.10	0	2.4	3.0
Transform	0.65	0.54	0.26	0.05	0	-	-	-	-	1.5	3.5

6.4.2.2 Sentences (Q_e tot P_w)

When fitting Eq. (6-5) to the context factors found by Dingemanse and Goedegebure (2019), the parameters c_{min} and α_c were used as free fitting parameters, which resulted in a root-mean-square error ($RMSE$) of 0.017 for α_c = 1.89 and c_{min} = 0. When using one free fitting parameter and fixing c_{min} to the value of 0.035 suggested by Bronkhorst *et al.* (2002), the $RMSE$ increased slightly to 0.019. Due to the small amount of data available, the latter value was chosen, corresponding to α_c = 2.00. This is lower than the value 2.69 found by Bronkhorst *et al.* (2002) for German sentence material, but higher than the value of 1.78 that was found by Smits and Zekveld (2021) who used the same speech corpus as the current work, but only selected 6-word sentences. The latter value was estimated using intelligibility data of the same sentence material, but at four different SNRs. A higher value of α_c means that there is a sharper drop of c_i at higher values of i.

The parameter α_l determines the relation between the values of c_1 for sentences of different lengths. Bronkhorst *et al.* (2002) found a value of 2.23 for meaningful German sentences. A higher value means larger differences between c_1-values for different sentence lengths. The above value was used as a starting point for the current estimations. Using the context parameters that resulted from the above parameters, the isolated phoneme score q_e corresponding to 50% sentence intelligibility was 48%. However, the estimated sentence intelligibility at this value for q_e ranged between 32% for 4-word sentences to 75% for 9-word sentences. The weighted average (using the data in the left panel of Fig. 6-2) of the intelligibility for all sentence lengths equaled 50%, but the variance was relatively large. Since Versfeld *et al.* (2000) selected the sentences to be equally intelligible, α_l was varied to minimize variance between sentence intelligibility

at values of q_e between 10% and 90%. This resulted in α_l = 1.56. This reduced the variance at 50% intelligibility and now ranged between 49% and 52%.

6.4.3 Model predictions

The model predictions are displayed in Fig. 6-5. The results were separated based on three noise types. Noises with an artificial envelope and an artificial fine structure are for instance interrupted noises or sinusoidally intensity-modulated noise. Both other noise types have speech-like envelopes, but differ in their fine structure, which is either artificial [e.g., ICRA-5 noise by Dreschler *et al.* (2001)] or speech-like [e.g., a competing speaker or the ISTS by Holube *et al.* (2010)]. All corresponding values are depicted in Table E-1 in Appendix E.

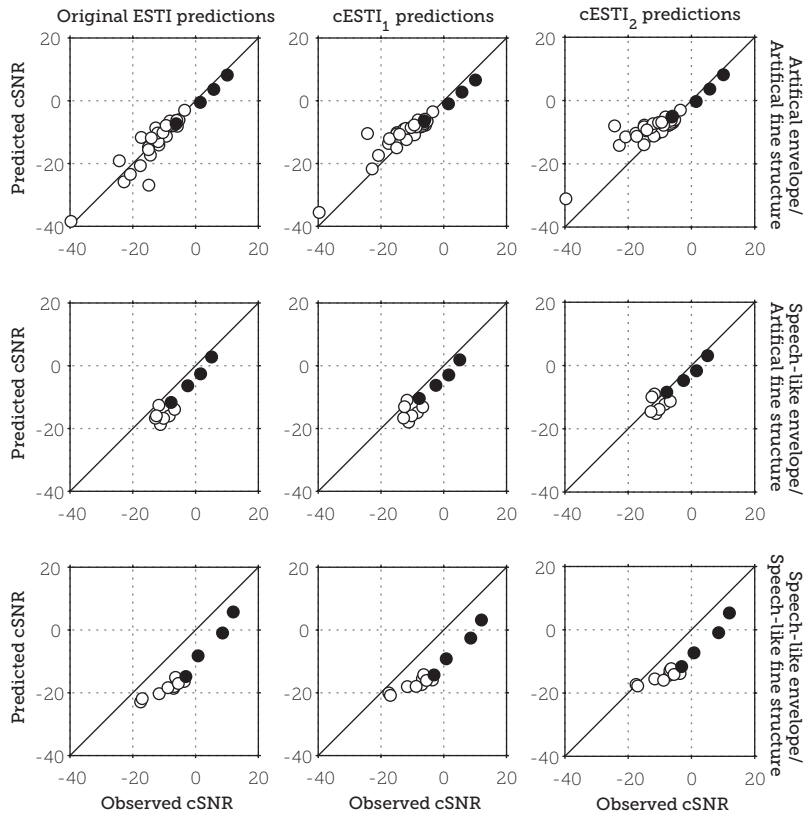


Fig. 6-5: Relation between the observed cSNR and the predicted cSNR using the ESTI, $cESTI_1$ or $cESTI_2$. Each row represents a different category of noises. Open symbols depict conditions without reverberation. Closed symbols depict conditions with reverberation. Corresponding values for R^2 *and* $RMSE$ *are found in* Table 6-3.

In Table 6-3, the corresponding values for R^2 are depicted, together with the root-mean-squared error ($RMSE$). Note that the used linear model to fit the observed and predicted data is represented by $y = x$, and not necessarily by the best linear fit $y = ax + b$. When the mean of the data is a better predictor than $y = x$, a (counterintuitive) negative value of R^2 is obtained. Since only a subset of the sentence material from chapter 3 was used, the values in the first column are Table 6-3 are different than the values found by Van Schoonhoven *et al.* (2019).

Table 6-3: Values for R^2 for three types of noise and three models. This value is based on $y = x$. The root-mean-square error ($RMSE$) is provided in parentheses. The corresponding data is depicted in Fig. 6-5 and Table E-1.

	Original ESTI	$cESTI_1$	$cESTI_2$
Artificial fluctuations / Artificial fine structure	0.88 (2.9)	0.84 (3.4)	0.68 (4.9)
Speech-like fluctuations / Artificial fine structure	0.19 (5.1)	0.36 (4.6)	0.71 (3.1)
Speech-like fluctuations / Speech-like fine structure	-0.36 (9.6)	-0.23 (9.1)	0.27 (7.1)

The $cESTI_1$ reached similar or better performance than the original ESTI for all noise types. No significant differences in prediction accuracy of male and female speakers were found. For noises with an artificial envelope and fine structure, $cESTI_2$ tends to overestimate the cSNR. This is especially true for noises where intelligibility is higher (and the cSNR is lower). A slight trend is visible where predictions of intelligibility with a cSNR lower than –10 to –5 dB deviate from the diagonal. On the other hand, for noises with speech-like modulations and artificial fine structure, $cESTI_2$ outperformed both the $cESTI_1$ and the classic ESTI. However, although the performance of both $cESTI_1$ and $cESTI_2$ is similar or better than the original ESTI, inaccuracies remain for noises with speech-like characteristics.

6.5 Discussion

The goal of the current chapter was to apply the cESTI-model to the sentence materials that were used to evaluate the original ESTI. Generally, a modest increase in prediction accuracy was observed, although inaccuracies remain for noises with speech-like envelopes.

Three steps were necessary to model sentence scores (see also Fig. 6-1 and Fig. 6-3). First, a transfer function was needed to convert the local ESTI-value to q_e. This transfer function was taken from Van Schoonhoven *et al.* (2023) and was adjusted based on the results of sentence intelligibility in SSN per study. Second, q_e was converted to Q_e using a 5-phoneme transform based on the work of Dingemanse and Goedegebure (2019). The values of the context parameters for this transform were based on the Q_e-values corresponding to the averaged cSNR in SSN across all studies used by Van Schoonhoven *et al.* (2019). Finally, Q_e was converted to a sentence score based on the context values presented by Dingemanse and Goedegebure (2019) for 6-word sentences. To calculated the context factors for the other sentence lengths, a nonlinear optimization procedure was used to fit Eqs. (6-5) and (6-6). All three steps are separately discussed below.

6.5.1 Transfer function

The transfer function between ESTI and q_e was based on the CVC-data by Van Schoonhoven *et al.* (2023). The average phoneme length of the CVC-word corpus is approximately three times longer than the average phoneme length of the sentences (223 ms versus 69 ms). Although the 5-phoneme transform aimed to adjust for the differences in context between the CVC-words and the isolated sentence words, it did not account for the difference in speech rate. A higher speech rate can lead to a decrease in intelligibility (e.g., Adams and Moore, 2009; Adams *et al.*, 2012; Saija *et al.*, 2014). Adams *et al.* (2012) found a significant difference in sentence intelligibility in noise when presenting the sentences at slow speed (120 words per minute or wpm), at average speech (170 wpm), or at high speed (234 wpm). Based on the average word length, the current sentences have a speech rate of 198 wpm and the CVC-words of 90 wpm. Also, Schlueter *et al.* (2014) found a 6-7 dB increase in cSNR when time-compressing matrix sentences to 30% of the original length. However, it is difficult to compare the effects of speech rate on sentence versus word intelligibility. Important differences are the use of context and working memory (Dingemanse and Goedegebure, 2019), and possibly forward masking.

Does the phoneme length influence intelligibility, provided that the speech quality (i.e., ESTI) is the same? Obviously, there must be a lower limit of the phoneme length at which a phoneme can still be perceived accurately. In her dissertation, Janse (2003) tested intelligibility in quiet of time-compressed meaningful and nonsense Dutch words (both mono- and disyllabic) at normal speed, and at 40% and 35% of the original duration (original word lengths and presentation levels were not provided). Intelligibility of mono- and disyllabic meaningful words decreased from 99% at normal rate to 66–69% at the fastest rate. Intelligibility of monosyllabic nonsense words decreased from 85% to 26%, and of disyllabic nonsense words from 78% to 10%.

Although the word material is different, the context factors j and k in quiet found by Bronkhorst *et al.* (1993) and Van Schoonhoven *et al.* (2022) were applied to the monosyllabic word scores. For meaningful monosyllabic words, q_e dropped from 98% to 73% between the normal and the fastest rate. The isolated phoneme score for nonsense monosyllabic words dropped from 95% to 66%. So, despite the large differences between meaningful and nonsense word scores, the isolated phoneme scores are relatively close for each speech rate. This might indicate that listeners use context similarly in normal and time-compressed speech, but reach lower word scores due to lower intelligibility of the isolated phonemes. Although time-compressed speech has different properties than natural speech, and more data is needed to investigate this more thoroughly, these results possibly indicate that a different transfer function between ESTI and q_e would be more appropriate.

6.5.2 Context factors

6.5.2.1 Transform (q_e to Q_e)

Using several constraints, the context factors for the 5-phoneme transform were found using a nonlinear optimization procedure. Although the new values were expected to be lower than the values of Dingemanse and Goedegebure (2019) for cochlear implant recipients, the drop in c_1 from 0.98 to 0.65 seems large. This discrepancy is possibly related to the inaccuracy of the transfer function. In the previous paragraph, it was suggested that the current transfer function might overestimate the intelligibility of isolated phonemes in sentences. During the optimization of the context factors, the goal was to relate the value of q_e (based on the ESTI-value at cSNR) to Q_e (based on 50% sentence intelligibility). When q_e is relatively high, the optimal context values are relatively low. Conversely, when q_e would be lower, the context factors would increase in comparison to the current values. The problem is that no reference values are available, except those of CI recipients. Both the transform and the transfer functions are unknowns, so additional research regarding these concepts is needed to resolve this issue.

Another cause of inaccuracy is the variable number of phonemes per word. As Dingemanse and Goedegebure (2019) already stated, it would be more accurate to calculate a weighted average using all transforms corresponding to the different number of phonemes per word. However, since this would result in too many parameters, the current approach using 5 phonemes was used. Note that the number of phonemes per word in the sentence material ranges from one (for *u*, the Dutch formal version of *you*) to 11 (e.g., for *scheidsrechter*, Dutch

for *referee*). The simplification of only 5-phoneme words inevitably leads to a decrease in accuracy, but it is not straightforward to predict the exact consequences. After all, the discrepancy is influenced by the combination of context values. and by the isolated phoneme score. A sufficient amount of intelligibility data would be needed for all word lengths to be able to estimate context factors with higher accuracy.

6.5.2.2 Sentences (Q_e to P_w)

Smits and Zekveld (2021) discussed a new context model and reviewed other context models. They used the same speech material as Dingemanse and Goedegebure (2019) and also used only the 6-word sentences. The context values they found were generally higher than the values used in the current study. Differences between both studies mainly concern the presentation of the sentences. Smits and Zekveld (2021) used existing intelligibility data that was obtained using all sentences in the original corpus. For the analyses, only 6-word sentences were used. On the contrary, Dingemanse and Goedegebure (2019) only presented 6-word sentences. Also, at least part of the data used by Smits and Zekveld (2021) was based on the scoring of keywords correct, instead of all words correct. Finally, a large part of the data by Smits and Zekveld (2021) was obtained using a 2-up/1-down and a 1-up/2-down scoring method leading to the SNRs corresponding to 29% and 71% respectively. This is also different from the stochastic approximation method applied by Dingemanse and Goedegebure (2019). Altogether, these differences indicate that there are risks involved when using context factors from literature, possibly due to differences in methodologies. Furthermore, the values of the context factors by Smits and Zekveld (2021) were influenced by the SNR at which the sentences were presented. The value of c_5 ranged from 0.30–0.35 at –7 dB SNR to 0.7 at –4 dB SNR. So, over the range of 3 dB, the value of this parameter doubled. Like Smits and Zekveld (2021), it is a possibility to extend our model with SNR- or q_e-dependent context factors to increase the accuracy. A complicating factor is the use of context in modulated noises, since the local SNR shows temporal variations.

6.5.3 Model predictions

Van Schoonhoven *et al.* (2019) hypothesized that the inaccuracies that were observed in the ESTI-predictions of intelligibility when using speech-like maskers were the result of modulation masking, informational masking, and/or context effects. To test the latter hypothesis, the cESTI was developed. Although an increase in accuracy was observed in the current study, model performance was still suboptimal under these conditions, especially for noises

with a speech-like fine structure like a competing speaker. The original goal of the ESTI was to improve the original STI with regard to non-stationary noises. It is possible that the limits of the STI methodology based on the MTF are reached. Possible alternatives for the MTF are the speech-based STI where speech is used as a probe signal (e.g., Payton and Braida, 2002; Payton and Shrestha, 2013). This presents the opportunity to compare the speech and noise envelopes, for instance by using linear regression (e.g., Ludvigsen *et al.*, 1990; Goldsworthy and Greenberg, 2004) or the normalized covariance (Holube and Kollmeier, 1996). However, to be able to model higher-order processes like auditory attention that are involved in informational masking are beyond the reach of signal-based intelligibility models like the STI.

6.6 Conclusion

In the current chapter, the cESTI-model was used to predict the intelligibility of sentences. For all noise types, model performance relative to the original ESTI-model remained similar or improved. However, performance was still suboptimal for noises with speech-like characteristics. The concept of the STI is based on the reduction of temporal modulations in speech. To a certain extent, the ESTI and cESTI are able to deal with temporally modulated noises, but complex interactions between the speech and noise envelopes, together with modulation masking and informational masking, remain beyond the reach of the current model design. However, resolving uncertainties about the shape of the transfer function between ESTI and q_e, and about the transform from q_e to Q_e might increase the accuracy of the model.

CHAPTER 7

GENERAL DISCUSSION

7.1 Main findings

The focus of the current thesis was to improve the Speech Transmission Index or STI (IEC60268-16, 2011) for conditions with non-stationary background noise. Based on the noise and/or reverberation of a transmission channel, the STI can take on a value between 0 and 1, providing information about the quality of the transmitted speech. This index value can be correlated with speech intelligibility by using a transfer function, making the STI a useful tool in the prediction of intelligibility. However, the classic STI performs suboptimally when background noise is non-stationary.

Chapter 2 focused on the measurement method, and it was found that a minimum impulse-to-noise ratio of +25 dB in fluctuating noise was needed for accurate STI measurements. In chapter 3, an extended version of the classic STI was introduced with the aim to predict 50% intelligibility of sentence materials in literature. The outcomes of this ESTI-mode were accurate for stationary maskers, maskers with artificial fluctuations, and maskers with real-life, non-speech-like modulations. However, maskers with a speech-like envelope introduced systematic errors in the model outcomes, probably due to a combination of modulation masking (MM), context effects, and/or informational masking (IM).

To deal with these inaccuracies, a context model was added to the ESTI-model in chapter 4. This cESTI-model was used to estimate the intelligibility of the phonemes in monosyllabic words. These isolated phoneme scores were then used as input to the context model in order to estimate word scores. It was found that model predictions for monosyllabic words improved using this method, especially for maskers with interruption rates below 5 Hz. In chapter 5, the cESTI-model was evaluated using both meaningful and nonsense monosyllabic words, resulting in an explained variance of 75% of the measured intelligibility data. When using an alternative forward masking function in the cESTI-model, the explained variance increased to 83%. In chapter 6, the cESTI-model was used to predict the intelligibility of the sentence material from chapter 3. The model accuracy improved in comparison to the original ESTI, but the performance was still relatively poor when using maskers with speech-like envelopes, probably as a result of MM and IM.

7.2 IEC standard

In the current work, the 4th edition of the IEC60268-16 (2011) was referenced mostly. However, since 2020 the 5th edition is available (IEC60268-16, 2020). In both editions, conditions with fluctuating noise are only mentioned briefly. The new standard clearly states in its scope that it does not cover conditions with fluctuating noise. In both editions, general comments regarding fluctuating noise are provided in the sections about limitations and the measurement procedure. Below, the differences between both editions are discussed.

The 5th edition states that the STI in impulsive or fluctuating background noise should be lower than 0.3 to ensure that the effect of the noise fluctuations on the probe signal is minimal. In edition 4, this threshold was stricter and set on 0.2. Also, it is advised in edition 5 that the probe signal level is 20 dB above the fluctuating background noise level, as opposed to the less strict 15 dB in edition 4. Although not explicitly stated, this presumably regards the direct measurement method. However, both editions state that the indirect method should be used when background noise is non-stationary. When using a sine-sweep to measure the STI (i.e., the indirect method) it is advised in edition 5 that the SNR per octave band is at least 20 dB. No criterion was provided in the 4th edition.

Both editions of IEC60268-16 predominantly focus on the measurement procedure with regard to fluctuating noise, and not on the fluctuating masker benefit. The differences between both editions are minor. No reference is provided for the suggested sweep-to-noise ratio of +20 dB in edition 5. In chapter 2 of the current work, an INR of +25 is suggested when performing measurements in fluctuating background noise using a sine-sweep. This corresponds to a sweep-to-noise ratio between −4 dB and +15 dB (95th percentile). The maximum sweep-to-noise ratio that was found in the results of chapter 2, was +17.5 dB. Based on these values, a minimum sweep-to-noise ratio of +20 dB is on the safe side, but is a reasonable suggestion. Overall, no changes were found in the 5th edition compared to the previous version that would influence the contents of the current thesis.

7.3 Strengths and limitations

7.3.1 Monaural presentation

During all psycho-acoustical experiments in chapters 3, 4, 5, and 6 speech and noise were presented monaurally to the listeners. This is in conjunction with the traditional STI (IEC60268-16, 2011; 2020), but is not representative for real-life listening conditions, where the head shadow effect and binaural squelch play an important role in speech intelligibility. For this reason, Van Wijngaarden and Drullman (2008) presented a binaural version of the classic STI. They used binaural cross-correlations to model interaural auditory processing, and a "better ear" principle to model the head shadow effect. Results showed that the relation between the binaural STI and word scores closely resembled the STI reference curve.

Van Wijngaarden and Drullman (2008) only calculated the interaural cross-correlograms for the octave bands centered around 500, 1000, and 2000 Hz, since binaural interactions beneficial for speech intelligibility take place between 500 and 1500 Hz (Akeroyd, 2006). The correlograms were calculated per 30 ms time window and the signal power was calculated for interaural time delays of maximally +/− 2 ms. For each time delay at these octave bands, the MTF was calculated using the direct measurement method. The MTFs that contributed to the highest overall STI were chosen. For the remaining octave frequency bands, the left and right ear MTFs were calculated and the highest MTF was chosen to base the STI on. Note that the presented material was filtered using measured and simulated binaural impulse responses. These impulse responses could also have been used to obtain the MTF using the indirect measurement method.

The time window lengths that were used for the current ESTI calculations equaled 2.8 ms, 2.0 ms, and 2.0 ms for octave band frequencies 500, 1000, and 2000 Hz respectively. This is short with respect to the 30 ms time windows the interaural cross-correlograms were based on, so it is questionable if this is a practical front end for a possible binaural ESTI. Longer ESTI time windows will result in lower prediction accuracy, especially when noise on- and offsets are abrupt. Van Wijngaarden and Drullman (2008) did mention that shorter binaural time windows will lead to less accurate estimates, but suggested that the differences were small. However, 2 ms windows seem too small, since the range of interaural time delays was also set to 2 ms.

Another option is to use the equalization cancellation method as a front end (Beutelmann and Brand, 2006; Beutelmann *et al.*, 2010). In their (short-time) Binaural Speech Intelligibility Model (BSIM and stBSIM) the SNR per frequency band was based on the difference between the signals at the left and right ear. The left ear signal was attenuated and delayed with a value that yielded the highest SNR after subtraction. The stBSIM used effective time window lengths of around 12 ms, but the authors argued that an extra time window of 100 ms was needed to simulate binaural "sluggishness" (e.g., Culling and Summerfield, 1998). They found an effective rectangular window length of around 100 ms that reflected the minimum integration time of the binaural auditory system. Thus, to obtain an accurate version of a binaural ESTI, different time constants

might have to be incorporated into the model, with the inevitable consequence of increased complexity of the calculation scheme. Given the objectives of the current work, this is undesirable.

7.3.2 Reverberation

7.3.2.1 Artificial impulse responses

In chapter 3, intelligibility measurements were done using noise and reverberation. The impulse responses that were used to reverberate the presented signals were artificial. They were constructed by multiplying exponentially decaying envelopes with fragments of white noise (George *et al.*, 2008). Since white noise has a flat spectrum, the reverberation did not cause any significant changes in the spectral content of the signals. Also, unwanted effects of room acoustics were avoided using this method. However, this choice is not representative for real-life listening conditions.

Rennies *et al.* (2014) studied the combined effect of noise and reverberation on speech intelligibility and listening effort by using the STI. They simulated reverberation by convolving the signals using white noise-based impulse responses (similar to our approach described in chapter 3) and by using more realistic, impulse responses that were simulated using room acoustics software. Real impulse responses were also used. When the MTF was only based on the SNR and on T_{60} (see Houtgast and Steeneken, 1985), intelligibility deviated for different types of reverberation at the same STI-value. Intelligibility was higher for more realistic reverberation compared to artificial reverberation. The problem was that not all properties of the impulse response were used for the STI calculations. When calculating the STI using the method by Schroeder (1981), intelligibility was more consistent with the STI-values. Clearly, the calculation of the MTF based on the normalized Fourier transform of the squared impulse response resulted in a more accurate value of the STI, also when using impulse responses based on white noise. Despite the fact that artificial impulse responses were used in the current study, the ESTI calculation using the method by Schroeder appears to be robust.

7.3.2.2 Reverberation combined with noise

In Table 3-3 and Fig. 3-7, it was shown that the classic STI- and ESTI-values were not constant across all reverberation times. Especially at T_{60} = 0.4 and 0.8 s the (E)STI was higher than the reference condition without reverberation. At reverberation times of 0.1 s and 1.2 s, the STI was similar to the reference condition. When performing a multivariate ANOVA, the ESTI of SSN and IN8 combined was significantly higher at T_{60} = 0.4 s than at 0 and 1.2 s ($p < 0.001$) and also higher than at 0.1 s ($p < 0.05$). The ESTI at 0.8 s was higher than at 1.2 s

($p < 0.01$) and at 0 s ($p < 0.05$). Note that ISTS was left out of this statistical analysis due to the contribution of informational masking.

To investigate the reason for this difference, the modulation reduction caused by reverberation and noise was assessed separately. This was done by calculating the STI in the various conditions solely based on MTF_{SNR} [according to Eq. (3-6)] or on MTF_{rev} [according to Eq. (3-5)]. It makes sense that the contribution of noise on the STI is largest at T_{60} = 0 s and smallest at T_{60} = 1.2 s. The results are shown in Fig. 7-1. The contributions of noise and reverberation on the total STI are equal at approximately 0.6 s. This is also the point where the STI deviates most from the reference value. So, at 50% intelligibility, the STI is highest when contributions of MTF_{rev} and MTF_{SNR} are equal. In other words, when noise and reverberation equally contribute to modulation reduction, this is more detrimental to intelligibility than modulation reduction due to noise or reverberation alone.

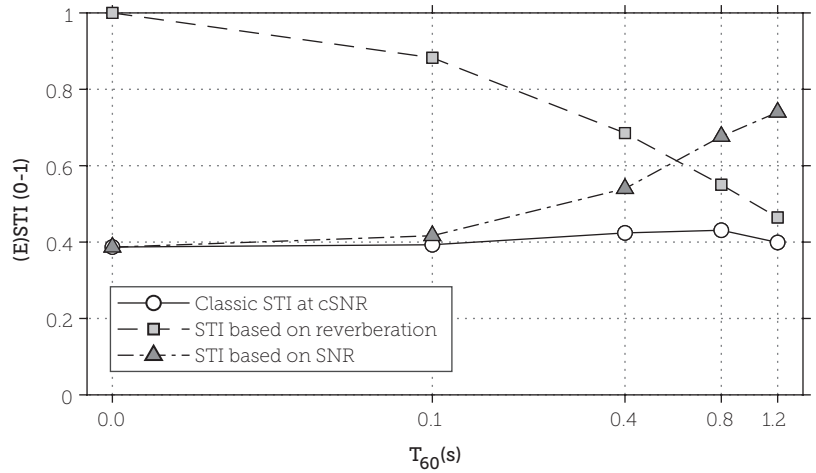


Fig. 7-1: The classic STI at cSNR (see also Fig. 3-7), together with the STI based on reverberation only, and the STI based on the SNR only. See Table 3-3 for the corresponding values of T_{60} and the cSNR.

Duquesnoy and Plomp (1980) tested sentence intelligibility for normally hearing and hearing-impaired subjects using a similar procedure as Van Schoonhoven *et al.* (2019) at reverberation times between 0 and 2.3 seconds. They found significantly higher STI-values at T = 0.4 s at 50% intelligibility in comparison to the other reverberation times. They varied the reverberation by recording the sentences in a room with variable acoustics and attributed this deviation to the

different frequency response at this reverberation time. Payton *et al.* (1994) mentioned that the STI overestimated intelligibility when speech was severely degraded by both noise and reverberation. It is possible that listeners can more easily handle one type of distortion, than two types simultaneously.

7.3.2.3 Adaptive procedure

Another aspect that deserves attention is the procedure to determine the SNR at 50% intelligibility. After an initial stepsize of 4 dB, a 2 dB stepsize was employed in the adaptive procedure. At T_{60} = 0 s, this means that the stepsize was equal to 0.067 STI units. However, when increasing the reverberation time, the contribution of the 2 dB stepsize was less dominant on the total STI. For T_{60} = 1.2 s, the stepsize was only 0.025 STI units (when *decreasing* the SNR with 2 dB) and 0.019 STI units (when *increasing* the SNR with 2 dB). This is equivalent to 0.74 and 0.58 dB respectively for the condition without reverberation. The asymmetry in the equivalent stepsize was caused by the log-transform when calculating the apparent SNR based on the MTF [see Eq. (3-8)]. This operation is the exact inverse of Eq. (3-6), where the MTF_{SNR} was calculated using the SNR. However, the MTF that was used to calculate the apparent SNR in Eq. (3-8) is the product of the MTF_{SNR} and the MTF_{rev}. Therefore, when reverberation is present, this asymmetry cannot be resolved.

So, the equivalent stepsize decreased at higher reverberation times *and* was biased towards lower SNRs. At T_{60} = 0.4 and 0.8 s, the stepsize in STI units was 10–20% larger when decreasing the SNR than when increasing the SNR. This bias made the intelligibility task more difficult and had the potential effect that the cSNR-values were higher than expected. However, although noise contributed less at higher reverberation times, one would expect this effect to increase at T_{60} = 1.2 s. However, this was not the case. An interesting option would be to adjust the stepsize based on the reverberation time, ensuring equivalent STI steps in the procedure. For the reverberation times tested here, this would result in a stepsize between 2 and 4 dB when decreasing the SNR, and a stepsize between 2 and 10 dB when increasing the SNR. Especially for T_{60} = 1.2 s, the stepsize is unrealistically large. However, a compromise might be found to increase the accuracy of the measurement when using reverberated speech.

7.3.3 Use of context

In chapters 4 and 5 a context model was added to the ESTI-model presented in chapter 3. This cESTI-model was based on the approach described by Bronkhorst *et al.* (1993) regarding interrupted speech. Their fundamental assumption was the existence of a fixed relation between the speech-time-fraction and

the intelligibility of isolated phonemes, expressed in the transfer function in Fig. 4-4. The cESTI-model took this approach one step further. It was based on the assumption that there exists a fixed relation between the ESTI-value per phoneme and the intelligibility of the same phoneme in isolation (q_e). As a consequence, the cESTI does not distinguish between speech information in a gap between noise bursts, and speech information as a result of masking noise. This was already discussed in section 4.5.

The data for low interruption frequencies presented in chapter 5 can be viewed in a different way (see Fig. 7-2). Here, the x-axis displays the fraction of the word that was completely audible to the listener. The axis is scaled to the lengths of the phonemes. So ⅓ and ⅔ represent the transition between the different phonemes. The value 0 means that the complete word was masked at either –18 or –9 dB SNR. The value ½ means that half of the word was masked and the other half was completely audible. The value 1 means that the entire word was completely audible. It can be viewed as a 'noise-curtain' that is opened step by step. Note that the graphs are only displayed from left to right, but data was used where the 'noise-curtain' was opened from either side. Therefore, the value ⅓ at the x-axis represents the data where the initial consonant of the word was audible and the final two phonemes were masked, combined with the data where the final consonant was audible and the first two phonemes were masked. A limitation of this approach is the presence of coarticulatory cues around the phoneme transitions at x = ⅓ and $x = ⅔$.

Based on Fig. 7-2 several observations can be made. At an SNR of –18 dB (top panels), no speech information was available during the noise peaks. This is illustrated by the word score of 0% when only (a part of) one consonant was audible. Note that context did not play a role in this situation. When the vowel became audible (at $x \geq ⅓$), a difference between meaningful and nonsense words is observed. When half of the vowel was audible (at $x = ½$), the meaningful word score increased to 39% (top-left panel), whereas the nonsense word score remained below 5% (top-right panel). This is most likely the result of a difference in context.

At an SNR of –9 dB (bottom two panels), differences between meaningful and nonsense words were larger. From the bottom-left panel, it is clear that the word score was around 18% when the complete word was masked by noise. Note that this value is slightly higher than the stationary noise data at –12 dB displayed in Fig. 5-2. However, when only a part of one consonant was completely audible, the intelligibility of the entire word increased to 70%. It is possible that this point is an outlier, but the meaningful word score was in any case larger than the nonsense word score when the listener had access to (part of) the first two phonemes (at $x \leq ⅔$). So, when the word was completely masked

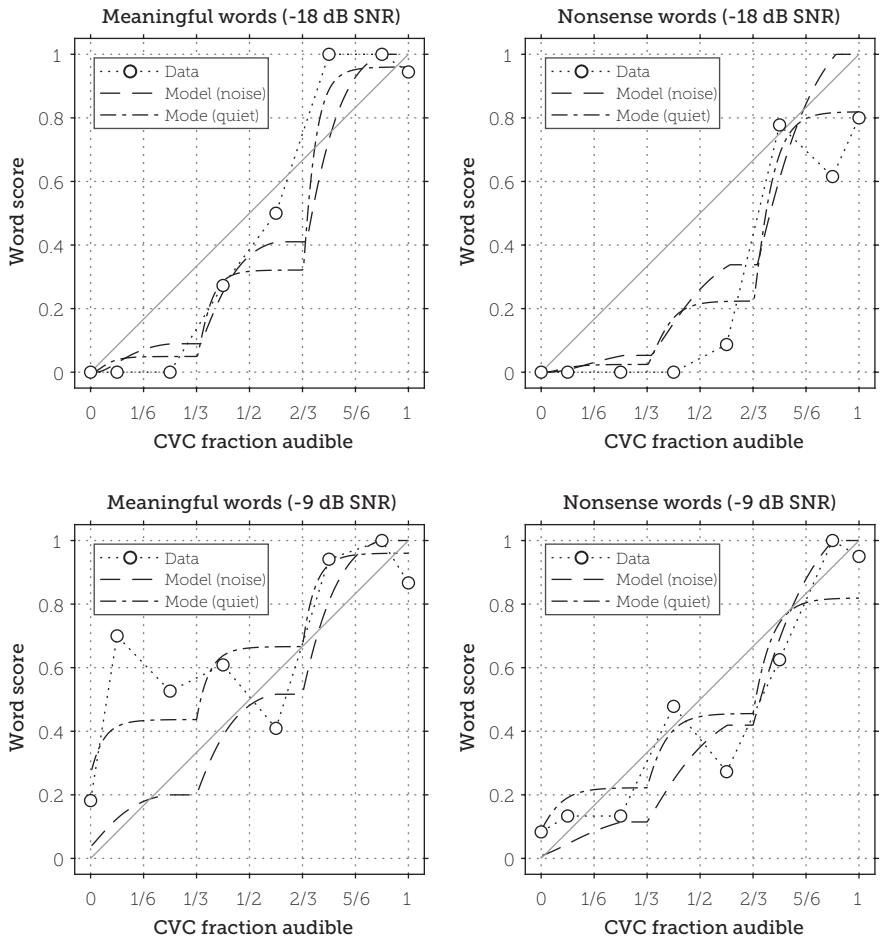


Fig. 7-2: Word scores at low interruption frequencies. The x-axis displays the fraction of the CVC-word that was not masked by noise. The x-axis is scaled to the length of the phonemes, so ⅓ and ⅔ represent the transitions between the different phonemes. Two model predictions are displayed per panel: using the transfer function and context factors for quiet, and for noise. Only predictions using the Boothroyd and Nittrouer model are displayed.

by noise at an SNR of –12 dB, intelligibility was between 0 and 20%. However, when part of a consonant became fully audible, the listener was capable of combining this extra information with the context of the word, which caused a clear increase in word score.

Model predictions using context factors in noise and in quiet are not very different in three out of the four conditions. However, when looking at the bottom-left panel, the model appears to better predict the data when context factors in quiet are used. This is partly a consequence of the difference in context factors between quiet and noise (see Table 4-2), but also of the difference in transfer functions (compare Fig. 4-4 and Fig. 5-4). Note that the explained variance is highest for the context factors in quiet (see Table 7-1). Here, predictions using the Bronkhorst model are also displayed.

Table 7-1: R^2-values for the model predictions using the context models by Boothroyd and Nittrouer, and by Bronkhorst. Both context factors in quiet and in noise were used. Also, the explained variance is displayed for meaningful and nonsense words separately and combined. Note that in Fig. 7-2 only predictions using the model by Boothroyd and Nittrouer are displayed.

	Boothroyd and Nittrouer model		Bronkhorst model	
	Noise	Quiet	Noise	Quiet
Total	0.77	0.90	0.78	0.83
Meaningful	0.70	0.90	0.68	0.86
Nonsense	0.81	0.90	0.84	0.79

This alternative approach confirms the statement from chapters 4 and 5 that context factors based on intelligibility in quiet are more suitable for low interruption rates than context factors based on intelligibility in noise.

7.3.4 Short time windows

In chapter 3 the choice was made to comply with Van Wijngaarden and Houtgast (2004) and use modulation bands between 0.63 and 31.5 Hz. The periods corresponding to the different modulation frequencies range from 31 ms to over 1.5 sec. In the ESTI-model, local STI-values were calculated for time windows as short as 2 ms, depending on the octave frequency band (see Table 3-1). These local values were then averaged, which resulted in one ESTI-value. A similar approach was used for the ESII (Rhebergen *et al.*, 2006), where audibility was assessed per time window. So, if speech is audible 50% of the time, a total of 50% of the speech information is available, which is intuitively correct. However, the ESTI uses modulations in the speech as a measure of the quality, and not the audibility. So, what is the significance of the time window lengths with respect to these speech modulations?

When speech is masked by interrupted noise at a low interruption rate, most speech modulations can be assessed during one glimpse. For example, at a rate

of 0.5 Hz, the glimpses are 1 second long, so all but the lowest two modulation frequency bands can be analyzed accurately between two noise blocks. Here, the parallel with the ESII is clear: if 50% of the speech is available, a total of 50% of the speech *modulations* is available. On the contrary, when glimpsing speech in interrupted noise (and also when glimpsing interrupted speech) at higher rates, the listener forms a complete picture of the speech by combining several glimpses (e.g., Miller and Licklider, 1950; Cooke, 2006). At these rates, modulations are not assessed between two noise blocks but are followed across the different glimpses, like in the visual analogy of the picket fence (Miller and Licklider, 1950). This listening across the glimpses is not accounted for in the ESTI-model. After all, averaging the local STI-values for a 16 Hz interrupted noise with 50% DC yields the same result as for a 2 Hz interrupted noise (ignoring forward masking).

For a better analysis of the choice of time windows, a closer look at the SII and STI is needed. When comparing the classic SII and STI, there are more similarities than differences. Even though the SII is audibility-based and the STI is based on modulations, and the SII does not account for reverberation. For simplicity, the current assumption is that speech and noise are spectrally matched, meaning that the SNR in each frequency band is the same. Also, a pure exponential impulse response is assumed, which means that reverberation is the same across frequency bands. This means that the apparent SNR only depends on the modulation frequency and the $\boldsymbol{MTI}$ is constant over all frequency bands. When the $\boldsymbol{MTI}$ does not depend on the octave frequency bands, this means that frequency weighting and redundancy correction can be discarded:

$$\begin{aligned} STI &= \sum_{q=1}^{Q} \alpha_q MTI_q - \sum_{q=1}^{Q-1} \beta_q \sqrt{MTI_q MTI_{q+1}} \\ &= \sum_{q=1}^{Q} \alpha_q MTI - \sum_{q=1}^{Q-1} \beta_q \sqrt{MTI^2} \\ &= MTI \left[\sum_{q=1}^{Q} \alpha_q - \sum_{q=1}^{Q-1} \beta_q \right] \stackrel{\text{def}}{=} MTI = \frac{1}{R} \sum_{r=1}^{R} \frac{SNR_{app_r} + 15}{30} \end{aligned} \tag{7-1}$$

Now, for the SII we can assume that speech does not exceed normal vocal effort by more than 10 dB. This means that L_q (Level Distortion Function) is equal to one. Since the band importance functions add up to one, the following relations hold:

$$\begin{aligned} SII &= \sum_{q=1}^{Q} L_q I_q \frac{SNR_q + 15}{30} = \frac{SNR + 15}{30} \sum_{q=1}^{Q} L_q I_q \\ &= \frac{SNR + 15}{30} \sum_{q=1}^{Q} I_q \stackrel{\text{def}}{=} \frac{SNR + 15}{30} \end{aligned} \tag{7-2}$$

with $SNR_q = E'_q - D_q$ and $L_q = 1 - \frac{E'_q - U_q - 10}{160}$

This results in the following relations for the STI and SII:

$$STI = \frac{1}{R} \sum_{r=1}^{R} \frac{SNR_{app_r} + 15}{30} \tag{7-3}$$

$$SII = \frac{SNR + 15}{30} \tag{7-4}$$

Without reverberation, the apparent SNR is not dependent on modulation frequency ($\boldsymbol{r}$), and thus, when speech and noise are spectrally matched and at normal vocal effort, the SII and STI are identical. Despite the fact that the SII was originally based on audibility and the STI on modulation reduction, the practical implementation converged over the years. Even though the choice of short time windows is theoretically not the most sound option for the ESTI, in practice lessons were learned from the success of the ESII. Because of the similarities, similar choices were made for the ESTI. Another goal of the current extension was to make the ESTI relatively easy to implement. When assessing interactions between speech and noise modulations more thoroughly, it is an option to implement a modulation filterbank (e.g., Jørgensen and Dau, 2011) or a speech-based version of the STI (e.g., Payton and Shrestha, 2013). However, the calculation scheme of the ESTI compared to the STI would have been drastically changed, which was not the objective of this research. An important advantage of the ESTI is that the calculation per time window exactly follows the calculation scheme of the classic STI.

7.3.5 Hearing threshold

At regular speech levels, the hearing threshold does not play a role in the classic STI. After all, in this situation modulation reduction is a supra-threshold phenomenon. However, for lower speech levels the hearing threshold becomes increasingly important. It is modelled as an internal noise (Pavlovic, 1987) and is defined in (IEC60268-16, 2011; 2020). Since the ESTI was designed to deal with non-stationary maskers, the hearing threshold plays a more prominent

role than in the classic STI. It serves as a virtual noise floor during the gaps in the noise, and as the lower limit of the exponentially decaying forward masking function [see Eq. (3-3)]. For the latter purpose, the estimate of the minimal audible pressure (MAP) by Killion (1978) was used.

To analyze the effect of the auditory threshold on the STI calculation, a point of reference is needed. It is important to define where the sound level is measured, for instance in a 6cc coupler, or at the position of the head in the free field (with the head absent). The exact point of reference is not of particular importance, as long as the same point is used throughout the calculations. In this paragraph, the sound pressure level at the eardrum is used for this purpose.

When presenting speech in stationary noise to a listener via headphones, the SNR at the eardrum is pretty much equal to the SNR originally presented. After all, the linear transformation as a result of the headphones and outer ear equally influences speech and noise. When presenting speech in quiet (or in interrupted noise, during the silent intervals), the situation becomes more complex, especially when speech is soft. The hearing threshold remains fixed, but the speech level is influenced by the transducer and the characteristics of the ear canal (Bentler and Pavlovic, 1989).

So, what would the effect of these adjustments be on the ESTI-values? To check this, the ESTI was calculated for SNRs between –60 and +30 dB with the noise level fixed at 65 dB (A). This was done for stationary SSN and 8 Hz interrupted noise. The speech and noise were filtered using a 2000th order FIR filter to account for the frequency response of the TDH39 headphones that were used in the majority of the experiments. Furthermore, the MAP by Killion (1978) was used as the internal noise instead of the values in Table A.2 of IEC60268-16 (2011). Finally, the output of the headphones was transformed to the sound pressure level at the eardrum using the transfer functions provided by Bentler and Pavlovic (1989) in column E of their Table 1. These steps were done *before* the calculation of the forward masking. Note that the transformation values for 125 and 8000 Hz were not available, but were visually extrapolated based on the existing data.

As expected, there were slight differences between both calculation schemes. The maximum differences using the SSN- and IN8-masker were less than 0.01 and 0.015 STI units respectively. These differences fall well within the accepted measurement error of 0.03 STI units (IEC60268-16, 2011). The larger error in interrupted noise was expected due to the largest effect of the hearing threshold. When zooming in on SSN, the *classic* STI was not influenced by the change in hearing threshold. This was expected. However, when calculating the ESTI with the alternative hearing threshold, but without forward masking, the same differences were observed. In other words, the observed differences in SSN due

to the alternative hearing threshold were caused by the temporal approach of the ESTI and not by the introduction of forward masking. This effect was caused by intrinsic modulations in stationary noise. For example, when calculating the sound level per time window in the 125 Hz octave band for SSN at 65 dB (A), the levels per window range between 40 and 80 dB. Since the auditory threshold in this band is 46 dB according to IEC60268-16 (2011), the calculations for this band were affected by the auditory threshold. In conclusion, although choosing one point of reference for all calculations would be a more thorough approach, the effects are relatively small and probably will not affect the main conclusions.

7.4 ESTI versus other models

7.4.1 Classic STI

When suggesting an alternative for a widely used, proven metric, it is important to focus on the changes in practice. After all, it is not desirable to change outcomes that were not inaccurate in the first place. In the current situation, this means that results for stationary noise with or without reverberation should ideally remain the same. Payton and Shrestha (2013) stated that any short-time version of an established metric such as the STI should approach the long-term results when averaged over many frames, unless the long-term STI is inaccurate. Van Wijngaarden and Drullman (2008) further suggested that STI parameters should not be tuned to a specific application and that added complexity of a modification should be proportional to the increase in accuracy.

When assuming a situation without background noise, the total MTF only depends on the reverberation and does not depend on time. The MTF-value for each time window will be exactly equal. As a result, the eventual ESTI-value will be the same as the classic STI-value, which is desirable and in line with the earlier statements.

For the effects of stationary noise, the classic STI and the ESTI should ideally retrieve the same results. In Fig. 7-3 the behavior of the STI and ESTI as a function of SNR in SSN is shown. Although both metrics follow the same general pattern, there are two important differences. First, in the linear portion of the relation between SNR and (E)STI, there is a systematic deviation of 0.024 STI units on average. The second difference is most clear at the boundaries of the linear portion around +/– 15 dB. Where the original STI shows discontinuities in these regions due to the clipping of the apparent SNR, the ESTI changes more gradually.

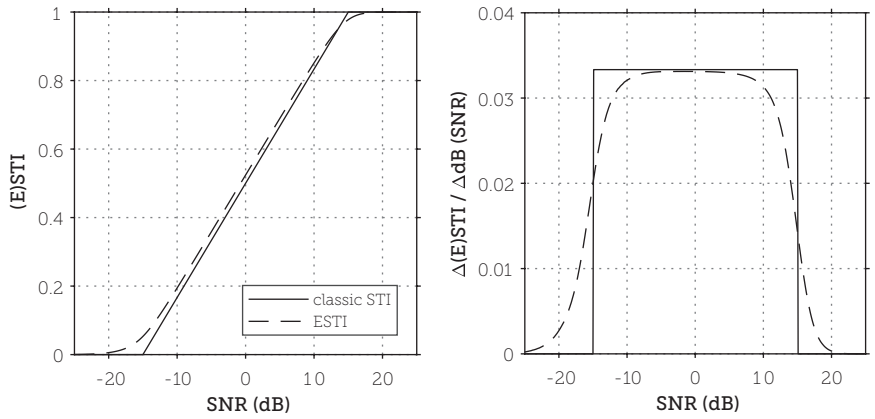


Fig. 7-3: Relation between SNR and (E)STI to illustrate the differences in behavior. In the right panel, the derivatives with respect to the SNR of the relations in the left panel are shown.

The first, systematic discrepancy is related to the fundamental difference between calculating the classic STI and the ESTI. With the classic STI (in case of noise, but without reverberation), the rms-value of the total signal is calculated, after which it is transformed to the STI. Basically, the SNR between –15 and +15 dB is linearly mapped to an index between 0 and 1. With the ESTI on the other hand, the rms per time window is calculated in order to calculate the local STI-value first. The same linear transformation thus takes place, but now per time window instead of per the complete signal. As opposed to the local rms-values, the local STI-values are *not* distributed normally and the skewness of the distribution causes the discrepancy between the ESTI and classic STI. The longer the time windows, the smaller this effect is, since peaks and valleys in the noise are averaged out. Also, the discrepancy depends on the type of stationary noise that is used, but is generally between 0.02 and 0.03 (see Table 7-2), which is equivalent to a difference in SNR less than 1 dB.

Another cause of the discrepancy between the STI and ESTI for SSN is the following. Given a long-term SNR of –15 dB, the classic STI would be 0 (disregarding octave weighting and redundancy correction). However, as mentioned earlier, SSN is not purely stationary and shows random temporal fluctuations (Stone *et al.*, 2011). As a result, the SNRs per time window are not all equal to –15 dB, but form a distribution around –15 dB. The local SNRs that fall below –15 dB are clipped to –15 dB, but the higher SNRs yield a local STI-value larger than 0. As a consequence, when averaging all local STI-values, the total ESTI-value will be slightly larger than 0. The opposite is also true: if the long-term SNR equals +15 dB, the ESTI will be slightly smaller than 1.

Table 7-2: The differences in STI and ESTI for different types of stationary noise (T_w = 2 ms).

Noise	Classic STI	ESTI
CVC meaningful	0.500	0.524
CVC nonsense	0.500	0.529
VU female	0.500	0.524
VU male	0.500	0.525
White noise	0.500	0.521

As is visible in Fig. 7-3 this also has an effect at SNRs above –15 dB and below +15 dB. Since the range of levels after windowing with the current choice of time windows can be as large as 30 dB for the lower octave band frequencies, this effect is also noticeable at SNRs close to 0 dB. Increasing the time window length would reduce this effect, but in chapter 3 it was shown that this reduces model accuracy.

Clearly, discrepancies between the STI and ESTI are not desirable. However, it is the question if these effects are entirely unwanted. The classic transfer function between SNR and STI contains discontinuities around +/– 15 dB due to the 'hard' clipping of the index. Between –15 and +15 dB, the STI increases by approximately 0.033 per dB SNR. Outside of this range, the SNR does not affect the STI-value. The ESTI, without discontinuities around the edges, appears more natural. After all, the cumulative amplitude distribution of speech is shaped like a regular performance intensity function (Drullman, 1995; Boothroyd, 2008; Rhebergen *et al.*, 2009). The amount of speech information that becomes available per dB increase in SNR is not distributed equally. So, a smooth transfer function between the SNR and a measure for speech quality is not necessarily a disadvantage. Besides this, small modulations in stationary noise are traditionally disregarded, but can certainly affect speech intelligibility (e.g., Stone *et al.*, 2012).

As opposed to the ESTI, the addition of context should not be viewed as a generalization of the STI. The first two chapters of this thesis aimed at extending the usability of the STI to conditions where noise is non-stationary. The end result of the calculations was still a single ESTI-value. This single value was used to predict speech intelligibility, but could also serve to quantify the acoustical and/or noise-related properties of a transmission channel. This is different when context is added, since the intelligibility per phoneme is then assessed based on the local STI-value. The transfer function that is used to relate the STI-value to intelligibility is now part of an intermediate step, and not of the final step. It must therefore be viewed as an addition to the existing measure, since the specific application is only the prediction of intelligibility. It cannot be used as a complete replacement of the STI or the ESTI.

7.4.2 Other models

Intelligibility models can be categorized in several ways. Feng and Chen (2022) distinguished between intrusive and nonintrusive models, based on the presence or absence of a reference signal respectively. The STOI (Taal *et al.*, 2011) is a typical intrusive model, since short-time temporal envelope segments of clean and degraded speech are compared by using a covariance metric. Nonintrusive models are usually less accurate, but are typically used when no reference signal is available, like when real-time monitoring of intelligibility is desirable. For example, in state-of-the-art hearing devices nonintrusive models can be used for the real-time adaptation of the device settings to optimize intelligibility when acoustic circumstances change (e.g., Falk *et al.*, 2015). Since real-time monitoring of intelligibility was not a goal of the (E)STI, this type of model is beyond the scope of the current research.

Furthermore, the distinguishment between conventional feature-based (using acoustical features), non-conventional data-based, and neurophysiological models can be made (Feng and Chen, 2022; Karbasi and Kolossa, 2022). This last category uses features derived from signals like EEG (Verschueren *et al.*, 2020) or oculometry (Favre-Félix *et al.*, 2018) to retrieve additional information about speech intelligibility. Since these models — or additions to existing models — do not use speech as the primary source of information, these are also beyond the scope of the current work.

Due to the large number of models that exist to predict speech intelligibility, it is virtually impossible to create an overview that is complete. In the sections below the focus lies on conventional feature-based models which are based on modulation reduction and audibility. Besides this, ASR-based models are discussed briefly since the number of applications and their importance rapidly increase.

7.4.2.1 Feature-based models based on audibility

The foundation for modelling speech intelligibility was laid at Bell Telephone Laboratories, New York in the form of the Articulation Index or AI (French and Steinberg, 1947; Fletcher and Galt, 1950). It is a speech audibility measure that later evolved into the Speech Intelligibility Index or SII (ANSI-S3.5, 1997). Several models exist that are based on the SII. Rhebergen and Versfeld (2005) and Rhebergen *et al.* (2006) introduced a temporal approach by calculating the SII using 4 ms windows to better deal with non-stationary maskers. This ESII-model formed the basis of the current ESTI method. Meyer and Brand (2013) proposed the ESIIsen, where speech was used as the test signal instead of stationary noise. Results were comparable to those of the original ESII.

Beutelmann and Brand (2006) also used the SII, but combined this with an Equalization Cancellation (EC) front-end to model binaural hearing. This model later evolved into the Binaural Speech Intelligibility Model (BSIM). Its predictive power was good for stationary noise and reverberant conditions, but the effect of reverberation on the speech itself was not taken into account since only near-field speech was used. Beutelmann *et al.* (2010) successfully introduced a short-time implementation of their model (stBSIM), similar to the ESII. They used an effective frame length of 12 ms, similar to the frame length by Rhebergen and Versfeld (2005) where the ESII was introduced without forward masking. Note that the introduction of forward masking to the ESII required shorter frame lengths of 4 ms (Rhebergen *et al.*, 2006). Beutelmann *et al.* (2010) observed a performance improvement compared to the original EC/SII-model, but still only used near-field speech. Rennies *et al.* (2011) modelled several extensions to the BSIM to deal with the detrimental effect of reverberation on intelligibility. They found better model predictions than the original BSIM when the effect of reverberation was strong. The stBSIM was not investigated in this study.

Lavandier and Culling (2010) chose a similar approach as Beutelmann and Brand (2006) but replaced the EC frontend with a BMLD (Binaural Masking Level Difference) estimation step. A high correlation was found between model predictions and speech intelligibility measurements in reverberation and stationary noise. Collin and Lavandier (2013) tested the same model but also for speech-modulated interferers. They based their effective window lengths of 12 ms on Beutelmann *et al.* (2010) and found a similar performance to their stBSIM.

7.4.2.2 Feature-based models based on modulations

Various alternative measures for the classic STI were proposed over the years. Notable examples are the speech-based STI (Ludvigsen *et al.*, 1993; Drullman, 1995; Hohmann and Kollmeier, 1995; Payton and Braida, 1999; 2002), the Covariance-based STI or CSTI (Ludvigsen *et al.*, 1990; Holube and Kollmeier, 1996; Goldsworthy and Greenberg, 2004), the quasi-stationary STI or QSTI (Schwerin and Paliwal, 2014) and the eSTI (Prodi and Visentin, 2019).

Payton *et al.* (1994) demonstrated that the intelligibility of clear speech is better than that of conversational speech, even though the long-term spectra are similar. Since the STI is not sensitive to differences in speaking style, Payton and Braida (1999) chose to use speech instead of modulated noise as a probe signal. However, the problem arose that the degradation of speech using noise caused spurious modulation components, leading to artifacts. To account for these artifacts, MTFs were truncated based on the coherence between the clean and degraded speech envelopes. A good correspondence between the

speech-based and classic STI was found. Goldsworthy and Greenberg (2004) discussed a variety of existing and novel speech-based STI approaches to deal with nonlinearly processed speech. Two notable approaches were an alternative implementation of the Envelope Regression or ER method by Ludvigsen *et al.* (1990) and the Normalized Correlation method, which was derived from the Normalized Covariance Method or NCM (Holube and Kollmeier, 1996). When using ER, the envelope of the probe and response signals are compared by means of linear regression. In NCM the normalized covariance between the envelopes of the probe and response signals are calculated. In both approaches, the output is used as a surrogate for the MTF when calculating the apparent SNR (see Eq. (*3-8*), where ***MTF*** would be substituted for the alternative metric).

Payton and Shrestha (2013) took it one step further and designed a short-time speech-based version of the STI based on the ER method suggested by Goldsworthy and Greenberg (2004). Speech was degraded using SSN, SSN combined with reverberation, and babble noise. Various time window lengths were used, ranging between 78 ms and 107 s (the length of the entire probe signal). The optimal value was 300 ms, since for shorter windows the ER algorithm followed the noise envelope during the pauses in the speech, leading to non-zero values. Evaluation of speech intelligibility was only done using SSN and showed a good correlation with the STI. Other short-time speech-based versions were developed by Schlesinger (2012) using 400 ms windows and Falk *et al.* (2010) using 256 ms windows.

The Short-Time Objective Intelligibility measure (STOI) was developed by Taal *et al.* (2011) and updated to the ESTOI by Jensen and Taal (2016). It is, like the CSTI (Goldsworthy and Greenberg, 2004), based on a covariance metric between clean and distorted speech. In contrast to the CSTI, the STOI calculates the correlation coefficient for short time segments. It was tested with window lengths varying between 20 – 30 ms and more than one second. The final model works with overlapping windows of 384 ms. The ESTOI uses the same window lengths, but calculates the correlation coefficients in the spectral domain instead of in the temporal domain. The authors claimed that the effects of time-modulated maskers were better captured in this way. According to Karbasi and Kolossa (2022), the STOI has become a widely used benchmark in the field of speech processing.

Another concept that is closely related to the STI was introduced by Dubbelboer and Houtgast (2008). They suggested the signal-to-noise ratio in the modulation domain (SNR_{mod}) to better deal with nonlinear distortions such as spectral subtraction. The SNR_{mod} is the ratio between the speech modulations and the modulation noise floor. This modulation noise floor was said to consist of the original noise modulations, but also contains modulations as a result of the

nonlinear interactions between speech and noise. The SNR_{mod} better dealt with these nonlinear interactions than the STI. A downside to this method was that the modulation noise floor could not be estimated from the clean speech or the noise separately, but had to be inferred from the noisy speech using an artificial probe signal. This approach made it hard to generalize the concept of the SNR_{mod}.

A similar approach was the Envelope Power Spectrum Model or EPSM (Dau *et al.*, 1999; Ewert and Dau, 2000) and used a modulation filter bank. At first, it was used to predict amplitude modulation detection based on the signal-to-noise ratio of the envelope power (SNR_{env}) at the output of the modulation filter. This model only considered target and noise modulations, not interaction modulations like the SNR_{mod}. Jørgensen and Dau (2011) introduced the speech-based EPSM or sEPSM to evaluate speech intelligibility. The results of the model were in accordance with speech intelligibility data for reverberated noisy speech (SSN) and spectral subtraction. Especially in the latter, nonlinear condition, the sEPSM outperformed the STI.

The multi-resolution sEPSM or mr-sEPSM (Jørgensen *et al.*, 2013) aimed to also include non-stationary maskers by estimating the SNR_{env} for each combination of modulation and peripheral filters in temporal segments. The lengths of these temporal segments were based on the modulation filter and ranged between 3.9 and 1000 ms. The mr-EPSM was again successfully applied to conditions with stationary interferers and reverberation, and spectral subtraction. It also accounted for fluctuating interferers with gradual modulations, like sinusoidally amplitude modulated noise, two interfering talkers, and ISTS. Other extensions of the EPSM were introduced to deal with phase distortions (Chabot-Leclerc *et al.*, 2014), the addition of a 'traditional' PSM branch with a decision backend to deal with more types of masking (Biberger and Ewert, 2016; 2017) and a binaural extension (Chabot-Leclerc *et al.*, 2016).

The spectro-temporal modulation index or STMI by Chi *et al.* (1999) is a biologically motivated model. It compares the cortical representations of the spectro-temporal modulations in clean and distorted speech. Elhilali *et al.* (2003) showed that it performed similarly to the classic STI for reverberated speech in stationary noise, but performed better for nonlinear distortions like phase jitter or linear phase-shifting. The main reason is that the speech envelope remains intact after these distortions, which explains the insensitivity of the STI. The STMI was updated to the OSTMI by Edraki *et al.* (2019) and the wSTMI by Edraki *et al.* (2021a) to enhance performance. Edraki *et al.* (2021b) and Edraki *et al.* (2022) also introduced a spectro-temporal glimpsing index (STGI) to better deal with modulated noises, noise reduction, and reverberation. The STGI outperformed other metrics like the ESII, ESTOI, and wSTMI on a variety

of distortions including reverberation, modulated noises, and competing speakers. However, the authors state that the STGI-model has 44 parameters fitted to the training dataset, which makes the model prone to overfitting.

7.4.2.3 Data-based models

Data-based models use machine learning techniques that are applied in automated speech recognition (ASR). These methods are used to either predict speech intelligibility directly, or to use the output (or internals) of an ASR-system for the same purpose. These models show great potential for a wide range of acoustic scenarios (e.g., Andersen *et al.*, 2018; Zezario *et al.*, 2020), provided that large training sets are available and extensive work is done on parameter optimization. The direct consequence is that these models generally have high complexity and low explainability, which is the main disadvantage compared to conventional feature-based models.

The Framework for Auditory Discrimination Experiments or FADE (Schädler *et al.*, 2015; Schädler *et al.*, 2016) is a model that uses the accuracy of the ASR transcription to estimate speech intelligibility. It used Mel Frequency Cepstral Coefficients (and their temporal derivatives) as a front-end, and whole-word Hidden Markov Models as a backend. It performed well on sentence tests and has been extended to model binaural hearing (Kollmeier *et al.*, 2016) and impaired hearing (Schädler *et al.*, 2018). However, since a FADE simulation for one condition required several hours of signals, a data-reduced version was developed (Hülsmeier *et al.*, 2021). This version reduced the amount of time to 30 minutes of recorded speech per condition and performed within 1 dB of FADE for speech in stationary noise. However, the difference in performance was approximately 5 dB for non-stationary noises. Spille *et al.* (2018) introduced a model based on machine learning with a front-end using features from an amplitude modulation filterbank and a deep neural net (DNN) based backend. This model outperformed the ESII, STOI, and mr-EPSM on various modulated maskers. The advantage compared to the FADE-model was that the DNN-based model was not trained using the same type of data as in the testing phase. However, this came at the cost of needing more training data and more computational resources.

Various other ASR-based models are available, but a complete discussion is beyond the scope of the current work. See Karbasi and Kolossa (2022) for a recent and extensive review. Speech intelligibility prediction using deep neural networks is rapidly developing and super-human performance in recognizing speech seems to be within reach (Nguyen *et al.*, 2020). It is therefore likely that ASR-based speech intelligibility prediction will further improve during the next decade and that conditions with fluctuating noises, binaural hearing, sensorineural hearing loss, and hearing aids will be modelled more accurately.

7.4.3 Model comparisons

When comparing models, a quantitative comparison is desirable. For this purpose, two existing models were selected from the literature: the ESTOI (Jensen and Taal, 2016) and the mr-sEPSM (Jørgensen *et al.*, 2013). This choice was partly based on the applicability to non-stationary noises, but also — pragmatically — on the online availability of the MATLAB-code.

Using these models, the cSNRs of the speech material in chapter 6 were estimated. Again, the metric was calibrated using the cSNR in SSN per study. Since the ESTOI is expected to have a monotonic increasing relation with intelligibility, the same approach as for the ESTI and cESTI was used. At cSNR the listener is expected to reach the same ESTOI-value in SSN as in non-stationary noises. Clean and distorted speech were used as the input signals.

For the mr-sEPSM the fitting procedure was more complex. Distorted speech and noise alone were used as input signals and were used to calculate the speech-to-noise envelope power ratio or SNR_{env}. To convert SNR_{env} to d' (the sensitivity index of an "ideal observer"), the parameter k was fitted according to Eq. (6) from Jørgensen *et al.* (2013) based on the cSNR in SSN per study[x]. To convert d' to an intelligibility score, several parameters that were suggested by Jørgensen and Dau (2011) based on Danish meaningful sentences were used in the current estimations. Especially the parameter σ_s, which was assumed to be primarily related to the redundancy of the speech material, might need extra tuning. For example, as opposed to the Dutch sentence material (Versfeld *et al.*, 2000), the level of the sentences of the Danish speech corpus was adjusted to equalize intelligibility (Nielsen and Dau, 2011). Finally, the fitted parameter k was applied to SNR_{env} for all non-stationary noise conditions from the same study. Since the ESTOI and the mr-EPSM use real speech as input signals, 100 sentences were selected from the total speech corpus for computational reasons.

The root-mean-square error ($RMSE$) between observed and predicted cSNRs are depicted in Table 7-3. The ESTOI had difficulties with reverberation, especially when T_{60} was 0.8 s or higher. Therefore, the $RMSE$ is depicted for all conditions combined, and for the conditions without reverberation (within parentheses). Since the spread of the data was relatively low, especially when the conditions with reverberation were omitted, the $RMSE$ was used instead of R^2.

x This parameter k should not be confused with the context factor from the Boothroyd and Nittrouer model.

Table 7-3: *RMSE*-values of model prediction with reverberation. In parentheses are the values when data with reverberation was removed. No conditions with reverberation were tested in the category in the bottom row, so no values between parentheses are provided.

	ESTI	$cESTI_1$	$cESTI_2$	ESTOI	mr-sEPSM
Artificial fluctuations / Artificial fine structure	2.9 (3.0)	3.4 (3.5)	4.9 (5.1)	5.5 (5.0)	11.5 (12.2)
Speech-like fluctuations / Artificial fine structure	5.1 (5.9)	4.6 (5.0)	3.1 (3.4)	4.8 (2.8)	4.1 (4.7)
Speech-like fluctuations / Speech-like fine structure	9.6 (9.6)	9.1 (8.5)	7.1 (6.4)	8.6 (3.1)	5.2 (2.4)
Real-life background noise	4.8	4.8	3.4	10.9	8.9

For noises with artificial fluctuations and fine structure, the ESTI performed best. Especially the mr-sEPSM had difficulties with abrupt fluctuations in the noise. As was already clear from chapter 3, the ESTI had problems when fluctuations are speech-like. Although the cESTI is a clear improvement, especially the ESTOI performs better in the absence of reverberation. Note that all models perform relatively poor for competing speakers and ISTS (third row) in the presence of reverberation. Taal *et al.* (2011) mention that the original STOI was not tested for conditions with reverberation.
Real-life background noise was added as an additional category and contained signals like multi-talker babble, music, and car noise (Rhebergen *et al.*, 2008; Francart *et al.*, 2011). These are more representative of daily listening conditions, but vary highly in their temporal and spectral characteristics. In this category, the ESTOI and mr-sEPSM perform surprisingly poor.
The model prediction using the ESTOI and mr-sEPSM were used with the standard set of available parameters. To fit the current speech data more accurately, extra tuning of these parameters is required. Results from the original publications of the model with similar noise types show a higher accuracy. However, the results from Table 7-3 clearly show a trend. The mr-sEPSM has problems with interrupted noises and the accuracy of the ESTOI diminishes when reverberation is present.

7.4.4 In summary

An added value of models like mr-sEPSM, the ESTOI, and the short-time speech-based STI-models, is the better applicability to nonlinear signal processing. Here, these models perform better than the classic STI (and also than the ESTI). However, the current extension of the STI was not focused on nonlinear processing, but was designed for room acoustics and aimed to

accurately model intelligibility for a wide range of non-stationary noises, including interrupted noises. One of the problems with other models that were designed for non-stationary noises, is the use of relatively long time windows. Interrupted noises at high rates still provide the listener with useable glimpses as short as 8–10 ms (Miller and Licklider, 1950; Rhebergen *et al.*, 2006). For example, only the filters of the modulation filterbank by Jørgensen *et al.* (2013) that are tuned to 128 and 256 Hz have windows that are short enough to capture these glimpses. However, modulations in speech at these frequencies play a minor role in intelligibility. The temporal window that corresponds to the dominant modulation frequency in speech of ~4 Hz (Houtgast and Steeneken, 1978) is approximately 250 ms long. The higher inaccuracies in interrupted noise of the mr-sEPSM shown in Table 7-3 are a consequence of this modelling choice. It is of course the question how relevant interrupted noises are in daily life. A competing speaker is a more frequent occurring form of distortion and both the ESTOI and the mr-sEPSM outperform the ESTI and cESTI, especially in the absence of reverberation.
It is obvious that a relatively simple acoustic measure like the STI and its derivations cannot compete with the powerful performance of ASR-based models. Not now, and certainly not within the near future, since the accuracy of methods based on machine learning will only improve. However, ASR-based models tend to be highly complex, and need a large amount of training data, often for each separate condition. In theory, these models are interesting, but the practical implementation is therefore not always straightforward. And this point is exactly the major strength of the STI: a single noise recording, together with a recording of an impulse response is sufficient to calculate the STI and directly draw conclusions. Besides its easy applicability, the STI is thoroughly evaluated over the past decades and clearly specified in IEC60268-16 (2011; 2020).

7.5 ESTI and hearing loss

The focus of the current study was on normally hearing subjects. Of course, in clinical practice, the STI can be a valuable tool to evaluate the (acoustical) circumstances in which hearing-impaired people have to function. What role can the ESTI play here? People with sensorineural hearing impairment experience less fluctuating masker benefit than normally hearing subjects (e.g., Festen and Plomp, 1990). The fluctuating masker benefit in people with normal hearing is adequately modelled by the ESTI, especially when noises do not have speech-like characteristics (see chapter 3).

To provide a brief overview of the possibilities of the ESTI and hearing loss, the intelligibility data by De Laat and Plomp (1983) was used. This data was later re-examined by Rhebergen *et al.* (2010). Subjects with normal, and mildly to moderately impaired hearing were tested using SSN and 10 Hz interrupted noise as a masker. In each condition, all subjects were tested with stationary noise levels of 65, 75, and 85 dB (A) and interrupted noise levels of 62, 72, and 82 dB (A). The adaptive procedure by Plomp and Mimpen (1979) was used to determine the cSNR.

In normally hearing subjects, the ESTI-value at cSNR was 0.37 (+/– 0.008) when using SSN as a masker. For the 10 Hz interrupted masker the average ESTI-value was 0.296 (+/-0.03). Here, the ESTI tended to increase with increasing noise levels, but overall, it was slightly lower for non-stationary noise than for SSN. The ESTI in SSN for hearing-impaired subjects was 0.494 (+/– 0.02). This corresponds well to the classic STI data found by Duquesnoy and Plomp (1980) who found values between 0.46 and 0.51 for the subgroup with similar auditory thresholds. In interrupted noise the average ESTI was 0.653 (+/– 0.05). In this case, the ESTI was heavily influenced by the noise level, with an increase of 0.01 STI-units per dB of noise. The average ESTI at cSNR was 0.15 units higher in interrupted noise than in stationary noise. This reflects the limited benefit hearing-impaired subjects receive from gaps in the noise.

Implementation of individual hearing thresholds seems to be an important factor here. A first order approximation was to adjust the hearing thresholds as they are currently used in the ESTI-model. A moderate, slightly sloping hearing loss that satisfied the average thresholds provided in Table 1 of Rhebergen *et al.* (2010) was superimposed on the auditory thresholds already available in the model. The resulting ESTI-values at cSNR in stationary noise were 0.356 (+/– 0.05) and in interrupted noise 0.331 (+/– 0.02). Note that only audibility and the forward masking function were adjusted here. The distortion factor as in the model by Plomp (1986) was not accounted for. Duquesnoy and Plomp (1980) stated that this distortion factor (i.e., the hearing loss in dB for speech in SSN) can be expressed as a shift of 0.033 STI units per dB distortion. In order to model intelligibility in fluctuating noise for hearing-impaired subjects more accurately, the individual distortion factor should also be taken into account. This is in line with the statement by Meyer and Brand (2013) that the pure tone audiogram does not seem to be sufficient to estimate the cSNR in non-stationary noises.

An important question here is how the ESTI should be applied to hearing-impaired subjects. One option is to predict intelligibility as discussed above. In order to do so, the individual hearing threshold and distortion factor could be included in the model. Another option is to calculate the ESTI according to the normal calculation scheme from chapter 3, and relate the resulting values to a

certain qualification interval. In Table I.1 of IEC60268-16 (2011), four qualification intervals are provided which are coupled to STI-values. Generally, a STI of 0.75–1.00 represent excellent circumstances, 0.60–0.75 good, 0.45–0.60 fair, 0.30–0.45 poor, and 0.00–0.30 bad. These are the labels for normally hearing persons. For aged listeners with an average hearing loss of 30 dB, a STI of 0.51 is classified as Bad-Poor, and 0.66 is classified as Poor-Fair. Higher qualifications cannot be achieved. Adjusted qualifications are necessary for the ESTI, since the current versions underestimate acoustic circumstances with non-stationary noises. After all, the hearing-impaired subjects described above needed a regular ESTI in SSN of 0.494 to reach 50% intelligibility. However, they needed an ESTI of 0.653 in interrupted noise.

7.6 Suggested applications

An important application of the ESTI is the suitability for conditions with non-stationary background noise. A good example is a classroom where a teacher experiences problems with speech intelligibility. To investigate the acoustic circumstances in which the teacher has to work, the STI can be used. In order to avoid inaccuracies due to non-stationary background noise, measurements are often done after school hours. However, several factors are not taken into account in this situation, such as the true noise characteristics, the influence of the children on the acoustical properties of the room and the daily noises from outside the classroom. With the ESTI, this limitation can be overcome by applying the indirect measurement method using the conditions described in chapter 2, and the ESTI calculation scheme of chapter 3. As a consequence, the results of the ESTI measurement will better reflect the actual circumstances of the classroom during school hours.

Another possibility that has opened up with the new calculation scheme of the ESTI, is a time-dependent visualization of the STI. As an example, an employer at a train station might experience difficulties in understanding speech, but only when trains pass by, or stop at the station. With the classic STI, the train noise would have caused measurement inaccuracies, invalidating the measured STI-value. With the indirect measurement approach as suggested in section 4.4 of IEC60268-16 (2011), this issue would be resolved. However, the fluctuations would probably average out in the separate noise recording, leading to a STI-value which is more representative for the quiet intervals between passing trains, but not for the situation where the actual problem occurs. The current ESTI approach presents the opportunity to visualize the ESTI as a function of time. The quiet intervals could then be analyzed separately from the moments of

passing trains. This can be the basis of a more thorough, quantitative assessment of the working place.
As mentioned earlier, the STI and SII converged over the years and their calculation schemes now show a lot of similarities. It is therefore a logical option to also investigate the added value of a context-based approach of the ESII. For example, Rhebergen *et al.* (2006) showed that the ESII could accurately predict 50% sentence intelligibility for speech in interrupted noise with rates of 8 Hz and higher. However, the ESII drastically overestimated intelligibility in 4 Hz interrupted noise, because context was not accounted for. A similar approach might resolve this issue and is relatively easy to implement.

7.7 Future of the Speech Transmission Index

The STI is a relatively simple but powerful measure, developed in the 1970s and 1980s. What is the role of the STI in the current landscape of complex, AI-based models? A simple search in Web of Science tells us that the term Speech Transmission Index was used as a topic in scientific papers almost 300 times since 1980. When analyzing this data more thoroughly, a trend is visible towards more uses of the term each year, with a plateau of approximately 17 per year since 2013 (see Fig. 7-4). Of course, the body of research in general has rapidly grown over the past forty years, so it is more accurate to display this number relative to all the publications available. Still, the number of mentions per million

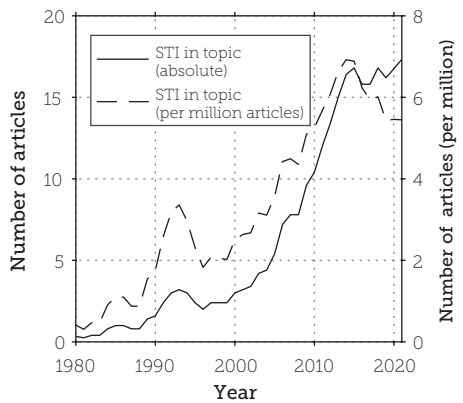


Fig. 7-4: Scientific publications found on Web of Science with the term "Speech Transmission Index" in the topic. The solid line represents the absolute number of publications per year (corresponding to the left y-axis). The dashed line represents the number of publications per million for each year (corresponding to the right y-axis).

articles on Web of Science shows an increasing trend until 2014. After that, a decrease is seen.
So, in the past decade a scientific paper about the STI is published almost every three weeks. Recent citations span a variety of topics, ranging from classroom acoustics (Leccese *et al.*, 2018) to open-plan offices (Cabrera *et al.*, 2018), and from the estimation of MTFs using neural networks (Duangpummet *et al.*, 2022) to the assessment of the recording quality of a new type of acoustic actuator (Dipassio *et al.*, 2022).
The STI is a metric that is subjected to ongoing research. It is likely that technologies based on artificial intelligence will play a role in the estimation and optimization of the STI. As mentioned earlier, novel models will continue to emerge with an increasing pace due to the explosion of algorithms driven by machine learning. However, the need for a simple and powerful measure of speech quality will remain. Since the STI is embedded within today's standards, it is likely that it will continue to be used for this purpose in the future. Of course, metrics evolve and the current research is a contribution to this evolution. But the modulation-based assessment of speech intelligibility will remain as important as it is today.

CHAPTER 8

GENERAL CONCLUSIONS

General conclusions

In section 1.4 the objectives of the current work were formulated. The primary goal was to increase the usability of the STI in non-stationary background noise. Besides this, the authors aimed to remain close to the original calculation scheme in order to minimize complexity and maintain robustness.
Generally speaking, this thesis consisted of three topics:

1) The circumstances under which (E)STI measurements in non-stationary noise should take place (chapter 2)
2) The introduction of a temporal extension of the classic STI to account for the increase in intelligibility when gaps in the noise are introduced (chapter 3)
3) The addition of context models to the Extended STI to account for noises with slow modulations (chapters 4, 5 and 6)

The ESTI-model was developed to deal with two limitations of the classic STI. First, the model needed to account for the fluctuating masker benefit. Besides this, the measurement method needed to be suitable for non-stationary background noises. The latter aspect was addressed in chapter 2. The indirect measurement method in non-stationary noise proved to be suitable to estimate the impulse response that was necessary to calculate the Modulation Transfer Function. This conclusion paved the way for the introduction of the temporal extension of the STI in chapter 3. When calculating the STI per time window and averaging all values across the entire signal to obtain one index, the prediction accuracy increased when compared to the classic STI. This was the case for all non-stationary noises and reverberation times that were analyzed, although predictions in noises with speech-like properties still showed inaccuracies.
The primary objective of this thesis was met in chapters 2 and 3. The combination of the suggested measurement conditions and the extension of the STI clearly increased the usability in non-stationary noises. The complexity of the calculation scheme did increase as a result of the introduction of forward masking and the temporal extension, but, in the opinion of the author, this increase is proportional to the higher accuracy.
The formulation and evaluation of a context-based version of the ESTI in chapters 4, 5 and 6 aimed to deal with the remaining inaccuracies discussed in chapter 3. The addition of context was successful in predicting the intelligibility of monosyllabic words in interrupted noise with low interruption rates. It was particularly accurate when predicting the typical dip in intelligibility that is often seen in interrupted noise around a rate of 1 Hz. However, the model still showed inaccuracies for maskers with speech-like characteristics in the prediction of sentence intelligibility.

The introduction of context to the ESTI marked an important deviation from the classic STI. The outcome of the classic STI and the ESTI is a single index value that can be used to predict intelligibility when a transfer function is available. This index value can also be used to evaluate the quality of speech transmission from a talker to a listener (which was the original purpose of the STI). On the contrary, the outcome of the cESTI is an estimate of intelligibility only and the index values are merely calculated as an intermediate step. In other words, the cESTI cannot be used as a direct replacement for the STI or ESTI, since it is not suitable to evaluate the quality of a transmission channel. However, this addition does show how a context model can be successfully used in combination with traditional methods for estimating speech intelligibility.

An important limitation of the ESTI-model is its poorer performance in non-stationary noises with speech-like characteristics. Stationary noises are relatively homogeneous, with the primary variation in their spectral content. Non-stationary noises are infinitely more variable, leading to difficulties in constructing a model that explains all outcomes. Intelligibility models are inherently an oversimplification of human speech recognition. Keeping a model relatively simple and usable is impossible to unite with a model that accurately predicts intelligibility in all possible conditions. Merely the concept of modulation reduction cannot account for complex interactions between the envelopes of speech and noise, or for modulation masking and informational masking. An important objective of the current research was to make the ESTI applicable in the same way the STI is being used today. Within that framework and despite the remaining inaccuracies, the ESTI led to a significant increase in prediction performance when compared to the classic STI.

References

Adams, E. M., Gordon-Hickey, S., Morlas, H., and Moore, R. (**2012**). "Effect of rate-alteration on speech perception in noise in older adults with normal hearing and hearing impairment," Am J Audiol **21**, 22-32.

Adams, E. M., and Moore, R. E. (**2009**). "Effects of Speech Rate, Background Noise, and Simulated Hearing Loss on Speech Rate Judgment and Speech Intelligibility in Young Listeners," J Am Acad Audiol **20**, 028-039.

Akeroyd, M. A. (**2006**). "The psychoacoustics of binaural hearing," Int J Audiol **45**, 25-33.

Andersen, A. H., Haan, J. M. d., Tan, Z. H., and Jensen, J. (**2018**). "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**, 1925-1939.

ANSI-S1.11 (**2004**) in *Specification for Octave-Band and Fractional-Octave-Band Analog and Digital Filters* (American National Standards Institute, New York).

ANSI-S3.5 (**1969**) in *Methods for Calculation of the Articulation Index* (American National Standards Institute, New York).

ANSI-S3.5 (**1997**) in *Methods for Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).

Apoux, F., and Bacon, S. P. (**2008**). "Selectivity of modulation interference for consonant identification in normal-hearing listeners," J Acoust Soc Am **123**, 1665-1672.

Baskent, D. (**2012**). "Effect of speech degradation on top-down repair: phonemic restoration with simulations of cochlear implants and combined electric-acoustic stimulation," J Assoc Res Otolaryngol **13**, 683-692.

Baskent, D., Eiler, C. L., and Edwards, B. (**2010**). "Phonemic restoration by hearing-impaired listeners with mild to moderate sensorineural hearing loss," Hear Res **260**, 54-62.

Bell, T. S., Dirks, D. D., and Trine, T. D. (**1992**). "Frequency-importance functions for words in high- and low-context sentences," J Speech Hear Res **35**, 950-959.

Benoit, C. (**1990**). "An intelligibility test using semantically unpredictable sentences towards the quantification of linguistic complexity," Speech Commun **9**, 293-304.

Bentler, R. A., and Pavlovic, C. V. (**1989**). "Transfer functions and correction factors used in hearing aid evaluation and research," Ear Hear **10**, 58-63.

Beranek, L., and Mellow, T. (**2012**). *Acoustics: Sound fields and Transducers* (Elsevier Inc., Waltham, MA, USA).

Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J Acoust Soc Am **120**, 331-342.

Beutelmann, R., Brand, T., and Kollmeier, B. (**2009**). "Prediction of binaural speech intelligibility with frequency-dependent interaural phase differences," J Acoust Soc Am **126**, 1359-1368.

Beutelmann, R., Brand, T., and Kollmeier, B. (**2010**). "Revision, extension, and evaluation of a binaural speech intelligibility model," J Acoust Soc Am **127**, 2479-2497.

Biberger, T., and Ewert, S. D. (**2016**). "Envelope and intensity based prediction of psychoacoustic masking and speech intelligibility," J Acoust Soc Am **140**, 1023.

Biberger, T., and Ewert, S. D. (**2017**). "The role of short-time intensity and envelope power for speech intelligibility and psychoacoustic masking," J Acoust Soc Am **142**, 1098.

Boothroyd, A. (**1968**). "Statistical theory of the speech discrimination score," J Acoust Soc Am **43**, 362-367.

Boothroyd, A. (**2004**). "Room Acoustics and Speech Perception," Semin Hear **25**, 155-166.

Boothroyd, A. (**2008**). "The Performance/Intensity Function: An Underused Resource," Ear Hearing **29**.

Boothroyd, A., and Nittrouer, S. (**1988**). "Mathematical treatment of context effects in phoneme and word recognition," J Acoust Soc Am **84**, 101-114.

Bosman, A. J. (**1989**). "Speech Perception by the Hearing Impaired," (University of Utrecht, Utrecht, The Netherlands).

Bosman, A. J., and Smoorenburg, G. F. (**1995**). "Intelligibility of Dutch CVC syllables and sentences for listeners with normal hearing and with three types of hearing impairment," Audiology **34**, 260-284.

Bronkhorst, A. W. (**2015**). "The cocktail-party problem revisited: early processing and selection of multi-talker speech," Atten Percept Psychophys **77**, 1465-1487.

Bronkhorst, A. W., Bosman, A. J., and Smoorenburg, G. F. (**1993**). "A model for context effects in speech recognition," J Acoust Soc Am **93**, 499-509.

Bronkhorst, A. W., Brand, T., and Wagener, K. (**2002**). "Evaluation of context effects in sentence recognition," J Acoust Soc Am **111**, 2874-2886.

Bronkhorst, A. W., and Houtgast, T. (**1990**). "STI approach for predicting the effect of fluctuating interference on speech intelligibility," J Acoust Soc Am **87**, S126.

Bronkhorst, A. W., and Plomp, R. (**1989**). "Binaural speech intelligibility in noise for hearing-impaired listeners," J Acoust Soc Am **86**, 1374-1383.

Bronkhorst, A. W., and Plomp, R. (**1990**). "A Clinical-Test for the Assessment of Binaural Speech-Perception in Noise," Audiology **29**, 275-285.

Brouwer, S., Van Engen, K. J., Calandruccio, L., and Bradlow, A. R. (**2012**). "Linguistic contributions to speech-on-speech masking for native and non-native listeners: language familiarity and semantic content," J Acoust Soc Am **131**, 1449-1464.

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

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M. N., Nasser, N. H. A., Elkholy, W. A. H., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G. I., Westerman, S., and Ludvigsen, C. (**1994**). "An International Comparison of Long-Term Average Speech Spectra," J Acoust Soc Am **96**, 2108-2120.

Cabrera, D., Yadav, M., and Protheroe, D. (**2018**). "Critical methodological assessment of the distraction distance used for evaluating room acoustic quality of open-plan offices," Applied Acoustics **140**, 132-142.

Chabot-Leclerc, A., Jørgensen, S., and Dau, T. (**2014**). "The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction," J Acoust Soc Am **135**, 3502-3512.

Chabot-Leclerc, A., MacDonald, E. N., and Dau, T. (**2016**). "Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain," J Acoust Soc Am **140**, 192.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., and Shamma, S. (**1999**). "Spectro-temporal modulation transfer functions and speech intelligibility," The Journal of the Acoustical Society of America **106**, 2719-2732.

Collin, B., and Lavandier, M. (**2013**). "Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers," The Journal of the Acoustical Society of America **134**, 1146-1159.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J Acoust Soc Am **119**, 1562-1573.

Culling, J. F., and Summerfield, Q. (**1998**). "Measurements of the binaural temporal window using a detection task," J Acoust Soc Am **103**, 3540-3553.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997a**). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," J Acoust Soc Am **102**, 2892-2905.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997b**). "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," J Acoust Soc Am **102**, 2906-2919.

Dau, T., Verhey, J., and Kohlrausch, A. (**1999**). "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers," J Acoust Soc Am **106**, 2752-2760.

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. Hartmann (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 359-363.

Desloge, J. G., Reed, C. M., Braida, L. D., Perez, Z. D., and Delhorne, L. A. (**2010**). "Speech reception by listeners with real and simulated hearing impairment: Effects of continuous and interrupted noise," J Acoust Soc Am **128**, 342-359.

Dieudonne, B., and Francart, T. (**2019**). "Redundant Information Is Sometimes More Beneficial Than Spatial Information to Understand Speech in Noise," Ear Hear **40**, 545-554.

Dingemanse, J. G., and Goedegebure, A. (**2019**). "The Important Role of Contextual Information in Speech Perception in Cochlear Implant Users and Its Consequences in Speech Tests," Trends Hear **23**, 2331216519838672.

Dipassio, T., Heilemann, M. C., and Bocko, M. F. (**2022**). "Audio Capture Using Structural Sensors on Vibrating Panel Surfaces," J Audio Eng Soc **70**, 1027-1037.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (**2001**). "ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. International Collegium for Rehabilitative Audiology," Audiology **40**, 148-157.

Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility," J Acoust Soc Am **97**, 585-592.

Duangpummet, S., Karnjana, J., Kongprawechnon, W., and Unoki, M. (**2022**). "Blind estimation of speech transmission index and room acoustic parameters based on the extended model of room impulse response," Applied Acoustics **185**, 108372.

Dubbelboer, F., and Houtgast, T. (**2007**). "A detailed study on the effects of noise on speech intelligibility," J Acoust Soc Am **122**, 2865-2871.

Dubbelboer, F., and Houtgast, T. (**2008**). "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility," J Acoust Soc Am **124**, 3937-3946.

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

Duquesnoy, A. J., and Plomp, R. (**1980**). "Effect of Reverberation and Noise on the Intelligibility of Sentences in Cases of Presbyacusis," J Acoust Soc Am **68**, 537-544.

Durlach, N. (**2006**). "Auditory masking: need for improved conceptual structure," J Acoust Soc Am **120**, 1787-1790.

Durlach, N. I., Mason, C. R., Kidd, G., Jr., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking," J Acoust Soc Am **113**, 2984-2987.

Edraki, A., Chan, W.-Y., Jensen, J., and Fogerty, D. (**2019**). *Improvement and Assessment of Spectro-Temporal Modulation Analysis for Speech Intelligibility Estimation*.

Edraki, A., Chan, W.-Y., Jensen, J., and Fogerty, D. (**2021a**). "Speech intelligibility prediction using spectro-temporal modulation analysis," IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 210-225.

Edraki, A., Chan, W.-Y., Jensen, J., and Fogerty, D. (**2022**). "Spectro-temporal modulation glimpsing for speech intelligibility prediction," Hearing Research **426**, 108620.

Edraki, A., Chan, W. Y., Jensen, J., and Fogerty, D. (**2021b**). "A spectro-temporal glimpsing index (STGI) for speech intelligibility prediction," in *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021* (International Speech Communication Association), pp. 2738-2742.

Egan, J. P. (**1948**). "Articulation testing methods," Laryngoscope **58**, 955-991.

Elhilali, M., Chi, T., and Shamma, S. A. (**2003**). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun **41**, 331-348.

Ewert, S. D., and Dau, T. (**2000**). "Characterizing frequency selectivity for envelope fluctuations," J Acoust Soc Am **108**, 1181-1196.

Falk, T. H., Parsa, V., Santos, J. F., Arehart, K., Hazrati, O., Huber, R., Kates, J. M., and Scollie, S. (**2015**). "Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools," IEEE Signal Processing Magazine **32**, 114-124.

Falk, T. H., Zheng, C., and Chan, W. Y. (**2010**). "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," IEEE Transactions on Audio, Speech, and Language Processing **18**, 1766-1774.

Favre-Félix, A., Graversen, C., Hietkamp, R. K., Dau, T., and Lunner, T. (**2018**). "Improving Speech Intelligibility by Hearing Aid Eye-Gaze Steering: Conditions With Head Fixated in a Multitalker Environment," Trends in Hearing **22**, 2331216518814388.

Feng, Y., and Chen, F. (**2022**). "Nonintrusive objective measurement of speech intelligibility: A review of methodology," Biomedical Signal Processing and Control **71**, 103204.

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.
Fogerty, D. (**2011**). "Perceptual weighting of individual and concurrent cues for sentence intelligibility: frequency, envelope, and fine structure," J Acoust Soc Am **129**, 977-988.
Fogerty, D. (**2014**). "Importance of envelope modulations during consonants and vowels in segmentally interrupted sentences," J Acoust Soc Am **135**, 1568-1576.
Fogerty, D., Bologna, W. J., Ahlstrom, J. B., and Dubno, J. R. (**2017**). "Simultaneous and forward masking of vowels and stop consonants: Effects of age, hearing loss, and spectral shaping," J Acoust Soc Am **141**, 1133.
Fogerty, D., Xu, J., and Gibbs, B. E., 2nd (**2016**). "Modulation masking and glimpsing of natural and vocoded speech during single-talker modulated noise: Effect of the modulation spectrum," J Acoust Soc Am **140**, 1800.
Francart, T., van Wieringen, A., and Wouters, J. (**2011**). "Comparison of fluctuating maskers for speech recognition tests," Int J Audiol **50**, 2-13.
French, N. R., and Steinberg, J. C. (**1947**). "Factors Governing The Intelligibility of Speech Sounds," J Acoust Soc Am **19**, 90-119.
George, E. L. J., Festen, J. M., and Goverts, S. T. (**2012**). "Effects of reverberation and masker fluctuations on binaural unmasking of speech," J Acoust Soc Am **132**, 1581-1591.
George, E. L. J., Festen, J. M., and Houtgast, T. (**2006**). "Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners," J Acoust Soc Am **120**, 2295-2311.
George, E. L. J., Festen, J. M., and Houtgast, T. (**2008**). "The combined effects of reverberation and nonstationary noise on sentence intelligibility," J Acoust Soc Am **124**, 1269-1277.
George, E. L. J., Goverts, S. T., Festen, J. M., and Houtgast, T. (**2010**). "Measuring the Effects of Reverberation and Noise on Sentence Intelligibility for Hearing-Impaired Listeners," J Speech Lang Hear R **53**, 1429-1439.
Gifford, R. H., Bacon, S. P., and Williams, E. J. (**2007**). "An examination of speech recognition in a modulated background and of forward masking in younger and older listeners," J Speech Lang Hear Res **50**, 857-864.
Goldsworthy, R. L., and Greenberg, J. E. (**2004**). "Analysis of speech-based Speech Transmission Index methods with implications for nonlinear operations," J Acoust Soc Am **116**, 3679-3689.
Hak, C. C. J. M., Hak, J. P. M., and van Luxemburg, C. J. (**2008**). "INR as an estimator for the decay range of room acoustic impulse responses," in *AES Convention* (Amsterdam).
Hak, C. C. J. M., Wenmaekers, R. H. C., and van Luxemburg, L. C. J. (**2012**). "Measuring Room Impulse Responses: Impact of the Decay Range on Derived Room Acoustic Parameters," Acta Acust United Ac **98**, 907-915.
Hohmann, V., and Kollmeier, B. (**1995**). "The effect of multichannel dynamic compression on speech intelligibility," J Acoust Soc Am **97**, 1191-1195.
Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (**2010**). "Development and analysis of an International Speech Test Signal (ISTS)," Int J Audiol **49**, 891-903.
Holube, I., and Kollmeier, B. (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J Acoust Soc Am **100**, 1703-1716.
Houtgast, T. (**1978**). "Geknikte nagalmcurven en verstaanbaarheid," Nederlands Akoestisch Genootschap **44**, 1-9.
Houtgast, T. (**1989**). "Frequency selectivity in amplitude-modulation detection," J Acoust Soc Am **85**, 1676-1680.
Houtgast, T., and Steeneken, H. J. M. (**1973**). "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica **28**, 66-73.
Houtgast, T., and Steeneken, H. J. M. (**1978**). "Modulation Transfer-Function as a Link between Room Acoustics and Speech-Intelligibility," J Acoust Soc Am **63**, S6-S6.

Houtgast, T., and Steeneken, H. J. M. (**1985**). "A Review of the MTF Concept in Room Acoustics and Its Use for Estimating Speech-Intelligibility in Auditoria," J Acoust Soc Am **77**, 1069-1077.
Houtgast, T., and Steeneken, H. J. M. (**2002**). "The roots of the STI approach," in *Past, Present and Future of the Speech Transmission Index*, edited by S. J. van Wijngaarden (TNO Human Factors, Soesterberg), pp. 3-12.
Houtgast, T., Steeneken, H. J. M., and Plomp, R. (**1980**). "Predicting Speech-Intelligibility in Rooms from the Modulation Transfer-Function .1. General Room Acoustics," Acustica **46**, 60-72.
Howard-Jones, P. A., and Rosen, S. (**1993**). "Uncomodulated glimpsing in "checkerboard" noise," J Acoust Soc Am **93**, 2915-2922.
Hülsmeier, D., Schädler, M. R., and Kollmeier, B. (**2021**). "DARF: A data-reduced FADE version for simulations of speech recognition thresholds with real hearing aids," Hearing Research **404**, 108217.
IEC60268-16 (**2011**). "Edition 4.0," in *Sound system equipment, Part 16: Objective rating of speech intelligibility by speech transmission index* (International Electrotechnical Commission, Geneva, Switzerland).
IEC60268-16 (**2020**). "Edition 5.0," in *Sound system equipment, Part 16: Objective rating of speech intelligibility by speech transmission index* (International Electrotechnical Commission, Geneva, Switzerland).
ISO3382-1 (**2009**) in *Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces* (International Organization for Standardization, Geneva).
ISO3382-2 (**2008**) in *Acoustics – Measurement of room acoustic parameters – Part 2: Reverberation time in ordinary rooms* (International Organization for Standardization, Geneva).
Janse, E. (**2003**). "Production and Perception of Fast Speech," (University of Utrecht, Utrecht, The Netherlands).
Jensen, J., and Taal, C. H. (**2016**). "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," IEEE/ACM Trans. Audio, Speech, Lang. Process **24**, 2009-2022.
Jesteadt, W., Bacon, S. P., and Lehman, J. R. (**1982**). "Forward masking as a function of frequency, masker level, and signal delay," J Acoust Soc Am **71**, 950-962.
Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J Acoust Soc Am **130**, 1475-1487.
Jørgensen, S., Ewert, S. D., and Dau, T. (**2013**). "A multi-resolution envelope-power based model for speech intelligibility," J Acoust Soc Am **134**, 436-446.
Karbasi, M., and Kolossa, D. (**2022**). "ASR-based speech intelligibility prediction: A review," Hear Res, 108606.
Kates, J. M. (**1987**). "The short-time articulation index," J Rehabil Res Dev **24**, 271-276.
Kidd, G., and Feth, L. L. (**1981**). "Patterns of residual masking," Hearing Research **5**, 49-67.
Killion, M. C. (**1978**). "Revised estimate of minimum audible pressure: where is the "missing 6 dB"?," J Acoust Soc Am **63**, 1501-1508.
Koelewijn, T., Zekveld, A. A., Festen, J. M., Ronnberg, J., and Kramer, S. E. (**2012**). "Processing load induced by informational masking is related to linguistic abilities," Int J Otolaryngol **2012**, 865731.
Kollmeier, B., Schädler, M. R., Warzybok, A., Meyer, B. T., and Brand, T. (**2016**). "Sentence Recognition Prediction for Hearing-impaired Listeners in Stationary and Fluctuation Noise With FADE: Empowering the Attenuation and Distortion Concept by Plomp With a Quantitative Processing Model," Trends Hear **20**.
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. (**1962**). "Validation of the Articulation Index," The Journal of the Acoustical Society of America **34**, 1698-1702.
Kwon, B. J., and Turner, C. W. (**2001**). "Consonant identification under maskers with sinusoidal modulation: masking release or modulation interference?," J Acoust Soc Am **110**, 1130-1140.
Lavandier, M., and Culling, J. F. (**2010**). "Prediction of binaural speech intelligibility against noise in rooms," J Acoust Soc Am **127**, 387-399.

Leccese, F., Rocca, M., and Salvadori, G. (**2018**). "Fast estimation of Speech Transmission Index using the Reverberation Time: Comparison between predictive equations for educational rooms of different sizes," Applied Acoustics **140**, 143-149.

Lochner, J. P. A., and Burger, J. F. (**1964**). "The influence of reflections on auditorium acoustics," Journal of Sound and Vibration **1**, 426-454.

Ludvigsen, C. (**1985**). "Relations among some psychoacoustic parameters in normal and cochlearly impaired listeners," J Acoust Soc Am **78**, 1271-1280.

Ludvigsen, C., Elberling, C., and Keidser, G. (**1993**). "Evaluation of a noise reduction method--comparison between observed scores and scores predicted from STI," Scand Audiol Suppl **38**, 50-55.

Ludvigsen, C., Elberling, C., Keidser, G., and Poulsen, T. (**1990**). "Prediction of Intelligibility of Non-linearly Processed Speech," Acta Otolaryngol **109**, 190-195.

Maalderink, T. M., Rhebergen, K. S., and Dreschler, W. A. (**2011**). "Psychometric measurements for speech intelligibility in different noise types (after wide dynamic range compression)," poster presented at: *International Symposium on Auditory and Audiological Research (ISAAR), Nyborg, Denmark*.

Merriam-Webster.com (**January 24, 2023**). "Noise."

Meyer, R. M., and Brand, T. (**2013**). "Comparison of Different Short-Term Speech Intelligibility Index Procedures in Fluctuating Noise for Listeners with Normal and Impaired Hearing," Acta Acust United Ac **99**, 442-456.

Miller, G. A., Heise, G. A., and Lichten, W. (**1951**). "The intelligibility of speech as a function of the context of the test materials," J Exp Psychol **41**, 329-335.

Miller, G. A., and Licklider, J. C. R. (**1950**). "The Intelligibility of Interrupted Speech," The Journal of the Acoustical Society of America **22**, 167-173.

Moore, B. C. (**2007**). *Cochlear Hearing Loss* (John Wiley & Sons, Ltd, Chichester).

Moore, B. C., 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.

Nabelek, A. K., Letowski, T. R., and Tucker, F. M. (**1989**). "Reverberant overlap- and self-masking in consonant identification," J Acoust Soc Am **86**, 1259-1265.

Nguyen, T.-S., Stüker, S., and Waibel, A. (**2020**). *Super-Human Performance in Online Low-latency Recognition of Conversational Speech*.

Nielsen, J. B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int J Audiol **50**, 202-208.

Noordhoek, I. M., and Drullman, R. (**1997**). "Effect of reducing temporal intensity modulations on sentence intelligibility," J Acoust Soc Am **101**, 498-502.

Olsen, W. O. (**1998**). "Average Speech Levels and Spectra in Various Speaking/Listening Conditions," American Journal of Audiology **7**, 21-25.

Patro, C., and Mendel, L. L. (**2016**). "Role of contextual cues on the perception of spectrally reduced interrupted speech," J Acoust Soc Am **140**, 1336.

Pavlovic, C. V. (**1984**). "Use of the articulation index for assessing residual auditory function in listeners with sensorineural hearing impairment," J Acoust Soc Am **75**, 1253-1258.

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

Payton, K. L., and Braida, L. D. (**1999**). "A method to determine the speech transmission index from speech waveforms," J Acoust Soc Am **106**, 3637-3648.

Payton, K. L., and Braida, L. D. (**2002**). "Computing the STI using speech as a probe stimulus," in *Past, Present and Future of the Speech Transmission Index*, edited by S. J. van Wijngaarden (TNO Human Factors, Soesterberg), pp. 125-137.

Payton, K. L., and Shrestha, M. (**2013**). "Comparison of a short-time speech-based intelligibility metric to the speech transmission index and intelligibility data," J Acoust Soc Am **134**, 3818-3827.

Payton, K. L., Uchanski, R. M., and Braida, L. D. (**1994**). "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," J Acoust Soc Am **95**, 1581-1592.

Picard, M., and Bradley, J. S. (**2001**). "Revisiting speech interference in classrooms," Audiology **40**, 221-244.

Plomp, R. (**1964**). "Rate of Decay of Auditory Sensation," The Journal of the Acoustical Society of America **36**, 277-282.

Plomp, R. (**1976**). "Binaural and Monaural Speech-Intelligibility of Connected Discourse in Reverberation as a Function of Azimuth of a Single Competing Sound Source (Speech or Noise)," Acustica **34**, 201-211.

Plomp, R. (**1986**). "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," J Speech Hear Res **29**, 146-154.

Plomp, R., and Duquesnoy, A. J. (**1980**). "Room acoustics for the aged," J Acoust Soc Am **68**, 1616-1621.

Plomp, R., and Mimpen, A. M. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Audiology **18**, 43-52.

Prodi, N., and Visentin, C. (**2019**). "An experimental study of a time-frame implementation of the Speech Transmission Index in fluctuating speech-like noise conditions," Applied Acoustics **152**, 63-72.

Rennies, J., Brand, T., and Kollmeier, B. (**2011**). "Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet," J Acoust Soc Am **130**, 2999-3012.

Rennies, J., Schepker, H., Holube, I., and Kollmeier, B. (**2014**). "Listening effort and speech intelligibility in listening situations affected by noise and reverberation," J Acoust Soc Am **136**, 2642-2653.

Rhebergen, K. S., Pool, R. E., and Dreschler, W. A. (**2014**). "Characterizing the Speech Reception Threshold in hearing-impaired listeners in relation to masker type and masker level," J Acoust Soc Am **135**, 1491-1505.

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., de Laat, J. A. P. M., and Dreschler, W. A. (**2010**). "Modelling the speech reception threshold in non-stationary noise in hearing-impaired listeners as a function of level," Int J Audiol **49**, 856-864.

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.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J Acoust Soc Am **120**, 3988-3997.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2008**). "Prediction of the intelligibility for speech in real-life background noises for subjects with normal hearing," Ear Hearing **29**, 169-175.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2009**). "The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise," J Acoust Soc Am **126**, 3236-3245.

Rosen, S., Souza, P., Ekelund, C., and Majeed, A. A. (**2013**). "Listening to speech in a background of other talkers: effects of talker number and noise vocoding," J Acoust Soc Am **133**, 2431-2443.

Saija, J. D., Akyurek, E. G., Andringa, T. C., and Baskent, D. (**2014**). "Perceptual restoration of degraded speech is preserved with advancing age," J Assoc Res Otolaryngol **15**, 139-148.

Schädler, M. R., Warzybok, A., Ewert, S. D., and Kollmeier, B. (**2016**). "A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception," J Acoust Soc Am **139**, 2708.

Schädler, M. R., Warzybok, A., Hochmuth, S., and Kollmeier, B. (**2015**). "Matrix sentence intelligibility prediction using an automatic speech recognition system," Int J Audiol **54 Suppl 2**, 100-107.

Schädler, M. R., Warzybok, A., and Kollmeier, B. (**2018**). "Objective Prediction of Hearing Aid Benefit Across Listener Groups Using Machine Learning: Speech Recognition Performance With Binaural Noise-Reduction Algorithms," Trends in Hearing **22**, 2331216518768954.

Schlauch, R. S., Ries, D. T., and DiGiovanni, J. J. (**2001**). "Duration discrimination and subjective duration for ramped and damped sounds," J Acoust Soc Am **109**, 2880-2887.

Schlesinger, A. (**2012**). "Transient-based speech transmission index for predicting intelligibility in nonlinear speech enhancement processors," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3993-3996.

Schlueter, A., Lemke, U., Kollmeier, B., and Holube, I. (**2014**). "Intelligibility of time-compressed speech: the effect of uniform versus non-uniform time-compression algorithms," J Acoust Soc Am **135**, 1541-1555.

Schroeder, M. R. (**1978**). "Modulation transfer functions: definition and measurement (abstract)," in *Acoustics, Speech, and Signal Processing* (IEEE), p. 180.

Schroeder, M. R. (**1981**). "Modulation transfer functions: definition and measurment," Acustica **49**, 179-182.

Schubotz, W., Brand, T., Kollmeier, B., and Ewert, S. D. (**2016**). "Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features," J Acoust Soc Am **140**, 524.

Schwerin, B., and Paliwal, K. (**2014**). "An improved speech transmission index for intelligibility prediction," Speech Commun **65**, 9-19.

Shafiro, V., Fogerty, D., Smith, K., and Sheft, S. (**2018**). "Perceptual Organization of Interrupted Speech and Text," J Speech Lang Hear Res **61**, 2578-2588.

Shailer, M. J., and Moore, B. C. (**1983**). "Gap detection as a function of frequency, bandwidth, and level," J Acoust Soc Am **74**, 467-473.

Sherbecoe, R. L., and Studebaker, G. A. (**1990**). "Regression equations for the transfer functions of ANSI S3.5-1969," The Journal of the Acoustical Society of America **88**, 2482-2483.

Shield, B., and Dockrell, J. E. (**2004**). "External and internal noise surveys of London primary schools," J Acoust Soc Am **115**, 730-738.

Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cogn Sci **12**, 182-186.

Smits, C., and Zekveld, A. A. (**2021**). "Approaches to mathematical modeling of context effects in sentence recognition," J Acoust Soc Am **149**, 1371.

Smits, R. (**2000**). "Temporal distribution of information for human consonant recognition in VCV utterances," J. Phon. **28**, 111-135.

Smits, R., Warner, N., McQueen, J. M., and Cutler, A. (**2003**). "Unfolding of phonetic information over time: a database of Dutch diphone perception," J Acoust Soc Am **113**, 563-574.

Spille, C., Kollmeier, B., and Meyer, B. T. (**2018**). "Comparing human and automatic speech recognition in simple and complex acoustic scenes," Computer Speech & Language **52**, 123-140.

Steeneken, H. J. M. (**2002**). "Standardisation of performance criteria and assessment methods for speech communication," in *Past, Present and Future of the Speech Transmission Index*, edited by S. J. van Wijngaarden (TNO Human Factors, Soesterberg), pp. 117-124.

Steeneken, H. J. M., and Houtgast, T. (**1980**). "A Physical Method for Measuring Speech-Transmission Quality," J Acoust Soc Am **67**, 318-326.

Steeneken, H. J. M., and Houtgast, T. (**1999**). "Mutual dependence of the octave-band weights in predicting speech intelligibility," Speech Commun **28**, 109-123.

Steeneken, H. J. M., and Houtgast, T. (**2002**). "Validation of the revised STIr method," Speech Commun **38**, 413-425.

Stone, M. A., Fullgrabe, C., Mackinnon, R. C., and Moore, B. C. (**2011**). "The importance for speech intelligibility of random fluctuations in "steady" background noise," J Acoust Soc Am **130**, 2874-2881.

Stone, M. A., Fullgrabe, C., and Moore, B. C. (**2010**). "Relative contribution to speech intelligibility of different envelope modulation rates within the speech dynamic range," J Acoust Soc Am **128**, 2127-2137.

Stone, M. A., Fullgrabe, C., and Moore, B. C. (**2012**). "Notionally steady background noise acts primarily as a modulation masker of speech," J Acoust Soc Am **132**, 317-326.

Stone, M. A., and Moore, B. C. (**2004**). "Side effects of fast-acting dynamic range compression that affect intelligibility in a competing speech task," J Acoust Soc Am **116**, 2311-2323.

Studebaker, G. A., Sherbecoe, R. L., and Gilmore, C. (**1993**). "Frequency-importance and transfer functions for the Auditec of St. Louis recordings of the NU-6 word test," J Speech Hear Res **36**, 799-807.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," IEEE Trans. Audio Speech Lang. Process. **19**, 2125-2136.

Van Esch, T. E., Kollmeier, B., Vormann, M., Lyzenga, J., Houtgast, T., Hallgren, M., Larsby, B., Athalye, S. P., Lutman, M. E., and Dreschler, W. A. (**2013**). "Evaluation of the preliminary auditory profile test battery in an international multi-centre study," Int J Audiol **52**, 305-321.

Van Schijndel, N. H., Houtgast, T., and Festen, J. M. (**1999**). "Intensity discrimination of Gaussian-windowed tones: indications for the shape of the auditory frequency-time window," J Acoust Soc Am **105**, 3425-3435.

Van Schoonhoven, J., Rhebergen, K. S., and Dreschler, W. A. (**2017**). "Towards measuring the Speech Transmission Index in fluctuating noise: Accuracy and limitations," J Acoust Soc Am **141**, 818.

Van Schoonhoven, J., Rhebergen, K. S., and Dreschler, W. A. (**2019**). "The Extended Speech Transmission Index: Predicting speech intelligibility in fluctuating noise and reverberant rooms," J Acoust Soc Am **145**, 1178.

Van Schoonhoven, J., Rhebergen, K. S., and Dreschler, W. A. (**2022**). "A context-based approach to predict speech intelligibility in interrupted noise: Model design," J Acoust Soc Am **151**, 1404.

Van Schoonhoven, J., Rhebergen, K. S., and Dreschler, W. A. (**2023**). "A context-based approach to predict speech intelligibility in interrupted noise: Model evaluation [Manuscript submitted for publication]."

Van Wieringen, A., and Wouters, J. (**2008**). "LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands," Int J Audiol **47**, 348-355.

Van Wijngaarden, S. J., Bronkhorst, A. W., Houtgast, T., and Steeneken, H. J. M. (**2004**). "Using the Speech Transmission Index for predicting non-native speech intelligibility," J Acoust Soc Am **115**, 1281-1291.

Van Wijngaarden, S. J., and Drullman, R. (**2008**). "Binaural intelligibility prediction based on the speech transmission index," J Acoust Soc Am **123**, 4514-4523.

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.

Verschueren, E., Vanthornhout, J., and Francart, T. (**2020**). "The Effect of Stimulus Choice on an EEG-Based Objective Measure of Speech Intelligibility," Ear Hearing **41**.

Verschuure, J., and Brocaar, M. P. (**1983**). "Intelligibility of interrupted meaningful and nonsense speech with and without intervening noise," Percept Psychophys **33**, 232-240.

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.

Wagenaar, W. A. (**1969**). "Note on the Construction of Digram-Balanced Latin Squares," Psychol Bull **72**, 384-386.

Warren, R. M. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**, 392-393.

Warzybok, A., Rennies, J., Brand, T., Doclo, S., and Kollmeier, B. (**2013**). "Effects of spatial and temporal integration of a single early reflection on speech intelligibility," J Acoust Soc Am **133**, 269-282.

Wu, Y. H., Stangl, E., Chipara, O., Hasan, S. S., Welhaven, A., and Oleson, J. (**2018**). "Characteristics of Real-World Signal to Noise Ratios and Speech Listening Situations of Older Adults With Mild to Moderate Hearing Loss," Ear Hear **39**, 293-304.

Zekveld, A. A., Rudner, M., Johnsrude, I. S., and Ronnberg, J. (**2013**). "The effects of working memory capacity and semantic cues on the intelligibility of speech in noise," J Acoust Soc Am **134**, 2225-2234.

Zezario, R., Fu, S.-W., Fuh, C.-S., Tsao, Y., and Wang, H.-m. (**2020**). *STOI-Net: A Deep Learning based Non-Intrusive Speech Intelligibility Assessment Model*.

Summary

The Speech Transmission Index or STI is a widely used metric that was designed for the evaluation of the quality of speech transmission from a talker to a listener. It can be applied in communication technology and room acoustics to evaluate the transmission channel without having to conduct speech intelligibility measurements for that specific condition. Examples of use are the design and construction of auditoria and theatres, and the evaluation of a working environment when problems with speech intelligibility occur.

The STI is a relatively simple but robust index between 0 and 1 that is based on the reduction of modulations in the speech as a result of background noise and/or reverberation. This modulation reduction correlates well with the degree of speech intelligibility. The foundation of the STI is the Modulation Transfer Function. This MTF is calculated based on recordings that were traditionally made using the *direct* measurement method, by using modulated noise as a probe signal. An alternative approach is the *indirect* measurement method, which uses a separate recording of the background noise in combination with an impulse response measurement. When the MTF is known, it is converted to a STI-value in several steps. A STI-value of 0 represents a situation where no intelligibility is possible at all. An index of 0.75 or higher corresponds to circumstances where intelligibility is good to excellent. An index lower than 0.30 corresponds to bad/poor intelligibility. Note that this only applies to normally hearing, native listeners. When a transfer function between the STI and intelligibility is known for a certain speech corpus, speech intelligibility can be estimated based on the STI-value.

Despite the robustness and thorough evaluation of the STI, there are several weaknesses. When nonlinear distortions like compression or spectral subtraction occur, the accuracy of the STI decreases. However, the current research focused on room acoustics and therefore only dealt with linear types of distortions. Another limitation of the STI is its usability when background noise is non-stationary. This was the primary motivation for the current research. The background of this limitation is twofold. First, the traditional, direct measurement method which uses modulated noise as a probe signal is sensitive to fluctuations in the background noise. This can lead to under- or overestimation of the STI. Second, speech intelligibility increases in normally hearing subjects when gaps in the noise are introduced. This is caused by the ability to glimpse speech in temporal regions where the speech is least affected by the background noise. The STI is unable to deal with this so-called fluctuating masker benefit, leading to a STI-value that is not representative for the actual circumstances. The goal of the current thesis was to increase the usability of the STI in non-stationary

background noises by addressing both the measurement method and the fluctuating masker benefit.

In **chapter 2**, the conditions under which the STI can be accurately measured were investigated. The focus was on the indirect measurement method where the MTF was derived from an impulse response measurement and a long-term noise recording. This method is less sensitive to fluctuations in the noise, but under which conditions the measurement can be done reliably is not known. To investigate this more thoroughly, two experiments were carried out. Impulse response measurements (using a sweep signal) and noise recordings were conducted in a room with variable absorption, different levels of stationary and fluctuating background noise, and different sweep levels. In order to extrapolate the experimental findings to other acoustical conditions, a large number of other recording conditions were simulated. The experiments and simulations showed that a minimum impulse-to-noise ratio of +25 dB (corresponding to a sweep-to-noise ratio of −4 to +15 dB) in non-stationary noise was needed to accurately measure the STI.

The Extended STI or ESTI was introduced in **chapter 3**. The aim of the revised model was to account for the increased intelligibility when gaps in the noise are introduced. The primary adaptation was the calculation of the STI per 2 ms time windows instead of for the long-term signal. The average of all local STI-values yielded one ESTI-value. To deal with rapid offsets of the noise, forward masking was also introduced in the model. The model parameters were tuned using newly measured sentence intelligibility data in normally hearing subjects. The point of 50% intelligibility (cSNR) was measured using an adaptive procedure. Speech was distorted using a combination of noise (stationary noise and two types of non-stationary noise) and five degrees of reverberation. Model evaluation was done using intelligibility data from 10 studies in the literature. The ESTI proved to predict intelligibility better than the classic STI for all non-stationary noises that were used. Prediction inaccuracies were observed when background noise had speech-like characteristics. When only the noise envelope was speech-like, the offset between observed and predicted cSNR was approximately 3 – 4 dB. When the fine structure of the noise was also speech-like, an extra offset of 5 – 7 dB was found. The authors hypothesized that these inaccuracies were the result of modulation masking, informational masking, and/or context effects.

In **chapter 4** the hypothesis was tested that these inaccuracies were caused by context effects. When glimpsing speech in non-stationary noise with high modulation rates, the listener has access to parts of all speech elements. These elements can be the phonemes as part of a word, but also the words as part of a sentence. When modulations in the noise are slow, the probability increases

that speech elements are fully masked by the longer noise bursts and are not intelligible. The listener can then rely on contextual information to "guess" the missed elements. To account for this mechanism in speech intelligibility, context was added to the ESTI. First, the ESTI was calculated per speech element instead of for the whole signal. Then, a transfer function was estimated that related the ESTI per element to the isolated element score. When the latter score was known, the intelligibility of the whole utterance could be estimated using a context model. To evaluate the performance of this context-based ESTI or cESTI-model, existing intelligibility data of meaningful monosyllabic words in interrupted noise was analyzed. The addition of two existing context models was compared, referred to as $cESTI_1$ (with the Bronkhorst model) and $cESTI_2$ (with the Boothroyd and Nittrouer model). The prediction accuracy of the new model clearly improved for interruption frequencies below 5 Hz. The model versions $cESTI_1$ and $cESTI_2$ resulted in comparable performance.

The cESTI-model was evaluated in **chapter 5** using newly measured CVC-words in stationary and interrupted noise. Both nonsense and meaningful words were used. Only $cESTI_2$ (using the more simple of the two context models from chapter 4) was used in this chapter. The model outperformed the ESTI-model for both meaningful and nonsense words. However, despite the increased performance, intelligibility of meaningful words at rates below 5 Hz was still underestimated. Higher context values might be more suitable at low interruption rates and led to an improved prediction accuracy. Furthermore, model predictions showed a clear drop off at interruption rates of 8 and 16 Hz, possibly related to an overestimation of the effect of forward masking. The model accuracy increased when an alternative forward masking function was used.

In chapters 4 and 5, the cESTI-model was evaluated using monosyllabic *words*. However, the original motivation for adding context to the model was the observed prediction inaccuracy in *sentences* masked by noises with speech-like characteristics in chapter 3. Therefore, **chapter 6** discussed the performance of the cESTI-model using the sentence intelligibility data from chapter 3. Both the $cESTI_1$ and $cESTI_2$ were again evaluated. Prediction accuracy of the $cESTI_1$ remained similar or increased in comparison with the ESTI predictions. On the contrary, the accuracy of the $cESTI_2$-model decreased for non-stationary background noises without speech-like characteristics. However, $cESTI_2$ outperformed $cEST_1$ for speech-like noises, like a competing speaker. In general, inaccuracies with regard to speech-like noises remained. Adjusting the values of the context factors and the transfer function might lead to improvement, but it is likely that modulation masking and informational masking remain an important factor in the discrepancies between intelligibility and model predictions.

Compared to many existing audibility-based models, the suitability of the ESTI for use in reverberant conditions is an important advantage. When compared to other modulation-based models, the use of short time windows makes the ESTI-model better suited for use in interrupted noise. However, several of these models tend to perform better when speech is masked by noises with speech-like characteristics. When compared to models based on machine learning, the simplicity and easy usage of the ESTI are the main advantages.
An important aspect that was largely left unaddressed in the current work is the applicability of the model for sensorineurally hearing-impaired persons. The use of the individual tone audiogram can be useful, but this does not account for the temporal and spectral distortions that usually occur with this type of hearing loss. An additional individual distortion factor based on the hearing loss for speech in noise might improve the model further.
The ESTI proved to be a valuable extension of the classic STI for the use in non-stationary noises. With only a single recording of the background noise and an impulse response measurement, a reliable ESTI-value can be obtained. This value can then be used to predict speech intelligibility in a variety of non-stationary background noises. The addition of context improved the model further, but at the cost of higher complexity. It did show how a context model can be successfully used in combination with traditional methods for estimating speech intelligibility.

Samenvatting

De Speech Transmission Index of STI is een veelgebruikte maat die ontworpen is voor de evaluatie van de kwaliteit van spraak tussen een spreker en een luisteraar. De STI kan worden toegepast in de communicatietechnologie en zaalakoestiek om een transmissiekanaal te evalueren zonder spraakverstaanbaarheidsmetingen te hoeven uitvoeren voor een specifieke conditie. Voorbeelden van het gebruik zijn het ontwerpen en de constructie van auditoria en theaters, en de evaluatie van een werkomgeving wanneer er problemen optreden bij het verstaan van spraak.
De STI is een relatief eenvoudige maar robuuste index tussen 0 en 1 die is gebaseerd op de vermindering van modulaties in de spraak als gevolg van achtergrondruis en/of nagalm. Deze modulatiereductie blijkt gecorreleerd met de mate van spraakverstaan. De basis van de STI is de Modulatie Transfer Functie. Deze MTF wordt berekend op basis van opnames die traditioneel werden gemaakt met behulp van de *directe* meetmethode, door een gemoduleerde ruis als meetsignaal te gebruiken. Een andere benadering is de *indirecte* meetmethode, waarbij een separate opname van de achtergrondruis wordt gemaakt in combinatie met een meting van de impulsrespons. Wanneer de MTF bekend is, wordt deze omgerekend naar een STI-waarde in verschillende stappen. Een STI-waarde van 0 staat voor een situatie waar spraakverstaanbaarheid onmogelijk is. Een index van 0.75 of hoger staat voor omstandigheden waarbij goede tot excellente spraakverstaanbaarheid mogelijk is. Een index onder de 0.30 duidt op slechte tot zeer slechte spraakverstaanbaarheid. Merk op dat dit enkel van toepassing is op normaalhorende luisteraars in de eigen moedertaal. Wanneer een transferfunctie tussen de STI en de spraakverstaanbaarheid bekend is voor een bepaald spraakcorpus, kan de spraakverstaanbaarheid geschat worden met behulp van de STI-waarde.
Ondanks de robuustheid en grondige evaluatie van de STI zijn er verschillende zwakke punten. Wanneer non-lineaire vervormingen zoals compressie of spectrale subtractie plaatsvinden, gaat de nauwkeurigheid van de STI achteruit. Echter, het huidige onderzoek richtte zich op zaalakoestiek, waardoor alleen vervormingen die lineair van aard zijn, werden meegenomen. Een andere beperking van de STI is de bruikbaarheid wanneer achtergrondruis niet stationair is. Dit aspect was de belangrijkste aanleiding voor het huidige onderzoek. De achtergrond van deze beperking is tweeledig. Ten eerste is de traditionele, directe meetmethode met gemoduleerde ruis als meetsignaal gevoelig voor fluctuaties in de achtergrondruis. Dit kan leiden tot onder- of overschatting van de STI. Ten tweede neemt de spraakverstaanbaarheid bij normaalhorende luisteraars toe wanneer er dips in de ruis worden

geïntroduceerd. Deze winst wordt veroorzaakt door het vermogen om spraakfragmenten waar te nemen op de momenten dat de spraak het minst wordt beïnvloed door de ruis. De STI kan niet omgaan met de winst in het verstaan door deze fluctuaties in de ruis, waardoor de uitkomst niet representatief is voor de werkelijke omstandigheden. Het doel van het huidige werk was om de toepasbaarheid van de STI in niet-stationaire achtergrondruizen te verbeteren door zowel de meetmethode als de winst door fluctuaties in de ruis te onderzoeken.

In **hoofdstuk 2** werden de condities onderzocht waaronder de STI nauwkeurig gemeten kan worden. Hierbij lag de focus op de indirecte meetmethode, waarbij de MTF werd afgeleid van een meting van de impulsrespons en een langdurige opname van de achtergrondruis. Deze methode is minder gevoelig voor fluctuaties in de ruis, maar onder welke omstandigheden de metingen betrouwbaar uitgevoerd kunnen worden, was niet bekend. Om dit verder te onderzoeken werden twee experimenten uitgevoerd. Metingen van de impulsrespons (met een zogenaamd *sweep-signaal*) en ruisopnames werden uitgevoerd in een ruimte met variabele absorptie, verschillende niveaus van stationaire en fluctuerende ruis, en verschillende niveaus van het sweep-signaal. Om de experimentele bevindingen te kunnen extrapoleren naar andere akoestische condities, werd een groot aantal andere omstandigheden gesimuleerd. De experimenten en simulaties toonden aan dat de minimale impuls-ruisverhouding van +25 dB (overeenkomend met een sweep-ruisverhouding van –4 tot +15 dB) in niet-stationaire ruis benodigd was om de STI nauwkeurig te kunnen meten.

De Extended STI of ESTI werd geïntroduceerd in **hoofdstuk 3**. Het doel van het aangepaste model was om rekening te houden met de verbeterde spraakverstaanbaarheid wanneer er dips in de ruis aanwezig zijn. De belangrijkste aanpassing was de berekening van de STI per tijdsinterval van 2 ms, in plaats van voor het volledige signaal. De uiteindelijke ESTI-waarde was het gemiddelde van alle lokale STI-waarden. Om rekening te houden met abrupte fluctuaties in de ruis werd ook voorwaartse maskering (*forward masking*) toegevoegd aan het model. De fijnafstelling van de modelparameters werd gedaan op basis van nieuw uitgevoerde spraakverstaanbaarheidsmetingen bij normaalhorenden waarbij zinnen werden aangeboden. Door middel van een adaptieve procedure werd de signaal-ruisverhouding gemeten waarbij 50% van de zinnen werden verstaan (cSNR). De spraak werd vervormd door verschillende ruizen te gebruiken (stationaire ruis en twee soorten niet-stationaire ruis), dan wel nagalm toe te voegen in vijf gradaties. Evaluatie van het model vond plaats met behulp van data van 10 studies uit de bestaande literatuur over spraakverstaanbaarheid. De ESTI voorspelde de spraakverstaanbaarheid beter dan de klassieke

STI in alle soorten niet-stationaire ruis die werden gebruikt. Onnauwkeurigheden in de voorspellingen werden geobserveerd wanneer achtergrondruis spraakachtige eigenschappen vertoonde. Wanneer enkel de omhullende karakteristieken van spraak vertoonde, was het verschil tussen de waargenomen en voorspelde cSNR ongeveer 3 – 4 dB. Wanneer de fijnstructuur van de ruis ook op spraak leek, werd er een extra verschil van 5 – 7 dB gevonden. De hypothese van de auteurs was dat deze onnauwkeurigheden het resultaat waren van modulatiemaskering, *informational masking* en/of contexteffecten.

In **hoofdstuk 4** werd de hypothese getest dat deze onnauwkeurigheden veroorzaakt werden door contexteffecten. Bij het waarnemen van spraakfragmenten in niet-stationaire ruis met hoge modulatiefrequenties heeft de luisteraar toegang tot delen van alle spraakelementen. Deze elementen kunnen fonemen als deel van een woord zijn, maar ook woorden als deel van een zin. Wanneer de modulaties in de ruis traag zijn, neemt de waarschijnlijkheid toe dat elementen in de spraak volledig gemaskeerd worden door de langere ruisfragmenten en daardoor niet verstaanbaar zijn. De luisteraar kan in dat geval terugvallen op contextuele informatie om het gemiste spraakelement te "raden". Om dit mechanisme bij het verstaan mee te nemen, werd context toegevoegd aan het ESTI-model. Eerst werd de ESTI per spraakelement berekend in plaats van voor het gehele signaal. Daarna werd een transferfunctie geschat om de ESTI per element te koppelen aan de elementscore in isolatie. Wanneer deze score bekend was, kon de verstaanbaarheid van de gehele uiting geschat worden met behulp van een context model. Om de prestatie te evalueren van dit op context gebaseerde ESTI-model (cESTI-model) werd bestaande spraakverstaanbaarheidsdata van betekenisvolle, monosyllabische woorden in onderbroken ruis geanalyseerd. De toevoeging van twee verschillende contextmodellen werd vergeleken, aangeduid als $cESTI_1$ (met het Bronkhorst contextmodel) en $cESTI_2$ (met het Boothroyd en Nittrouer contextmodel). De nauwkeurigheid van de voorspellingen van het nieuwe model verbeterde aanzienlijk voor interruptiefrequenties lager dan 5 Hz. De prestaties van de twee modelversies $cESTI_1$ en $cESTI_2$ waren vergelijkbaar.

Het cESTI-model werd geëvalueerd in **hoofdstuk 5** met behulp van nieuw gemeten CVC-woorden in stationaire en onderbroken ruis. Zowel nonsens als betekenisvolle woorden werden gebruikt. Alleen $cESTI_2$ (met het eenvoudigere van de twee contextmodellen uit hoofdstuk 4) werd gebruikt in dit hoofdstuk. Het model presteerde beter bij zowel nonsens als betekenisvolle woorden. Echter, ondanks de verbeterde prestaties werd de verstaanbaarheid van betekenisvolle woorden bij interruptiefrequenties onder de 5 Hz nog steeds onderschat. Hogere contextwaarden bleken mogelijk meer geschikt te zijn bij lage interruptiefrequenties en leidden tot een verbeterde nauwkeurigheid van

de voorspellingen. Verder lieten modelvoorspellingen een duidelijke afname zien bij interruptiefrequenties van 8 en 16 Hz, mogelijk gerelateerd aan een overschatting van het effect van voorwaartse maskering. De nauwkeurigheid van het model nam toe bij het gebruik van een alternatieve functie voor de voorwaartse maskering.

In de hoofdstukken 4 en 5 werd het cESTI-model uitsluitend geëvalueerd met behulp van monosyllabische *woorden*. Echter, de oorspronkelijke reden voor het toevoegen van context was de onnauwkeurigheid van de voorspellingen van *zinnen* die gemaskeerd werden door ruizen met spraakachtige kenmerken in hoofdstuk 3. Om deze reden werd in **hoofdstuk 6** de prestatie van het cESTI-model onderzocht op basis van de spraakverstaanbaarheidsdata van zinnen uit hoofdstuk 3. Zowel de $cESTI_1$ als de $cESTI_2$ werden opnieuw geëvalueerd. De nauwkeurigheid van de voorspellingen van $cESTI_1$ bleef vergelijkbaar of nam toe in vergelijking met de ESTI-voorspellingen. Daarentegen nam de nauwkeurigheid van het $cESTI_2$-model juist af voor niet-stationaire achtergrondruizen zonder spraakachtige kenmerken. Echter, $cESTI_2$ presteerde beter dan $cESTI_1$ bij spraakachtige ruizen, zoals bijvoorbeeld een andere spreker. In het algemeen bleven de voorspellingen met betrekking tot spraakachtige ruizen relatief onnauwkeurig. Het aanpassen van de contextwaarden en de transferfuncties kunnen mogelijk tot een verbetering leiden, maar het is waarschijnlijk dat modulatiemaskering en informational masking een belangrijke rol blijven spelen in de discrepantie tussen spraakverstaanbaarheid en de modelvoorspellingen.

In vergelijking met veel modellen gebaseerd op hoorbaarheid, is de toepasbaarheid van de ESTI in condities met nagalm een belangrijk voordeel. Vergeleken met andere modellen gebaseerd op spraakmodulaties, maakt het gebruik van korte tijdsintervallen de ESTI beter geschikt voor onderbroken ruis. Echter, een aantal van deze modellen lijkt beter te presteren wanneer spraak gemaskeerd wordt door ruizen met spraakachtige kenmerken. Wanneer de ESTI vergeleken wordt met modellen gebaseerd op machine learning, zijn de eenvoud en het gebruiksgemak van de ESTI de belangrijkste voordelen.

Een belangrijk aspect dat nauwelijks genoemd werd in het huidige werk is de toepasbaarheid van het model bij mensen met een perceptief gehoorverlies. Het gebruik van het individuele toonaudiogram zou bruikbaar kunnen zijn voor deze groep, maar deze benadering houdt geen rekening met de temporele en spectrale vervorming die normaliter optreedt bij dit type gehoorverlies. Het toevoegen van een vervormingsfactor op basis van het gehoorverlies voor spraak in ruis kan het model mogelijk verder verbeteren.

De ESTI is een waardevolle toevoeging op de klassieke STI voor het gebruik in niet-stationaire ruizen. Met een enkele opname van de achtergrondruis en een

impulsresponsmeting kan een betrouwbare ESTI worden berekend. Dit resultaat kan vervolgens worden ingezet om de spraakverstaanbaarheid te voorspellen in een verscheidenheid aan niet-stationaire achtergrondruizen. Het toevoegen van context verbeterde het model verder, maar wel ten koste van een hogere complexiteit. Het liet wel zien hoe een contextmodel succesvol gebruikt kan worden in combinatie met traditionele methodes voor het voorspellen van spraakverstaan.

Dankwoord

Daar ligt ie dan! Een kleine 15 jaar nadat ik in oktober 2008 voor het eerst aan een promotietraject begon bij Klinische en Experimentele Audiologie op het AMC. Wie had dat ooit gedacht? Ikzelf heb het nogal eens moeilijk gehad met het vertrouwen in een goed eindresultaat. En Wouter, daarvoor wil jou in het bijzonder bedanken: voor het tomeloze vertrouwen in mijn kunnen en de kansen die je me hebt geboden om tot dit resultaat te komen. Dat was lang niet altijd vanzelfsprekend, en daardoor ben ik extra blij dat jij de juiste omstandigheden hebt gecreëerd die tot dit boekje hebben geleid.

Koen, jouw werk over de ESII en je eerste ideeën over de extensie van de STI waren het begin van dit proefschrift. Ik waardeer jouw visie en vaardigheid om, ongeacht de omstandigheden, je focus op het eindresultaat te houden. Dat heeft me zeker tijdens de mindere momenten erg geholpen om door te zetten. En ik heb daarnaast altijd genoten van onze gemeenschappelijke muzikale interesses en concertbezoeken.

Verder gaat mijn grote dank uit naar het Hensius-Houbolt Fonds voor de mede-financiering van dit onderzoek. Zonder deze steun voor mijn project was dit proefschrift er nooit geweest. Ook wil ik de leden van mijn promotiecommissie hartelijk danken voor het kritisch doorlezen en beoordelen van het manuscript.

Lieve Tessa en Margriet, wat ontzettend tof dat jullie naast me staan op 9 november! Tess, in goede en slechte tijden ben je er voor me. Als klankbord, als steun of gewoon voor de gezelligheid. En wat er ook speelt, of hoe druk je ook bent, die ruimte is er altijd. Margriet, je bent me al voorgegaan en nu zijn de rollen omgedraaid. Sinds 2002 zijn we redelijk gelijk opgetrokken met de studie, het vakgebied en het onderzoek. Je spontaniteit en je eigengereidheid werkten altijd verfrissend. Het is gek dat onze wegen op werkgebied nu gaan scheiden, maar gelukkig zien we elkaar daarbuiten nog regelmatig.

In de loop der jaren heb ik flink wat kamer- en afdelingsgenoten versleten. De laatste periode was het proces extra pittig door het thuiswerken en daardoor het gebrek aan gezellige gesprekken en lekkere koffie. Dus heel veel dank aan Bastiaan (voor je onorthodoxe en creatieve kijk op zo'n beetje alles), Femke (voor je – soms onbedoeld grappige – rake opmerkingen), Hiske (voor je betrokkenheid), Ilja (voor je aanstekelijke enthousiasme), Inge (voor de koffie en de koekjes), László (voor je ongekende technische kennis), Maaike (voor gewoon zeggen waar het op staat), Marjolijn (voor je spontaniteit en je erfenis.

Immers, MarjoProggie is nog steeds in gebruik!), Marya (voor je prachtige voorbeeld van doorzettingsvermogen), Mirjam (voor je nuchtere gezelligheid), Monique B. (voor je luisterend oor en fijne samenwerking), Monique L. (voor je positiviteit en optimisme), Rolph (voor het wegwijs maken in de audiologie), Simon (voor het delen van de smart), Thamar (voor je kritische blik en goede adviezen) en Thijs (voor je humor). Verder dank aan al mijn klinische collega's op het AMC, bij Pento Utrecht en in het UMC Utrecht. Niet alleen voor de betrokkenheid bij en interesse in mijn onderzoek, maar ook voor het creëren van de omstandigheden die zorgden dat ik al die jaren aan mijn onderzoek kon blijven werken.

Lieve vrienden, dank voor alle afleiding, avondjes, trips, tips, gesprekken, wandelingen en etentjes. Xounianen, dankzij onze jaarlijkse Ardennen-trip voel zelfs ík me steeds weer even 18.

Lieve Aam, lieve Bart, door de jaren heen heb ik voor veel keuzes gestaan. Grote en kleine, sommige meer succesvol dan andere. Bij ieder van die keuzes heb ik altijd jullie betrokkenheid en steun ervaren, zowel op de voorgrond als op de achtergrond. Zo ook gedurende dit lange proces.

En tot slot. Lieve Loes, zonder jouw steun was ik nooit zover gekomen. Dit is van toepassing op veel meer dan alleen dit boekje. Lieve Teun en Guus, jullie plaatsen alles in het juiste perspectief.

Curriculum vitae

Jelmer van Schoonhoven was born on May 18th 1981 in Makkum. He moved to Huizen at an early age and graduated in 1999 from the OSG de Huizermaat. After a year in Australia, he started his study at the University of Twente in 2001, where he received his Masters' degree in Biomedical Engineering on October 9th 2008. The subject of his thesis was The Influence of Vagus Nerve Stimulation on the Interictal EEG. After his study he moved to Amsterdam to start at a research position in the department of Clinical and Experimental Audiology at the Academic Medical Centre. After a brief hiatus he started as a Medical Physics Expert – Audiologist in training in 2011, also in the AMC. During this period, he contributed to research regarding bilateral use of hearing aids and of cochlear implants. During a research project as part of his training, he started working with Dr. K.S. Rhebergen on the first experiments with regard to the measurement of the STI in non-stationary noise. These experiments laid the basis for the current thesis. Since 2015, he has been working elsewhere, but continued his research on the ESTI under the supervision of Dr. K.S. Rhebergen and Prof. dr. ir. W.A. Dreschler.

PhD portfolio

Name PhD student: ir. J. van Schoonhoven
PhD period: 2011–2023
Name PhD supervisor: Prof.dr. ir. W.A. Dreschler
Name PhD co-supervisor: Dr. K.S. Rhebergen

PhD training

	Year	ECTS
General courses		
BROK (Basiscursus Regelgeving Klinisch Onderzoek)	2013	1
Statistics course	2014	2
Scientific writing in English	2013	1
Management course	2014	0.5
Specific courses		
Signal Processing in Acoustics and Audio	2011	5
Acoustics course	2012	2
Evoked Response Audiometry course	2012	2
Seminars, workshops and master classes		
Weekly research meetings	2011–2016	5
Annual KKAu meetings	2011–2022	3
Annual NVKF meetings	2011–2022	1
Biannual NVA meetings	2011–2022	3
Biannual meeting audiologist trainees	2011–2015	2
Hearing and Genes symposium	2012	0.3
Hearing aid workshop	2012	0.5
Hearing aid workshop	2013	0.5
PACT day (Platform for Audiological Clinical Testing)	2013	0.3
Symposium Sig Soli	2014	0.3
Symposium Vera Prijs	2014	0.3
Presentations		
A Test Battery To Assess The Benefits Of Bilateral Amplification With Hearing Aids part I (oral); *DGA conference, Germany*	2012	1
The Effectiveness of Bilateral Cochlear Implants for Severe to Profound Deafness in Children and Adults (oral); *Objective Measures, Amsterdam*	2012	1
The Extended Speech Transmission Index, to predict the speech intelligibility in fluctuating and reverberant background noise (oral); *International Hearing Aid Research Conference, United States*	2014	1

	Year	ECTS
Accuracy of STI measurements in fluctuating noise using the impulse response (poster); *International Hearing Aid Research Conference, United States*	2014	0.5
The Extended Speech Transmission Index, to predict the speech intelligibility in fluctuating and reverberant background noise (oral); *Meeting Werkgroep Auditief Systeem (WAS), Groningen*	2015	0.5
The Extended Speech Transmission Index, to predict the speech intelligibility in fluctuating and reverberant background noise (oral); *Meeting Dutch Society of Audiology (NVA), Utrecht*	2015	0.5
Context Effects, Modelling Speech Recognition in Interrupted Speech and Noise (oral); *Meeting Werkgroep Auditief Systeem (WAS), Leiden*	2018	1
Speech intelligibility and context effects, The Extended Speech Transmission Index (oral), *Conference Audiological Research Cores in Europe (ARCHES) France*	2019	1
(Inter)national conferences		
DGA conference, Germany	2012	1
Objective Measures congress, the Netherlands	2012	1.5
Pediatric hearing aid conference, United States	2013	1
International Hearing Aid Research Conference, United States	2014	1.5
Speech in Noise conference, the Netherlands	2016	0.7
Pediatric hearing aid conference, United States	2016	1
Congress on Bone Conduction Hearing, the Netherlands	2017	0.3
Conference Audiological Research Cores in Europe (ARCHES), France	2019	0.5

Teaching

	Year	ECTS
Lecturing		
Bi-monthly lecture for medical students on audiology	2011–2015	5
Invited lecture on insertion gain	2012	0.3
Tutoring, Mentoring		
Tutoring audiologists in training	2018–2022	3
Other		
Co-authorship of chapter in the online Dutch audiology textbook (adiologieboek.nl)	2015	1

Publications

Peer reviewed

The effectiveness of bilateral cochlear implants for severe-to-profound deafness in children: a systematic review. *Sparreboom, M., van Schoonhoven, J., van Zanten, B. G., Scholten, R. J., Mylanus, E. A., Grolman, W., Maat, B.* Otol Neurotol. 2010. 31(7): 1062-1071

The effectiveness of bilateral cochlear implants for severe-to-profound deafness in adults: a systematic review. *van Schoonhoven, J., Sparreboom, M., van Zanten, B. G., Scholten, R. J., Mylanus, E. A., Dreschler, W. A., Grolman, W., Maat, B.* Otol Neurotol. 2013. 34(2): 190-198.

Selecting Appropriate Tests to Assess the Benefits of Bilateral Amplification With Hearing Aids. *van Schoonhoven, J., Schulte, M., Boymans, M., Wagener, K. C., Dreschler, W. A., Kollmeier, B.* Trends Hear. 2016. 20: 1-16

Towards measuring the Speech Transmission Index in fluctuating noise: Accuracy and limitations. v*an Schoonhoven, J., Rhebergen, K. S., Dreschler, W. A.* J Acoust Soc Am. 2017. 141(2): 818-827.

The Extended Speech Transmission Index: Predicting speech intelligibility in fluctuating noise and reverberant rooms. *van Schoonhoven, J., Rhebergen, K. S., Dreschler, W. A.* J Acoust Soc Am. 2019. 145(3): 1178-1194.

A context-based approach to predict speech intelligibility in interrupted noise: Model design. *van Schoonhoven, J., Rhebergen, K. S., Dreschler, W. A.* J Acoust Soc Am. 2022. 151(2): 1404-1415.

Appendices

Appendix A: List of abbreviations

Table A-1: All general abbreviation that are used in the current work, except for the speech intelligibility models. These are described in Table A-2.

α	Parameter of the transfer function between the $ESTI$ or STF and q_e. This parameter primarily influences the point of intersection with the x-axis [see Eqs. (4-9) and (5-3)]
ASR	Automatic Speech Recognition
β	Parameter of the transfer function between the $ESTI$ or STF and q_e. This parameter primarily influences the slope of the curve and the intersection with the x-axis [see Eqs. (4-9) and (5-3)]
γ	Parameter of the transfer function between the $ESTI$ or STF and q_e. This parameter primarily influences the intersection with right vertical axis [see Eqs. (4-9) and (5-3)]
CE	Context Effects
c_i	Context factors according to Bronkhorst *et al.* (1993), which represent the probability of correctly guessing one of the elements that were missed in the sensory stage. See Eqs. (4-6), (4-7), and (C-1) – (C-5)
cSNR	Critical SNR (SNR at 50% speech intelligibility)
DC	Duty cycle of interrupted speech or noise. In practice, this is the similar to the STF. However, DC is a global property of the interrupted speech, whereas STF is the local fraction of speech perceived
EDT	Early Decay Time
EM	Energetic Masking
ER	Envelope Regression
F_{int}	Interruption frequency of interrupted speech or interrupted noise
FMB	Fluctuating Masker Benefit
IIR	Infinite Impulse Response
IM	Informational Masking
IN8	Interrupted Noise with an interruption rate of 8 Hz
INR	Impulse-to-Noise Ratio
ISTS	International Speech Test Signal (Holube *et al.*, 2010)
j	Context factor according to Boothroyd and Nittrouer (1988), which represents the number of statistically independent parts in a whole. Mathematically, it is the ratio between the log probability of a whole (p_w) and an element (p_e), both in context. See Eqs. (4-3) and (5-2)
k	Context factor according to Boothroyd and Nittrouer (1988), which represents the proportional increase in number of channels of statistically independent information as a result of context. Mathematically it is the ratio between the log error probabilities of an element in context (1-p_e) and without context (1-q_e). See Eqs. (4-2) and (5-1)

Table A-1: Continued.

MLS	Maximum Length Sequence
MM	Modulation Masking
MT	Masked Threshold. Parameter used in the forward masking function in Eqs. (3-3) and (5-5)
MTF	Modulation Transfer Function. See Eq. (3-7)
MTF_{rev}	Part of the Modulation Transfer Function influenced by reverberation. See Eq. (3-5)
MTF_{SNR}	Part of the Modulation Transfer Function influenced by noise. See Eq. (3-6)
MTI	Modulation transfer index. See Eq. (3-9)
NCM	Normalized Covariance Method
p_c	Consonant score in context
p_e	Element score in context. In the current work always of a phoneme as part of a word
p_v	Vowel score in context
p_w	Score of a whole. In the current work always of a word
P_w	Score of a whole. In the current work always of a sentence
$p_{w,n}$	n-phoneme score (when n equals the number of elements, $p_{w,n} = p_w$)
q_c	Consonant score in isolation
q_e	Element score in isolation. In the current work always of a phoneme as part of a word
Q_e	Element score in isolation. In the current work always of a word as part of a sentence
Q_i	The total probability of missing i elements without context in the sensory stage of the model of Bronkhorst *et al.* (1993) as described in Eqs. (4-4), (4-5), and (C-6) – (C-9)
q_v	Vowel score in isolation
rms	Root mean square
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
SNR	Signal-to-Noise Ratio
SNR_{env}	Speech-to-noise envelope power ratio, as used in the EPSM and related models.
SNR_{mod}	Signal-to-Noise Ratio in the modulation domain (Dubbelboer and Houtgast, 2008)
SSN	Stationary Speech-shaped Noise
STF	Fraction of speech perceived. In practice this is the similar to DC. However, DC is a global property of the interrupted speech, whereas STF is the local fraction of speech perceived

Table A-2: All speech intelligibility models that are referred to in the current work.

AI	Articulation index, later known as the SII (French and Steinberg, 1947; Fletcher and Galt, 1950; Kryter, 1962; ANSI-S3.5, 1969)
BSIM	Binaural Speech Intelligibility Index, based on the EC/SII (Beutelmann *et al.*, 2009)
cESTI	Method to predict speech intelligibility in interrupted noise using the ESTI combined with the Boothroyd and Nittrouer context model (Van Schoonhoven *et al.*, 2022)
$cESTI_1$	Method to predict speech intelligibility in interrupted noise using the ESTI combined with the Bronkhorst context model (Van Schoonhoven *et al.*, 2022)
$cESTI_2$	Identical to cESTI, but with the subscript 2 when used in combination with $cESTI_1$
$cSTF_1$	Precursor of the $cESTI_1$, using the STF instead of the ESTI in combination with the Bronkhorst context model to model interrupted speech (Van Schoonhoven *et al.*, 2022).
$cSTF_2$	Precursor of the $cESTI_2$, using the STF instead of the ESTI in combination with the Boothroyd and Nittrouer context model to model interrupted speech (Van Schoonhoven *et al.*, 2022)
CSTI	Covariance-based STI (Ludvigsen *et al.*, 1990; Holube and Kollmeier, 1996; Goldsworthy and Greenberg, 2004)
EC/SII	Binaural speech intelligibility model using Equalization Cancellation in combination with the SII (Beutelmann and Brand, 2006)
EPSM	Envelope-Power Spectrum Model (Dau *et al.*, 1999; Ewert and Dau, 2000)
ESII	Extended Speech Intelligibility Index (Rhebergen and Versfeld, 2005; Rhebergen *et al.*, 2006)
ESIIsen	ESII using sentences as input (Meyer and Brand, 2013)
eSTI	Extended Speech Transmission Index (Prodi and Visentin, 2019)
ESTI	Extended Speech Transmission Index (Van Schoonhoven *et al.*, 2019)
ESTOI	Extended Short-Time Objective Intelligibility measure (Jensen and Taal, 2016)
FADE	Framework for Auditory Discrimination Experiments (Schädler *et al.*, 2015; Schädler *et al.*, 2016)
GPSM	General power spectrum model (Biberger and Ewert, 2016; 2017)
mr-sEPSM	Multi-resolution speech-based Envelope-Power Spectrum Model (Jørgensen *et al.*, 2013)
NCM	Normalized Covariance Method (Holube and Kollmeier, 1996; Goldsworthy and Greenberg, 2004)
QSTI	Quasi-stationary STI (Schwerin and Paliwal, 2014)
RASTI	Rapid STI, using nine modulation frequencies in two octave bands (now obsolete)

Table A-2: Continued.

sEPSM	Speech-based Envelope-Power Spectrum Model (Jørgensen and Dau, 2011)
SII	Speech intelligibility Index, formerly known as the AI (Pavlovic, 1984; Pavlovic, 1987; ANSI-S3.5, 1997)
stBSIM	short-time Binaural Speech Intelligibility Model (Beutelmann *et al.*, 2010)
STI	Speech transmission index (Houtgast and Steeneken, 1985; Houtgast, 1989; Steeneken and Houtgast, 2002; IEC60268-16, 2011)
STIPA	Faster method of measuring the STI by applying two unique modulation frequencies to each of the seven octave bands
STGI	Spectro-temporal Glimpsing Index (Edraki *et al.*, 2022)
STMI	Spectro-temporal Modulation Index (Chi *et al.*, 1999)
STOI	Short-Time Objective Intelligibility measure (Taal *et al.*, 2011)

Appendix B: ESTI predictions

Table B-1: Observed cSNR-values (cSNR obs) and predicted cSNR-values (cSNR pred) of various studies using Dutch speech material. Per study, the ESTI-value in stationary noise without reverberation is used to predict the cSNR in fluctuating noises and/or reverberant conditions. *, George et al. (2008) did not test with SSN. Therefore, the SSN results from Rhebergen et al. (2010) and George et al. (2006) served as reference conditions. VU98: Sentence material by Versfeld et al. (2000); DC: Duty Cycle, followed by the percentage on-time of the noise; MD: Modulation Depth, followed by the MD of the noise interruptions in dB; ICRA (International Collegium of Rehabilitative Audiology): Noises as described by Dreschler et al. (2001); LIST: Leuven intelligibility sentence test (Van Wieringen and Wouters, 2008); HF: High Frequency; LF: Low Frequency.

Study	Speech	Noise	T_{60} (s)	cSNR obs (dB)	ESTI	cSNR pred (dB)
Data current study	VU98 female	Speech-shaped stationary noise (VU98 female)	0.0	-3.4	0.4115	*ref*
			0.1	-2.5	0.4162	-2.6
			0.4	1.2	0.4391	0.1
			0.8	5.3	0.4370	3.7
			1.2	7.2	0.4021	8.2
		Interrupted noise (8 Hz. DC50)	0.0	-15.0	0.4191	-15.6
			0.1	-6.2	0.4404	-7.3
			0.4	1.5	0.4593	-0.5
			0.8	5.8	0.4458	3.6
			1.2	10.1	0.4266	8.2
		ISTS	0.0	-7.0	0.6754	-18.6
			0.1	-3.1	0.6513	-14.8
			0.4	0.8	0.5577	-8.2
			0.8	8.6	0.5057	-1.0
			1.2	12.0	0.4442	5.7

Table B-1: Continued.

Study	Speech	Noise	T_{60} (s)	cSNR obs (dB)	ESTI	cSNR pred (dB)
Rhebergen et al. (2006)	VU98 female	Speech-shaped stationary noise (VU98 female)	0.0	-5.5	0.3419	*ref*
		Interrupted noise (4 Hz DC50)		-14.9	0.4640	-26.9
		Interrupted noise (8 Hz DC40)		-22.8	0.3972	-25.8
		Interrupted noise (8 Hz DC45)		-20.8	0.3848	-23.4
		Interrupted noise (8 Hz DC50)		-17.6	0.3848	-20.7
		Interrupted noise (8 Hz DC55)		-14.4	0.3820	-17.3
		Interrupted noise (8 Hz DC60)		-11.7	0.3886	-14.1
		Interrupted noise (16 Hz DC50)		-15.0	0.3377	-14.8
		Interrupted noise (32 Hz DC50)		-11.1	0.3331	-10.8
		Interrupted noise (64 Hz DC50)		-7.5	0.3574	-8.0
		Interrupted noise (128 Hz DC50)		-5.4	0.3654	-6.1
		Sinusoidal Intensity Modulated noise (8 Hz)		-5.8	0.4164	-8.1
		Sinusoidal Intensity Modulated noise (16 Hz)		-7.1	0.3693	-7.9
		Sinusoidal Intensity Modulated noise (32 Hz)		-6.5	0.3747	-7.5
		Sinusoidal Intensity Modulated noise (64 Hz)		-6.2	0.3643	-6.9
		Sinusoidal Intensity Modulated noise (128 Hz)		-5.3	0.3687	-6.1
		Sawtooth noise (8 Hz increasing)		-9.3	0.4018	-11.3
		Sawtooth noise (8 Hz decreasing)		-11.9	0.3728	-13.0
	VU98 male	Speech-shaped stationary noise (VU98 male)	0.0	-4.1	0.3905	*ref*
		Multitalker babble		-0.1	0.5529	-5.1
		ICRA5 fluctuating noise (250 ms gap length)		-8.4	0.5789	-16.0
		ICRA Dutch speaker noise (100 ms gap length)		-6.7	0.5823	-13.9
		Swedish speaker		-6.8	0.6564	-18.0
		Dutch speaker		-3.6	0.7071	-16.4
		ISTS		-6.4	0.6134	-15.1
Francart et al. (2011)	VU98 female	Speech-shaped stationary noise (VU98 female)		-3.6	0.4042	*ref*
		Multitalker babble		-1.1	0.5883	-6.9
		ICRA5 fluctuating noise (250 ms gap length)		-11.2	0.5834	-18.7
		ICRA Dutch speaker noise (100 ms gap length)		-10.2	0.5662	-16.8
		Swedish speaker		-11.6	0.6099	-20.3
		Dutch speaker		-8.8	0.6418	-18.4
		ISTS		-5.5	0.6885	-17.0
	LIST female	Speech-shaped stationary noise (LIST female)		-8.4	0.2463	*ref*
		Multitalker babble		-1.8	0.4771	-9.9
		ICRA5 fluctuating noise (250 ms gap length)		-14.2	0.4480	-23.2
		ICRA Dutch speaker noise (100 ms gap length)		-13.1	0.4279	-21.1
		Swedish speaker		-15.2	0.4618	-24.8
		Dutch speaker		-13.4	0.4651	-23.0
		ISTS		-9.0	0.5407	-21.2
Festen and Plomp (1990)	Plomp female	Speech-shaped stationary noise (Plomp female)	0.0	-4.4	0.3808	*ref*
		Speech-shaped stationary noise (Plomp male)		-7.0	0.3255	-5.3
		Speech modulated noise (Plomp female)		-10.6	0.4782	-14.8
		Speech modulated noise (Plomp male)		-12.8	0.4720	-16.4
	Plomp male	Speech-shaped stationary noise (Plomp male)		-4.1	0.3955	*ref*
		Speech-shaped stationary noise (Plomp female)		-3.2	0.4259	-4.1
		Speech modulated noise (Plomp male)		-7.2	0.5844	-14.2
		Speech modulated noise (Plomp female)		-7.4	0.5575	-14.3
Rhebergen et al. (2014)	VU98 female	Speech-shaped stationary noise (VU98 female)	0.0	-2.8	0.4323	*ref*
		Interrupted noise (8 Hz DC50)		-24.3	0.3981	-19.1
		Interrupted noise (8 Hz DC25)		-39.7	0.4055	-38.4
Rhebergen et al. (2010)	Plomp female	Speech-shaped stationary noise (Plomp female)		-4.8	0.3662	*ref*
				-4.5	0.3772	-4.8
				-4.3	0.3842	-4.8

Table B-1: Continued.

Study	Speech	Noise	T_{60} (s)	cSNR obs (dB)	ESTI	cSNR pred (dB)
Rhebergen et al. (2010)		Interrupted noise (10 Hz DC50)		-20.6	0.2630	-14.2
				-23.6	0.3050	-18.9
				-28.3	0.3195	-24.2
George et al. (2006)	VU98 male	Speech-shaped stationary noise (VU98 male)	0.0	-4.1	0.3918	*ref*
		Speech modulated noise (VU98 male)		-11.7	0.4148	-12.6
		Interrupted noise (16 Hz DC50)		-17.3	0.2802	-11.7
		Interrupted noise (16 Hz DC75)		-6.4	0.3907	-6.4
		Interrupted noise (16 Hz DC50 MD15)		-8.8	0.3660	-8.0
		Interrupted noise (32 Hz DC50)		-12.7	0.2781	-8.7
*George et al. (2008)**	VU98 male	Speech modulated noise (VU98 male)	0.1	-7.8	0.4910	-11.7
			0.3	-2.5	0.4810	-6.4
			0.5	1.6	0.4724	-2.6
			1.0	5.1	0.4214	2.8
	Plomp female	Speech modulated noise (Plomp female)	0.1	-15.4	0.3355	-14.0
			0.3	-9.0	0.3423	-7.9
			0.5	-5.3	0.3370	-3.9
			1.0	-0.6	0.3346	1.4
Van Esch et al. (2013)	VU98 female	ICRA1 stationary noise	0.0	-5.0	0.3929	*ref*
		ICRA4 fluctuating noise (250ms gap length)		-12.8	0.4941	-16.7
Rhebergen et al. (2008)	VU98 female	Speech-shaped stationary noise (VU98 female)	0.0	-6.5	0.3088	*ref*
		Crowd		-6.0	0.4416	-10.5
		Car noise		-6.3	0.4529	-11.1
		Construction		-7.5	0.3961	-10.6
		Music		-11.5	0.3014	-11.2
		Frogs and insects		-12.5	0.3016	-12.2
		Speech modulated noise (VU98 female)		-12.5	0.3858	-15.9
Maalderink et al. (2011)	VU98 female	Forward speech (Plomp male)		-17.5	0.4330	-22.9
		Backward speech (Plomp male)		-17.0	0.4211	-21.9
		Ramming piles		-16.5	0.4413	-21.5
		Hens and birds		-12.8	0.3757	-15.2
		Machine gun		-24.5	0.5229	-33.7
		Speech-shaped stationary noise (VU98 female)	0.0	-3.4	0.4083	*ref*
				-2.8	0.4208	*ref*
		Speech-shaped stationary noise (LF gaps)		-3.5	0.3924	-3.0
				-2.9	0.4049	-2.4
		Speech-shaped stationary noise (HF gaps)		-7.0	0.3606	-5.5
				-6.2	0.3759	-4.8
		Interrupted noise (8 Hz DC50)		-12.1	0.3576	-10.3
				-8.2	0.3643	-6.5
		Interrupted noise (LF gaps)		-10.4	0.4014	-10.1
				-6.0	0.4290	-6.2
		Interrupted noise (HF gaps)		-14.1	0.3426	-11.8
				-9.4	0.3727	-7.9

Appendix C: Equations Bronkhorst context model

Model equations for CVC-words (Bronkhorst *et al.*, 1993) where recognition probabilities of the initial and final consonants are assumed equal (Bosman, 1989).

$$p_{w,3} = Q_0 + c_1Q_1 + c_1c_2Q_2 + c_1c_2c_3Q_3 \quad \text{(C-1)}$$

$$p_{w,2} = (Q_1 + 2c_2Q_2 + 3c_2c_3Q_3)(1 - c_1) \quad \text{(C-2)}$$

$$p_{w,1} = (Q_2 + 3c_3Q_3)(1 - 2c_2 - c_1c_2) \quad \text{(C-3)}$$

$$p_{w,0} = Q_3(1 - 3c_3 - 3c_2c_3 - c_1c_2c_3) \quad \text{(C-4)}$$

$$p_e = \frac{1}{3}q_v + \frac{2}{3}q_c + \frac{1}{3}c_1Q_1 + \frac{2}{3}c_2Q_2 + c_3Q_3 \quad \text{(C-5)}$$

with

$$Q_0 = q_c^2q_v \quad \text{(C-6)}$$

$$Q_1 = q_c^2(1 - q_v) + 2q_cq_v(1 - q_c) \quad \text{(C-7)}$$

$$Q_2 = q_v(1 - q_c)^2 + 2q_c(1 - q_c)(1 - q_v) \quad \text{(C-8)}$$

$$Q_3 = (1 - q_c)^2(1 - q_v) \quad \text{(C-9)}$$

Appendix D: Data by Miller and Licklider

Table D-1: Interrupted speech data by Miller and Licklider (1950). Word scores are depicted and were read from Fig. 4 in the original paper. NA represents values that were not measured. F represents the interruption frequency of the speech and DC represents the duty cycle.

F (Hz)	DC=12.5%	DC=25%	DC=50%	DC=75%
0.1	NA	NA	0.47	NA
0.22	NA	NA	0.44	NA
0.46	NA	NA	NA	NA
1.0	0.03	0.09	0.44	0.81
2.2	NA	NA	0.62	NA
4.6	NA	NA	0.84	NA
10	NA	0.67	0.84	0.94
22	0.24	0.66	0.89	NA
46	0.08	0.66	0.90	NA
$1.0x10^2$	0.05	0.64	NA	0.96
$2.2x10^2$	0.02	0.32	0.80	NA
$4.6x10^2$	0.02	0.46	0.70	NA
$1.0x10^3$	0.20	0.47	0.74	0.88
$2.2x10^3$	0.51	0.74	0.87	NA
$4.6x10^3$	0.83	0.90	0.96	NA

Table D-2: Interrupted noise data by Miller and Licklider (1950). Word scores are depicted and were read from Fig. 8 in the original paper. NA represents values that were not measured. ***F*** represents the interruption frequency of the speech and SNR represents the long-term signal to noise ratio. Note that the SNR is 3 dB higher than in the original paper, since Miller and Licklider documented the local SNR during the on-time of the noise.

F **(Hz)**	**SNR=-15 dB**	**SNR=-6 dB**	**SNR=+3 dB**	**SNR=+12 dB**
0.1	0.49	0.57	0.79	0.90
0.22	0.43	0.61	0.78	0.93
0.46	NA	NA	NA	NA
1.0	0.60	0.67	0.80	0.90
2.2	0.66	0.80	0.86	0.92
4.6	0.72	0.85	0.90	0.96
10	0.72	0.78	0.93	0.95
22	0.69	0.78	0.90	0.92
46	0.34	0.70	0.86	0.92
$1.0x10^2$	0.03	0.36	0.76	0.92
$2.2x10^2$	0.01	0.36	0.73	0.88
$4.6x10^2$	0.00	0.18	0.62	0.85
$1.0x10^3$	0.01	0.15	0.68	0.91
$2.2x10^3$	0.01	0.26	0.56	0.93
$4.6x10^3$	0.00	0.38	0.69	0.86

Appendix E: cESTI predictions

Table E-1: Observed cSNR-values of various studies using Dutch speech material and predicted cSNR-values using the ESTI, cESTI1 and cESTI2. VU98: Sentence material by Versfeld et al. (2000); DC: Duty Cycle, followed by the percentage on-time of the noise; MD: Modulation Depth, followed by the MD of the noise interruptions in dB; SIM: Sinusoidally Intensity Modulated; ICRA (International Collegium of Rehabilitative Audiology): Noises as described by Dreschler et al. (2001); HF: High Frequency; LF: Low Frequency.

Study	Speech	Noise	T60 (s)	cSNR (dB)	ESTI	$cESTI_1$	$cESTI_2$
Van Schoonhoven *et al.* (2019)	VU98 female	Interrupted noise (8 Hz DC50)	0.0	-15.0	-15.6	-10.6	-8.3
			0.1	-6.2	-7.3	-6.4	-5.2
			0.4	1.5	-0.5	-1.0	-0.3
			0.8	5.8	3.6	2.8	3.6
			1.2	10.1	8.2	6.6	8.1
		ISTS	0.0	-7.0	-18.6	-17.4	-14.4
			0.1	-3.1	-14.8	-14.3	-11.7
			0.4	0.8	-8.2	-9.2	-7.3
			0.8	8.6	-1.0	-2.6	-0.9
			1.2	12.0	5.7	3.2	5.3
Rhebergen *et al.* (2006)	VU98 female	Interrupted noise (4 Hz DC50)	0.0	-14.9	-26.9	-10.1	-7.7
		Interrupted noise (8 Hz DC40)		-22.8	-25.8	-21.7	-14.2
		Interrupted noise (8 Hz DC45)		-20.8	-23.4	-17.4	-11.5
		Interrupted noise (8 Hz DC50)		-17.6	-20.7	-13.6	-10.4
		Interrupted noise (8 Hz DC55)		-14.4	-17.3	-11.8	-9.6
		Interrupted noise (8 Hz DC60)		-11.7	-14.1	-10.8	-9.0
		Interrupted noise (16 Hz DC50)		-15.0	-14.8	-15	-14.0
		Interrupted noise (32 Hz DC50)		-11.1	-10.8	-11.1	-10.7
		Interrupted noise (64 Hz DC50)		-7.5	-8.0	-8.3	-7.9

Table E-1: Continued.

Study	Speech	Noise	T60 (s)	cSNR (dB)	ESTI	$cESTI_1$	$cESTI_2$
Rhebergen *et al.* (2006)	VU98 female	Interrupted noise (128 Hz DC50)		-5.4	-6.1	-6.5	-6.1
		SIM noise (8 Hz)		-5.8	-8.1	-7.8	-7.0
		SIM noise (16 Hz)		-7.1	-7.9	-8.2	-7.8
		SIM noise (32 Hz)		-6.5	-7.5	-7.8	-7.4
		SIM noise (64 Hz)		-6.2	-6.9	-7.3	-6.9
		SIM noise (128 Hz)		-5.3	-6.1	-6.5	6.1
		Sawtooth noise (8 Hz increasing)		-9.3	-11.3	-10.9	-9.9
		Sawtooth noise (8 Hz decreasing)		-11.9	-13.0	-12.4	-11.2
Francart *et al.* (2011)	VU98 male	ICRA5. noise (250 ms gap length)	0.0	-8.4	-16.0	-14.9	-12.3
		ICRA noise (100 ms gap length)		-6.7	-13.9	-13.1	-11.3
		Swedish speaker		-6.8	-18.0	-15.5	-12.9
		Dutch speaker		-3.6	-16.4	-15.9	-13.8
		ISTS		-6.4	-15.1	-14.2	-12.3
	VU98 female	ICRA5. noise (250 ms gap length)		-11.2	-18.7	-18.0	-15.2
		ICRA noise (100 ms gap length)		-10.2	-16.8	-15.9	-13.9
		Swedish speaker		-11.6	-20.3	-18.0	-15.5
		Dutch speaker		-8.8	-18.4	-18.0	-16.0
		ISTS		-5.5	-17.0	-16.1	-14.1
Rhebergen *et al.* (2014)	VU98 female	Interrupted noise (8 Hz DC50)	0.0	-24.3	-19.1	-10.4	-8.0
		Interrupted noise (8 Hz DC25)		-39.7	-38.4	-35.6	-31.1
George *et al.* (2006)	VU98 male	Speech modulated noise (VU98 male)	0.0	-11.7	-12.6	-11.0	-9.0
		Interrupted noise (16 Hz DC50)		-17.3	-11.7	-12.1	-11.3
		Interrupted noise (16 Hz DC75)		-6.4	-6.4	-6.7	-6.3
		Interrupted noise (16 Hz DC50 MD15)		-8.8	-8.0	-8.3	-7.8
		Interrupted noise (32 Hz DC50)		-12.7	-8.7	-9.1	-8.6
George *et al.* (2008)*	VU98 male	Speech modulated noise (VU98 male)	0.1	-7.8	-11.7	-10.4	-8.4
			0.3	-2.5	-6.4	-6.2	-4.7
			0.5	1.6	-2.6	-2.9	-1.7
			1.0	5.1	2.8	1.9	3.1
Van Esch *et al.* (2013)	VU98 female	ICRA4 fluct. noise (250ms gap length)		-12.8	-16.7	-16.7	-14.6
Rhebergen *et al.* (2008)	VU98 female	Speech modulated noise (VU98 female)	0.0	-12.5	-15.9	-13.1	-9.9
		Forward speech (Plomp male)		-17.5	-22.9	-20.1	-17.3
		Backward speech (Plomp male)		-17.0	-21.9	-20.8	-17.8
Maalderink *et al.* (2011)	VU98 female	Interrupted noise (8 Hz DC50)	0.0	-12.1	-10.3	-12.1	-10.3
				-8.2	-6.5	-6.1	-5.2
		Interrupted noise (LF gaps)		-10.4	-10.1	-8.5	-7.0
				-6.0	-6.2	-5.9	-5.0
		Interrupted noise (HF gaps)		-14.1	-11.8	-10.7	-9.3
				-9.4	-7.9	-7.7	-6.9