Toward a clinically viable spectro-temporal modulation test for predicting supra-threshold speech reception in hearing-impaired listeners

Zaar, Johannes; Simonsen, Lisbeth Birkelund; Dau, Torsten; Laugesen, Søren

Published in:
Hearing Research

Link to article, DOI:
10.1016/j.heares.2022.108650

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Hearing Research 427 (2023) 108650

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research Paper

Toward a clinically viable spectro-temporal modulation test for predicting supra-threshold speech reception in hearing-impaired listeners

Johannes Zaar[a,b,*], Lisbeth Birkelund Simonsen[c], Torsten Dau[b], Søren Laugesen[c]

[a] Eriksholm Research Centre, DK-3070 Snekkersten, Denmark
[b] Hearing Systems Section, Department of Health Technology, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
[c] Interacoustics Research Unit, DK-2800, Kgs. Lyngby, Denmark

ARTICLE INFO

Article history:
Received 28 February 2022
Revised 5 November 2022
Accepted 12 November 2022
Available online 21 November 2022

Keywords:
Hearing impairment
Speech intelligibility
Spectro-temporal modulation
Psychoacoustics
Noise
Clinical Test

ABSTRACT

The ability of hearing-impaired listeners to detect spectro-temporal modulation (STM) has been shown to correlate with individual listeners' speech reception performance. However, the STM detection tests used in previous studies were overly challenging especially for elderly listeners with moderate-to-severe hearing loss. Furthermore, the speech tests considered as a reference were not optimized to yield ecologically valid outcomes that represent real-life speech reception deficits. The present study investigated an STM detection measurement paradigm with individualized audibility compensation, focusing on its clinical viability and relevance as a real-life supra-threshold speech intelligibility predictor. STM thresholds were measured in 13 elderly hearing-impaired native Danish listeners using four previously established (noise-carrier based) and two novel complex-tone carrier based STM stimulus variants. Speech reception thresholds (SRTs) were measured (i) in a realistic spatial speech-on-speech set up and (ii) using co-located stationary noise, both with individualized amplification. In contrast with previous related studies, the proposed measurement paradigm yielded robust STM thresholds for all listeners and conditions. The STM thresholds were positively correlated with the SRTs, whereby significant correlations were found for the realistic speech-test condition but not for the stationary-noise condition. Three STM stimulus variants (one noise-carrier based and two complex-tone based) yielded significant predictions of SRTs, accounting for up to 53% of the SRT variance. The results of the study could form the basis for a clinically viable STM test for quantifying supra-threshold speech reception deficits in aided hearing-impaired listeners.

1. Introduction

For almost a century, measuring pure-tone detection thresholds at a range of frequencies has been the most common method for assessing auditory function (for a review see Jerger, 2018). The definition of hearing loss is still largely based on pure-tone audiometry, i.e., the audiogram, and the same holds true for the prescription of hearing-aid amplification. While the audiogram is undoubtedly a powerful descriptor of auditory function, it does have its limits in the sense that there is no straightforward relationship between the ability to detect soft pure tones and supra-threshold deficits that listeners experience when trying to make sense of audible but complex acoustic information. This concerns especially speech perception in background noise, which represents a major challenge that hearing-impaired (HI) and elderly listeners face (for a review see Lopez-Poveda, 2014).

Plomp (1986) categorized the effect of hearing dysfunction on speech intelligibility based on two main components: audibility and distortion. According to Plomp's model, both components can manifest themselves in elevated speech reception thresholds (SRTs) in noise, meaning that listeners with either or both of these deficits require a higher signal-to-noise ratio (SNR) to reach a given speech-intelligibility level (commonly 50% correct). The audibility component in Plomp's definition is related to fully or partially inaudible speech signals and can potentially be overcome by means of amplification. In contrast, the distortion component constitutes a supra-threshold deficit that exists for fully audible speech stimuli and thus cannot be counteracted with amplification. Several studies have attempted to account for such supra-threshold aspects in auditory perception using psychoacoustic tests, mainly fo-

* Corresponding author at: Eriksholm Research Centre, DK-3070 Snekkersten, Denmark.
E-mail address: jozr@eriksholm.com (J. Zaar).

https://doi.org/10.1016/j.heares.2022.108650

cusing on frequency resolution, loss of compression, and/or temporal processing (e.g., Strelcyk and Dau, 2009; Johannesen et al., 2014; Thorup et al., 2016). The implicit assumption in this context is usually that the audiogram does not predict supra-threshold deficits because it is considered to be audibility-related only. However, pure-tone sensitivity may still to some extent be related to some supra-threshold deficits. For instance, recent work based on a data-driven classification approach using a multitude of psychoacoustic tests (Sanchez-Lopez et al., 2020) indicated that the audiogram may be a more powerful predictor of supra-threshold hearing than traditionally assumed.

It is still not fully understood which auditory deficits contribute to the supra-threshold speech reception deficits in HI listeners. However, it may be possible to design a clinical test that reflects the real-life speech reception difficulty experienced by HI listeners after audibility has been restored (e.g. by means of hearing aids). This would have the advantage of being applicable in clinical practice as well as presumably being language independent, whereas direct measurements of speech reception with hearing aids in realistic conditions (requiring multiple loudspeakers combined with a rather complicated set up and procedure) are hardly clinically feasible and have to be defined for and conducted in the listener's first language. Such a psychoacoustic, clinical test could provide a diagnostic tool for hearing-aid selection, counselling, and hearing-aid processing prescription (e.g., regarding noise reduction and speech enhancement).

One candidate for such a test is the spectro-temporal modulation (STM) test. Chi et al. (1999) defined a stimulus consisting of an amplitude modulation pattern that combines both spectral modulations in cycles/octave (c/o) and temporal modulations in Hertz (Hz), resulting in upward or downward moving ripples, and a broadband noise carrier signal that the pattern is imposed on. The test measures the just-detectable modulation depth required to distinguish a modulated from an unmodulated carrier signal, with normal-hearing (NH) listeners showing a low-pass characteristic in relation to both the temporal and spectral modulation frequency. Chi et al. (1999) linked these STM patterns to speech perception using an auditory model employing spectro-temporal modulation filters (the spectro-temporal modulation index, see also Elhilali et al., 2003), arguing that speech can be characterized in a physiologically plausible way using a multitude of such spectro-temporal modulation "basis functions".

Bernstein et al. (2013) measured STM detection performance in 8 NH and 12 HI listeners and compared the measured thresholds to the listeners' speech intelligibility in stationary speech-shaped noise (SSN) measured at an SNR of 0 dB. STM thresholds were obtained for all combinations of four spectral (0.5, 1, 2, and 4 c/o) and three temporal (4, 12, and 32 Hz) modulation frequencies using a four-octave wide pink noise carrier. The test was monaurally administered using a 1-up/3-down tracking rule (approaching 79.4% correct; see Levitt, 1971) in a two-alternative forced choice (2-AFC) paradigm. In an attempt to overcome audibility limitations in the HI listeners, the stimuli in both the STM test and the speech test (using IEEE sentences) were presented at high sound pressure levels (SPLs) of 86 and 92 dB, respectively. Bernstein et al. (2013) selected the STM condition with 2 c/o and 4 Hz because it induced the largest performance difference between the NH and HI groups. They demonstrated that the STM thresholds in this condition were an excellent predictor of the HI listeners' speech intelligibility, explaining 61% of the speech intelligibility variance in the HI group and thus substantially more than an audiogram-based speech intelligibility prediction using the speech intelligibility index (SII; 40% variance explained).

Mehraei et al. (2014) followed up on the Bernstein et al. (2013) study using the same listeners and test procedures, but narrower, one-octave wide noise carriers with flanking noise to mask potential edge cues. The resulting STM thresholds were compared to SRTs obtained in stationary- as well as fluctuating-noise conditions (obtained from the same data set as used by Bernstein et al., 2013). Compared to Bernstein et al. (2013), Mehraei et al. (2014) showed a higher predictive power of the audiogram-based SII (60% for SSN and 72% for modulated noise), presumably due to the use of SRTs instead of percent correct at a single SNR. The STM stimulus variant centered at 1 kHz with 2 c/o and 4 Hz was found to be a significant predictor of SRTs beyond the audiogram in both noise conditions, suggesting an important role of the low-frequency portion of the stimulus, albeit with a reduced predictive power as compared to the broadband stimulus considered in Bernstein et al. (2013).

Bernstein et al. (2016) measured monaural STM detection thresholds in 154 HI listeners with mild-to-moderate hearing loss and compared them with SRTs measured in SSN and four-talker babble using a matrix sentence test. Importantly, this study did not use high speech SPLs, as opposed to Bernstein et al. (2013) and Mehraei et al. (2014), but rather added individualized simulated hearing-aid amplification based on the across-ear average audiogram and a nominal speech level of 65 dB SPL. The STM measurement procedure was similar to that of Bernstein et al. (2013); however, only one stimulus variant (with 2 c/o and 4 Hz, upward moving) was considered and low-pass filtered at 2 kHz, motivated by the finding from Mehraei et al. (2014) that the low-frequency portion of the stimulus was crucial. The presentation level was 85 dB SPL. In contrast to Bernstein et al. (2013), 53 listeners (i.e., about a third) were not able to obtain a threshold in the adaptive procedure. This problem was tackled by employing a constant-stimuli approach, measuring percent correct at full modulation and extrapolating the result towards physically unattainable modulation depths greater than 0 dB. However, the extrapolation was based on the assumption that these listeners exhibited the same slope of the underlying psychometric function as the group of all remaining listeners, which may not hold and thus result in limited precision. Bernstein et al. (2016) reported that, after averaging SRTs across various conditions, the audiogram-related predictor (high-frequency average) explained 31% of the SRT variance and the STM thresholds 28%, amounting to 44% when combined. Thus, the predictive power of the STM test was lower than in the previous studies. An additional analysis revealed that the STM test did not provide any predictive power for a group of listeners that were older than 65 years and had a high-frequency hearing loss of more than 53 dB Hearing Level (HL), for a majority of which the extrapolation method described above had been applied.

Thus, while the STM test has been shown to predict supra-threshold speech intelligibility, it is unclear whether the test is fully applicable in elderly HI people as it might be too challenging overall. Furthermore, the speech tests considered in the previous investigations used artificial masking noise types, provided no spatial information, were undertaken at very high SPLs, and/or used closed-set matrix sentences as speech material, all of which strongly limit the ecological validity of the measured speech intelligibility outcomes (see Keidser et al., 2020).

In the present study, a modified STM test paradigm was used and six different STM stimulus variants were explored with the aim to facilitate the STM test whilst optimizing its predictive power with respect to ecologically valid SRTs. The motivation behind the study was to work toward a clinically viable supra-threshold test that (i) all listeners are able to complete, including elderly and severely HI listeners, and that (ii) reflects real-life speech reception difficulties. Numerous modifications were made to the STM measurement paradigm and four previously established STM stimulus variants Bernstein et al. (2013; 2016) were explored alongside two novel stimulus variants that are based on complex-tone carriers instead of noise carriers. A spatialized speech-on-

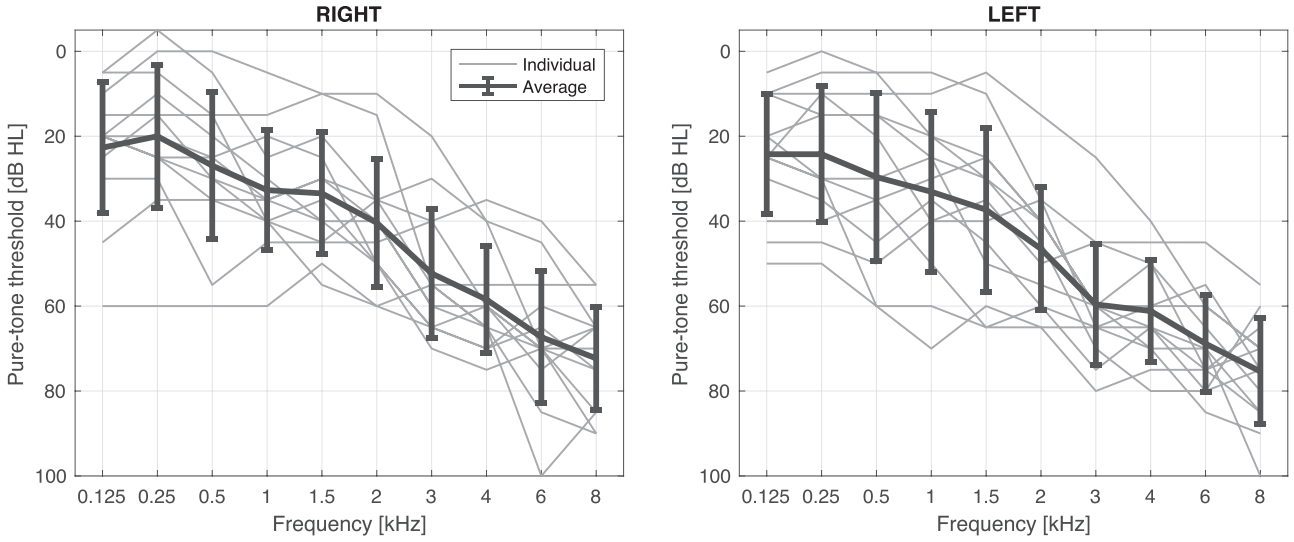

Fig. 1. Pure-tone thresholds obtained for the right and left ear of the 13 individual listeners (gray lines) along with mean and standard deviation across listeners (thick black lines).

speech test with open-set sentences and individualized audibility compensation was designed to obtain ecologically valid SRTs representative of aided listening; additionally, a non-spatialized SSN condition was included for comparison. All tests were conducted with a group of elderly HI listeners. The resulting STM thresholds and SRTs were analyzed at the group level as well as using the individual data to assess the predictive power of the STM thresholds with respect to the SRTs.

2. Method

2.1. Listeners

Thirteen HI listeners (10 male, 3 female) aged between 61 and 82 years (mean: 72.2 years; standard deviation: 5.2 years) were recruited from the test-person database of the Hearing Systems Section at the Technical University of Denmark. All listeners were native speakers of Danish. They underwent standard clinical audiometry with pure-tone thresholds measured at 0.125, 0.25, 0.5, 1, 1.5, 2, 3, 4, 6, and 8 kHz. The individual and average audiograms for the right and left ear, respectively, are shown in the two panels of Fig. 1. The listeners exhibited a range of mild to severe hearing losses, which were largely symmetric (defined as thresholds within 15 dB between ears for a minimum of 8 of the 10 test frequencies). All listeners provided informed consent and were offered financial compensation. All experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

2.2. Spectro-temporal modulation detection test

2.2.1. Stimulus design

The STM stimulus is based on the combination of a carrier signal and a modulator signal. The modulator introduces spectro-temporal modulation, i.e., modulation applied jointly along the (log) frequency axis and the time axis, resulting in a ripple pattern that moves upward or downward over time. The modulator M is mathematically defined as (see also Chi et al., 1999):

$$M(x,t) = \ m \cdot \sin(2\pi \cdot [f_{temp} \cdot t + f_{spec} \cdot x] + \varphi) \qquad (1)$$

where $x = \log_2(f/f_0)$ is the position along the logarithmic frequency axis defined relative to the lower boundary frequency f_0 of the spectrum, t represents time, m stands for the modulation depth ($m = 0$: unmodulated; $m = 1$: fully modulated), f_{temp} denotes the temporal modulation frequency in Hz, f_{spec} denotes the spectral modulation frequency in c/o, and φ represents the starting phase (in radians). The moving direction of the ripples is upward if f_{temp} is negative and downward if f_{temp} is positive. The full stimulus $S(x,t)$ is generated by multiplying the carrier $C(x,t)$, defined in the same time-frequency space as the modulator, with the modulator $M(x,t)$ according to:

$$S(x,t) = C(x,t) \cdot (1 + M(x,t)) \qquad (2)$$

Summation across the logarithmic frequencies x yields the stimulus time signal $s(t)$.

The present study considered six different STM stimulus variants that differ both in the carrier and the modulator. The details are described below; an overview of the physical characteristics and parameters is provided in Table 1. An auditory-inspired time-frequency representation of the stimulus variants, obtained using a gammatone filterbank and Hilbert envelope extraction, is provided in Fig. 2.

Three of the variants were related to Bernstein et al. (2013) and thus based on a 4-octave wide noise carrier with a bandwidth between 354 Hz and 5656 Hz, generated using 4000 random-phase sinusoidal components with equidistant spacing along the logarithmic frequency axis x (resulting in a "pink" spectrum). This carrier was combined with three different modulators for the three stimulus variants (termed $\text{Noisy}_{1c/o}$, $\text{Noisy}_{2c/o}$, and $\text{Noisy}_{4c/o}$, see also Table 1 and top panels of Fig. 2), defined by spectral modulation frequencies f_{spec} of 1, 2, and 4 c/o and a temporal modulation frequency f_{temp} of 4 Hz, with upward-moving ripples. The upward moving direction was chosen to facilitate STM detection, as it has been reported to yield somewhat lower thresholds than the downward direction (e.g., Chi et al., 1999).

In addition, the stimulus considered by Bernstein et al. (2016) was generated by band-limiting the above-mentioned noise carrier to the frequency band between 354 and 2000 Hz (resulting in 2499 components and 2.5 octaves) and combining it with a modulator with f_{spec} of 2 c/o and f_{temp} of 4 Hz (upward-moving ripples; $\text{Noisy-LP}_{2c/o}$, see also Table 1 and bottom-left panel of Fig. 2).

Lastly, two novel STM stimulus variants were generated based on a 4-octave wide 100-Hz complex-tone carrier within the frequency band between 354 Hz and 5654 Hz using 54 random-phase sinusoidal components with equidistant (100-Hz) spacing along

Table 1
The six different STM stimulus variants considered in the present study. Their names are shown in the top row; the remaining rows indicate the properties of the stimulus variants. The moving direction was upward in all cases.

	$Noisy_{1c/o}$	$Noisy_{2c/o}$	$Noisy_{4c/o}$	$Noisy\text{-}LP_{2c/o}$	$Tonal_{1c/o}$	$Tonal_{2c/o}$
Carrier	Noise	Noise	Noise	Noise	100-Hz Tone	100-Hz Tone
Modulator	1 c/o; 4 Hz	2 c/o; 4 Hz	4 c/o; 4 Hz	2 c/o; 4 Hz	1 c/o; 4 Hz	2 c/o; 4 Hz
Frq. range	354–5656 Hz	354–5656 Hz	354–5656 Hz	354–2000 Hz	354–5654 Hz	354–5654 Hz

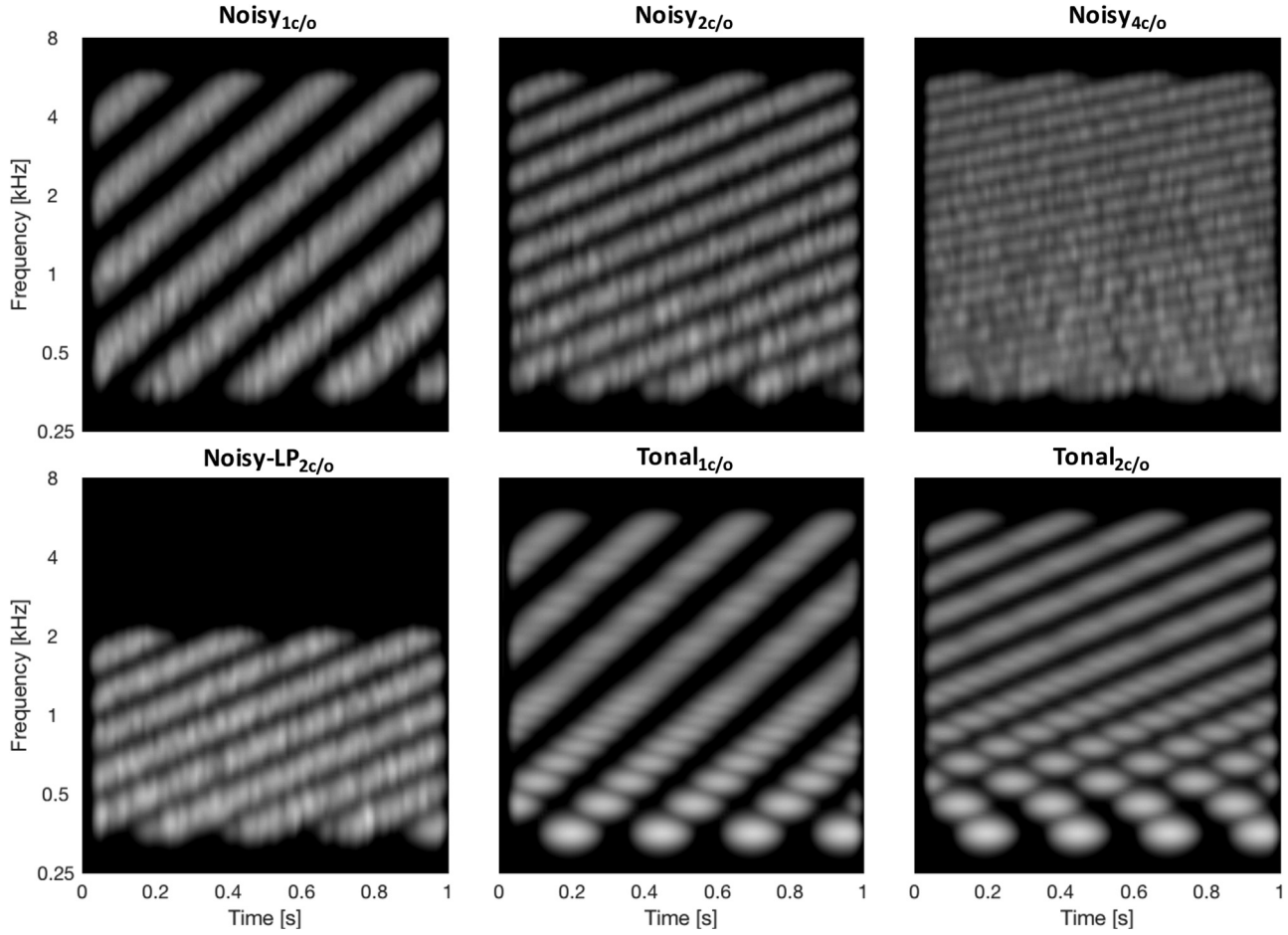

Fig. 2. Auditory spectrograms of STM stimulus variants considered in the present study. Top left to right: 4-octave wide random noise carriers modulated with 4 Hz and 1, 2, and 4 c/o. Bottom left: band-limited random noise carrier modulated with 4 Hz and 2 c/o. Bottom middle and right: 4-octave wide 100-Hz complex tone carrier modulated with 4 Hz and 1 c/o (middle) and 2 c/o (right). The moving direction of the ripples was upward for all variants.

the linear frequency axis. The power of the sinusoidal components was weighted by $1/f$ to obtain a pink spectrum comparable to the noise-carrier spectra described above. This carrier was combined with two modulators for the two stimulus variants (termed $Tonal_{1c/o}$ and $Tonal_{2c/o}$, see also Table 1 and bottom middle and bottom left panels of Fig. 2), defined by spectral modulation frequencies f_{spec} of 1 and 2 c/o and a temporal modulation frequency f_{temp} of 4 Hz (upward-moving ripples). These novel stimulus variants were designed to be closely related to actual speech signals (in an abstract fashion) not only in terms of the modulation patterns (cf. Chi et al., 1999; Elhilali et al., 2003) but also regarding the carrier signal, which is mostly periodic in natural speech (despite the presence of turbulent consonant sounds) with an average fundamental frequency roughly between 100 and 200 Hz. Another motivation for using the complex-tone carrier was to avoid the intrinsic low-frequency temporal fluctuations contained in the noise carrier, which potentially mask the imposed 4-Hz modulation (cf. Dau et al., 1999), as the complex-tone carrier shows no fluctuations below 100 Hz.

2.2.2. Adaptation of measurement paradigm

The overall measurement paradigm was designed with the aim to facilitate the adaptive STM threshold measurement such that even the poorest-performing listeners would be able to complete the task and obtain a threshold. This section describes the crucial modifications of the paradigm as compared to previous studies and the rationales behind them.

While all referenced studies measured STM detection monaurally, i.e., separately for the two ears, the STM test in the current study was administered binaurally. The idea behind this was to measure the entire system in “one shot” to obtain a performance level that is representative of speech comprehension in real life, where both ears are also used jointly. This means that listeners were able to use their better ear and/or make use of binaural summation, both of which should theoretically facilitate the STM detection task.

Whereas Bernstein et al. (2013), Mehraei et al. (2014), and Bernstein et al. (2016) presented the stimuli at fixed high presentation levels to all listeners, individualized linear amplification

was provided here to ensure audibility irrespective of the hearing loss whilst avoiding unnecessarily high levels that may deteriorate performance (Magits et al., 2019) or lead to discomfort. The amplification was based on a "sufficiently audible" approach (cf. Humes, 2007) that ensures a minimum of 15 dB sensation level (SL) in each 3rd-octave frequency band.

Bernstein et al. (2013), Mehraei et al. (2014), and Bernstein et al. (2016) used a 2-AFC procedure combined with a 1-up/3-down rule, tracking the 79.4% point on the psychometric function (Levitt, 1971). In contrast, a 3-AFC procedure was applied in the present study combined with a 1-up/2-down rule, tracking the 70.7% point on the psychometric function. This choice has two main implications. Firstly, 3-AFC allows to listen for the interval that is different from the other two, whereas 2-AFC requires an internal "template" of the imposed modulation (as opposed to the intrinsic fluctuations of a noise carrier), which may lead to confusion and thus worse performance. Secondly, the 1-up/2-down rule tracks a lower percent-correct point on the underlying psychometric function than the 1-up/3-down rule and the resulting threshold will thus be lower for a given performance level. Both of these aspects are expected to facilitate the threshold measurement.

Bernstein et al. (2013), Mehraei et al. (2014), and Bernstein et al. (2016) presented the individual intervals at presentation levels that were roved (i.e., randomized) by ±2.5 dB around the nominal presentation level.[1] This was done in order to reduce the effectiveness of a potential loudness cue. In the present study, no level roving was applied because the probability of a loudness cue playing a role in these stimuli (presented at the same level) was considered to be very low, whereas it was expected that the level roving may confuse listeners and thus make the task more difficult. Furthermore, different listeners may be confused to different degrees by the level roving depending on their cognitive abilities, which could provide an additional unwanted source of variability to the measured thresholds.

Finally, whereas Bernstein et al. (2013), Mehraei et al. (2014), and Bernstein et al. (2016) chose an interval duration of 500 ms, an interval duration of 1 s was chosen here in order to provide a longer observation time window to facilitate STM detection.

2.2.3. Procedure and apparatus

The STM test was administered using the AFC software package (version 1.40, Ewert, 2013) for psychoacoustic testing, running under Matlab on a Windows-based PC. STM thresholds were adaptively measured using a 3-AFC procedure with a 1-up/2-down tracking rule. The tracking variable was defined as the modulation depth m in dB full scale (FS), i.e., $L_m = 20 \cdot \log_{10}(m)$. In each trial, three intervals were played in random order. Two intervals contained the carrier only and one interval contained the modulated carrier. The starting phases of the carrier components were drawn from a uniform distribution in the interval $[-\pi, \pi]$ for each trial but identical ("frozen") across the three intervals within a given trial, such that the only physical difference across intervals in any given trial was the imposed modulation. The starting phase φ (see eq. (1)) of the STM profile was also drawn from a uniform distribution in the interval $[-\pi, \pi]$ for each trial. Each interval had a duration of 1 s with 50-ms fade in/out raised-cosine ramps. The inter-stimulus interval (pause between intervals) was 0.5 s. The initial value of the tracking variable was 0 dB FS, i.e., fully modulated; according to the 1-up/2-down tracking rule, L_m was increased after an incorrect response, unchanged after a correct response, and reduced after two successive correct responses. In case of an incorrect response at 0 dB FS the tracking variable remained at 0 dB FS; after 3 consecutive incorrect responses at 0 dB FS, the run was aborted. The initial step size was 4 dB, which was reduced to 2 dB after the first upper reversal and further reduced to 1 dB after the second upper reversal. Once the 1-dB step size was reached, the procedure continued until 8 (4 lower and 4 upper) reversals were reached, at which point it terminated. The resulting threshold was defined as the mean across the tracking-variable values at the last 8 reversals.

The listeners were seated in a sound-attenuating booth in front of a computer screen and bilaterally presented with the stimuli using Sennheiser (Wedemark, Germany) HDA200 headphones. The stimuli were played at a sampling rate of 48 kHz and a resolution of 16 bits, D/A converted through an RME (Haimhausen, Germany) Fireface UCX, and amplified through an SPL (Niederkrüchten, Germany) Phonitor mini stereo headphone amplifier. The frequency response of the headphone amplifier and the headphones was equalized based on a Brüel&Kjær (Nærum, Denmark) head and torso simulator (HATS) measurement. The stimuli were presented at a nominal broadband level of 65 dB SPL at the center-of-head position in a virtual free field, simulated by applying the diffuse-field-to-eardrum transformation from Moore et al. (2008), which yields a broadband level of 69.8 dB SPL at the eardrum. Linear amplification was applied where applicable to compensate for the individual hearing loss, such that listeners received a minimum of 15 dB SL in each 3rd-octave frequency band within the stimulus frequency range (cf. Humes, 2007). The amplification filter was obtained via comparison between the 3rd-octave band stimulus levels and an interpolated version of the across-ear average audiogram added to the minimum audible pressure (MAP) levels for NH (Moore et al., 2008). The listeners provided their responses using a touch screen, a computer keyboard, or a computer mouse, according to their preference. They received visual feedback after each response (correct/incorrect). At the beginning of a test session, a short training run was provided by means of a simple temporal amplitude-modulation detection task (using broadband noise with a 4-Hz sinusoidal temporal modulation) to familiarize the listeners with the procedure. For each stimulus variant, three adaptive measurements were conducted and the mean of the resulting three thresholds was considered as the final result.

2.3. Speech test

2.3.1. Spatial set up, stimuli, and conditions

The experiment took place in a quiet room designed in accordance with IEC 268-13 (1985). The room was middle-sized (7.52×4.74×2.76 m) and had a reverberation time T_{30} of about 0.5 s, corresponding to a typical living-room environment. Five Dynaudio (Skanderborg, Denmark) BM6 loudspeakers were positioned along a circle with a diameter of 2.5 m at azimuth angles of 0°, ±100°, and ±155° (see Fig. 3). The listeners were seated in a chair equipped with a headrest such that their viewing direction was 0° azimuth, their center of head was in the center of the loudspeaker arrangement and their ears were at the same height as the loudspeakers.

Speech intelligibility was measured based on the Danish Hearing in Noise Test (HINT, Nielsen and Dau, 2011). The target speech consisted of 5-word meaningful, grammatically correct Danish sentences uttered by a male speaker. Ten balanced lists of 20 sentences each and 3 lists of 20 training sentences are available in the Danish HINT corpus. The maskers consisted of (i) an SSN that had the same long-term spectral characteristics as the target speech and (ii) recordings of running speech spoken by four different male speakers that were also filtered to have the same long-term spectrum as the target. Two acoustical conditions were considered, termed multi-talker babble (MTB) separated and SSN co-located. In both conditions, the target speech sentences were presented from

[1] Bernstein et al. (2016) did not mention the level roving in their article but did apply it, as indicated by a copy of the experimental code provided by the first author of the study.

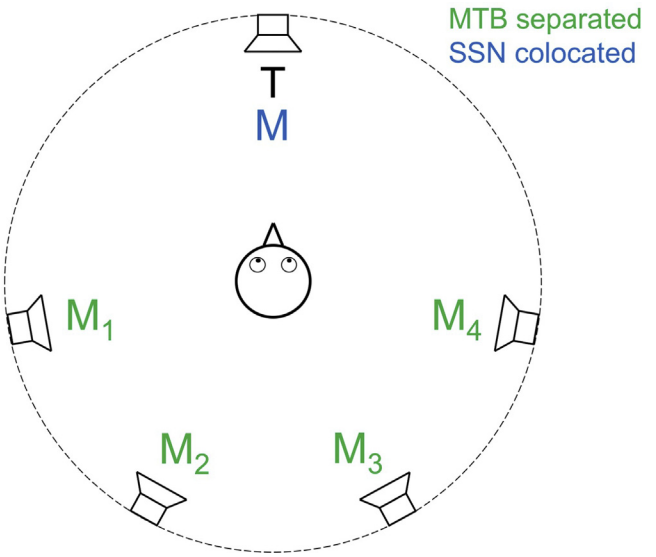

Fig. 3. Sketch of speech test set up. T refers to the target speech. M refers to the masker signals, where the green-colored M_1- M_4 show the positions of maskers in the multi-talker babble (MTB) separated condition and the blue-colored M relates to the masker position in the SSN co-located condition.

the viewing direction (0° azimuth, see black "T" in Fig. 3). In the MTB separated condition, the four running speech interferers were separately mixed with uncorrelated low-level SSN (−6 dB relative to the running speech interferers' long-term level) and presented from −100°, −155°, 155°, and 100°, respectively (see green-colored "M_1"- "M_4" in Fig. 3). In the SSN co-located condition, the SSN was played from the same loudspeaker as the target speech (0° azimuth; see blue-colored "M" in Fig. 3).

2.3.2. Procedure and apparatus

The listeners were seated on the chair in the center of the loudspeaker arrangement and instructed to use the headrest to maintain a static head position. They were asked to verbally repeat the target-sentence words they had understood into a microphone placed in front of them, slightly below the head. The responses were manually scored by an audiologist (the second author) who is a native speaker of Danish. The speech test was run using a dedicated software under Matlab (Mathworks, Natick, Massachusetts, USA) on a PC. All sound was delivered by an RME Fireface UCX soundcard at a sampling rate of 48 kHz and a resolution of 16 bits. The speech-test stimuli were amplified using customized Bang & Olufsen (Struer, Denmark) amplifiers and the talkback microphone was routed to the soundcard's headphone output such that the audiologist could listen to the responses on Sennheiser HDA200 headphones. The target speech was presented at a nominal level of 70 dB SPL (C-weighted) relative to the center-of-head position, whereas the masker levels were adapted to modify the SNR. SRTs at the 50%-sentences-correct level were tracked by adjusting the SNR according to sentence correct scoring (see Nielsen and Dau, 2011), where the first sentence was presented at a low SNR and repeated in increasing SNR steps of 4 dB until it was correctly identified (all five words). For sentences 2–4, the SNR was increased/decreased by 4 dB after an incorrect/correct response, respectively. Then, an average across the SNRs used in the previous 4 presentations was made and the SNR was adjusted from there in steps of 2 dB.

In analogy to the approach used in the STM test, linear amplification was applied where applicable to compensate for the individual hearing loss. The amplification filter was obtained via comparison between the 3rd-octave band long-term target speech levels and an interpolated version of the across-ear average audiogram added to the NH minimum audible field (MAF) levels (Moore et al., 2008) to guarantee a minimum SL. This minimum target speech level was set at 15 dB SL in each 3rd-octave band centered at frequencies between 0.125 and 3 kHz. Due to the reduction in speech energy towards higher frequencies, and to protect the tweeters of the loudspeakers, this rule was relaxed in the high-frequency region (4 kHz: 12 dB SL; 6 kHz: 8 dB SL; 8 kHz: 4 dB SL). Furthermore, the maximum playback level from the individual loudspeakers was limited to 78 dB SPL in any 3rd-octave band to protect the loudspeakers. However, the weighted average gain deficit (band-importance weighted according to Pavlovic, 1987, Table II, rightmost column) was only 1.64 dB for the most affected listener and only 0.54 dB on average across all listeners.

A first training run was conducted with 20 sentences (one training list) in the MTB separated condition, starting at a SNR of 0 dB. Then, a second training run was conducted in the MTB separated condition, this time using 40 sentences (two training lists) and starting at an SNR 6 dB below the SRT measured in the first training run. Next, the MTB separated SRT measurement was conducted, using 40 sentences (HINT lists 7/8 or 9/10, balanced across listeners) and starting at an SNR 6 dB below the SRT measured in the second training run. Lastly, the SSN co-located SRT measurement was conducted, using 40 sentences (HINT lists 9/10 or 7/8, balanced across listeners) and starting at an SNR of −6 dB. The order of conditions was not counterbalanced (i) because the MTB separated condition represented the main speech-test condition, whereas the SSN co-located condition was tested mainly as a reference and (ii) to avoid an interaction between subjects and effects of presentation order (e.g. fatigue, training), which was undesirable for the correlation analyses considered in this study.

2.3.3. Data post-processing

In the standard HINT procedure (Nielsen and Dau, 2011), the SRT is defined as the average across the last 17 SNRs presented. Alternatively, all the measured data can be exploited to increase the precision of the SRT estimate, (i) using responses obtained with all 40 considered sentences and (ii) considering these responses at the words-correct level (0, 20, 40, 60, 80, or 100%) instead of the coarser sentence-correct level (correct/incorrect). To this end, SRTs were estimated using a maximum likelihood approach following the method suggested by Rønne et al. (2017). First, the midpoints of the psychometric functions for each listener and condition were estimated based on the percentages of words correct obtained for all sentences. To ensure the robustness of the SRT estimate, a fixed slope was assumed for these functions, which was the median slope across HI listeners from Rønne et al. (2017). As suggested by Rønne et al. (2017), the 77% point on the resulting *words-correct* psychometric function was used for the final SRT estimate as it, on average, corresponds to the 50% *sentence-correct* SRT.

2.4. Experimental design

The experiment was structured in three separate sessions of maximally 2 h each. In the first session, listeners were briefed on the different parts of the experiment and signed a consent form. An otoscopy was conducted and listeners with more than two-thirds occlusion due to earwax were re-scheduled and asked to see their doctor for ear cleaning. After passing the otoscopy check, an audiogram was measured using an Interacoustics (Middelfart, Denmark) AC40 clinical audiometer, unless a recent audiogram (maximally 1 year old) was available. After a short break, the speech test was run. Finally, listeners were introduced to the STM test, receiving instructions and trying first a simple temporal amplitude-modulation detection task (4 Hz) and then one run

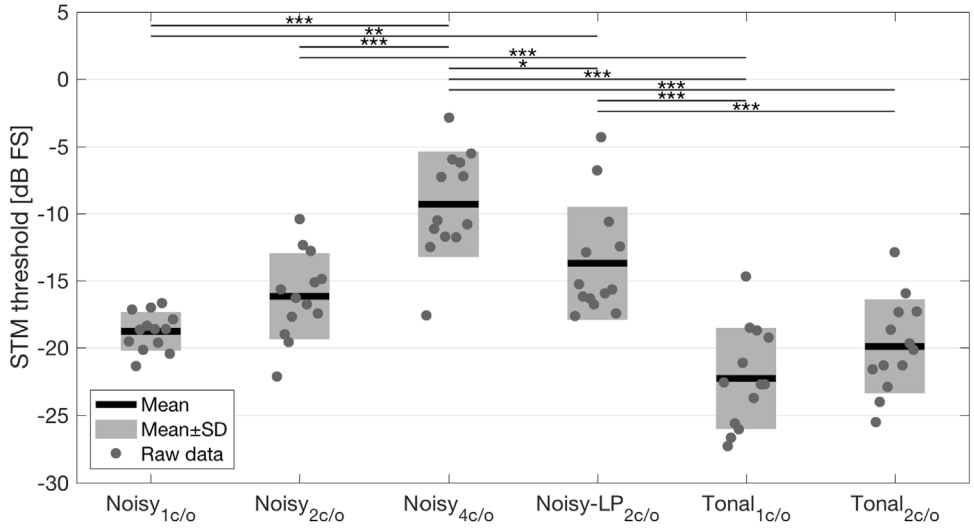

Fig. 4. STM thresholds for the considered stimulus variants. The black lines depict the group mean, the shaded areas represent ±1 standard deviation around the mean, and the gray dots show the individual results (with a slight horizontal offset for better visibility). The black horizontal lines in the top part indicate significant differences according to a one-way ANOVA with a Tukey-Kramer post-hoc test (with *: $p<0.05$, **: $p<0.01$, and ***: $p<0.001$).

with the $\text{Noisy}_{1c/o}$ STM stimulus variant. The second and third sessions each consisted of an initial otoscopy to check for ear wax, followed by a quick procedural training run using a 4-Hz temporal amplitude modulation detection task to reorientate listeners with the procedure using an easy task. Then, three blocks of three adaptive STM threshold runs were conducted, each block with a given STM stimulus variant. The order of STM conditions was randomized. Listeners were required to take short breaks in between the experimental blocks.

3. Results and analysis

3.1. Spectro-temporal modulation detection

Fig. 4 shows the STM thresholds obtained with the six considered stimulus variants in terms of group averages (black lines), standard deviations (gray shaded areas), and individual thresholds (gray dots). Importantly, all listeners were able to complete each task, i.e., each individual adaptive STM track converged to a threshold. The first three thresholds on the left of Fig. 4 show the results obtained for the three STM variants with a broadband noise carrier and a spectral modulation frequency of 1, 2, and 4 c/o ($\text{Noisy}_{1c/o}$, $\text{Noisy}_{2c/o}$, $\text{Noisy}_{4c/o}$, also considered by Bernstein et al., 2013), respectively. Next, the thresholds obtained with the low-pass filtered noise carrier variant with 2 c/o ($\text{Noisy-LP}_{2c/o}$, also considered by Bernstein et al., 2016) are shown. The two rightmost thresholds represent the conditions with a broadband 100-Hz complex tone carrier and a spectral modulation frequency of 1 and 2 c/o, respectively.

A one-way ANOVA (normally distributed residuals verified) showed a significant effect of stimulus variant [$F(5,72)=24.09$, $p<10^{-10}$] and a Tukey-Kramer post-hoc test revealed that 9 out of 15 possible pairwise comparisons were significant, with 7 of them being highly significant ($p<0.001$, see also Fig. 4). Broadly speaking, Fig. 4 indicates that the thresholds increased with increasing spectral modulation frequency, although it should be noted that the post-hoc test did not show significant differences between $\text{Noisy}_{1c/o}$ and $\text{Noisy}_{2c/o}$, $\text{Noisy}_{1c/o}$ and $\text{Tonal}_{2c/o}$, or $\text{Tonal}_{1c/o}$ and $\text{Tonal}_{2c/o}$, indicating that the clearest threshold increase was induced by the spectral modulation frequency of 4 c/o ($\text{Noisy}_{4c/o}$). Furthermore, the type of carrier seemed to have an influence as the thresholds were overall lower for the tonal carrier (two rightmost conditions shown in Fig. 4) than for the noise carriers (remaining conditions). However, no significant post-hoc comparisons were obtained between stimulus variants with similar spectral modulation frequency but different carriers, except between $\text{Noisy-LP}_{2c/o}$ and $\text{Tonal}_{2c/o}$, where both the type (noise vs. complex tone) and the bandwidth (354–2000 Hz vs. 354–5654 Hz) of the carriers differed. Nonetheless, an alternative analysis on an unbiased subset of the data (omitting the "odd" stimulus variants $\text{Noisy-LP}_{2c/o}$ and $\text{Noisy}_{4c/o}$) by means of a two-way ANOVA with the factors carrier type (noise or complex tone) and spectral modulation frequency (1 c/o or 2 c/o) indicated significant effects of both carrier type [$F(1,48)=17.87$, $p = 0.0001$] and spectral modulation frequency [$F(1,48)=8.66$, $p = 0.0054$], but no significant interaction between the effects [$F(1,48)=0.02$, $p = 0.8964$].

All variants induced comparable across-listener standard deviations of about 3–4 dB, with the notable exception of $\text{Noisy}_{1c/o}$ (leftmost bar in Fig. 4), which induced a substantially smaller standard deviation of 1.4 dB. An F-test indeed revealed that the across-listener variance measured in the $\text{Noisy}_{1c/o}$ condition was significantly smaller at the $p = 0.01$ level than the second smallest variance (measured in the $\text{Noisy}_{2c/o}$ condition), whereas the across-listener variances measured in the remaining five conditions were not significantly different from each other at the $p = 0.05$ level.

3.2. Speech intelligibility

Fig. 5 shows the SRTs measured in the speech test. The left panel depicts individual SRTs (gray dots) as well as group averages (black lines) and standard deviations (gray shaded areas) for the two considered conditions. It can be observed that the MTB separated condition yielded, on average, higher SRTs (0.9 dB) than the SSN co-located condition (−0.5 dB). A one-way ANOVA (normally distributed residuals verified) revealed that this difference was significant [$F(1,24)=5.61$, $p = 0.0263$]. The standard deviation across listeners almost doubled in the MTB separated condition (1.9 dB) as compared to the SSN co-located condition (1 dB). An F-test revealed that the across-listener variances were significantly different from each other at the p=0.05 level.

The right panel of Fig. 5 shows a scatter plot of the individual SRTs, i.e., the MTB separated SRTs measured for the individual lis-

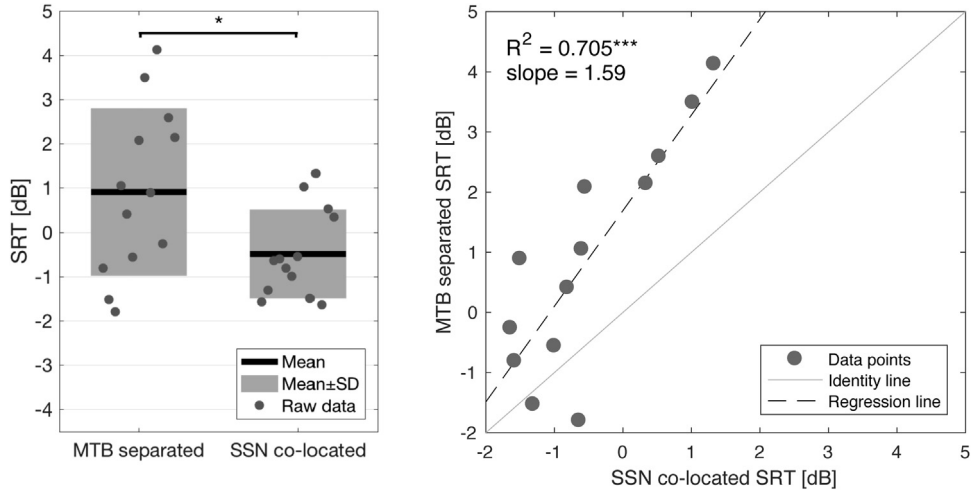

Fig. 5. Speech test results. Left panel: SRTs measured in the MTB separated and SSN co-located conditions. The black lines depict the group mean, the shaded areas represent ±1 standard deviation around the mean, and the gray dots show the individual results (with a slight horizontal offset for better visibility). The black horizontal line in the top part indicates a significant difference (p<0.05) according to a one-way ANOVA. Right panel: individual SRTs measured in the MTB separated condition versus the SRTs measured in the SSN co-located condition (gray dots). The solid gray line indicates equivalence between the two conditions; the dashed line represents a regression line fit to the data, shown along with the corresponding R^2 (p<0.001) and slope estimates.

teners versus the corresponding SSN co-located SRTs (gray dots). It can be seen that the SRTs obtained in the two conditions were strongly correlated (R^2=0.705, p = 0.0003). However, the correlation appeared to be strongly driven by the four listeners with the highest SRTs, whereas the remaining nine listeners showed almost no correlation between the two conditions (R^2=0.098, p = 0.413). A linear regression line fit considering all data points demonstrates again that the SRT value range was strongly expanded in the MTB separated condition as compared to the SSN co-located condition, as evidenced by the dashed regression line (slope of 1.59 dB/dB), especially in comparison with the solid identity line (slope of 1 dB/dB).

3.3. Correlation analysis

To investigate to what extent the performance in the different STM tests was related to speech-reception performance, a correlation analysis was conducted. In addition to the six STM variants, the 4-PTA (average across pure-tone thresholds at 0.5, 1, 2, and 4 kHz) was considered as a pure-tone audiometry-based predictor. Fig. 6 shows the Pearson's correlation coefficients between the resulting seven potential predictors (from left to right) and the SRTs for MTB separated (black bars) and SSN co-located (gray bars), respectively. All predictors were positively correlated with the SRTs measured in both conditions. However, the correlations were consistently higher for the MTB separated condition, which is in line with the larger across-listener SRT variability in the MTB separated condition demonstrated in Fig. 5. Furthermore, the correlations differed substantially across the different predictors, with the by far lowest values observed for Noisy$1_{c/o}$ (leftmost bars in Fig. 6), which is in line with the small across-listener variability shown in leftmost bar in Fig. 4. The highest correlations[2] for the two speech test conditions were observed for Tonal$_{2c/o}$ vs. MTB separated (r = 0.73) and Noisy-LP$_{2c/o}$ vs. SSN co-located (r = 0.59). Despite the applied audibility compensation, the 4-PTA also yielded high correlations with the SRTs (0.67 for MTB separated and 0.55

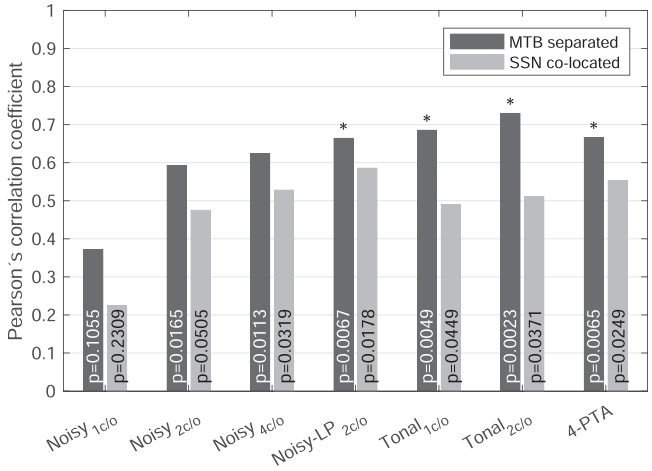

Fig. 6. Pearson's correlation coefficient between different potential predictors (six STM test variants and 4-PTA, from left to right) and SRTs measured in MTB separated (black bars) as well as SSN co-located conditions (gray bars). The p-values resulting from a test assessing positive correlation are shown in the corresponding bars. Significant correlations are indicated with * (p_{HB}<0.05 after correction for multiple comparisons).

for SSN co-located). A significance test with a Holm-Bonferroni correction for multiple comparisons revealed that the predictors Noisy-LP$_{2c/o}$, Tonal$_{1c/o}$, Tonal$_{2c/o}$, and 4-PTA showed a significant positive correlation with the SRTs obtained in the MTB separated condition (p<0.05, see asterisks in Fig 6). In contrast, none of the correlations between any of the predictors and the SSN co-located SRTs was significant, with many being just above the p_{HB} = 0.05 limit (after the Holm-Bonferroni correction).

Fig. 7 shows the relationships between the significant predictor-outcome pairs by means of scatter plots. Thus, MTB separated SRTs are shown as a function of STM thresholds measured with the Noisy-LP$_{2c/o}$ (top left panel), Tonal$_{1c/o}$ (top right panel), and Tonal$_{2c/o}$ (bottom left panel) stimulus variants, as well as in relation to the 4-PTA (bottom right panel). The individual data (black dots) are shown along with linear regression line fits (dashed lines)

[2] Note that the present study does not have the statistical power to determine whether these correlation coefficients were significantly different from the other observed correlation coefficients.

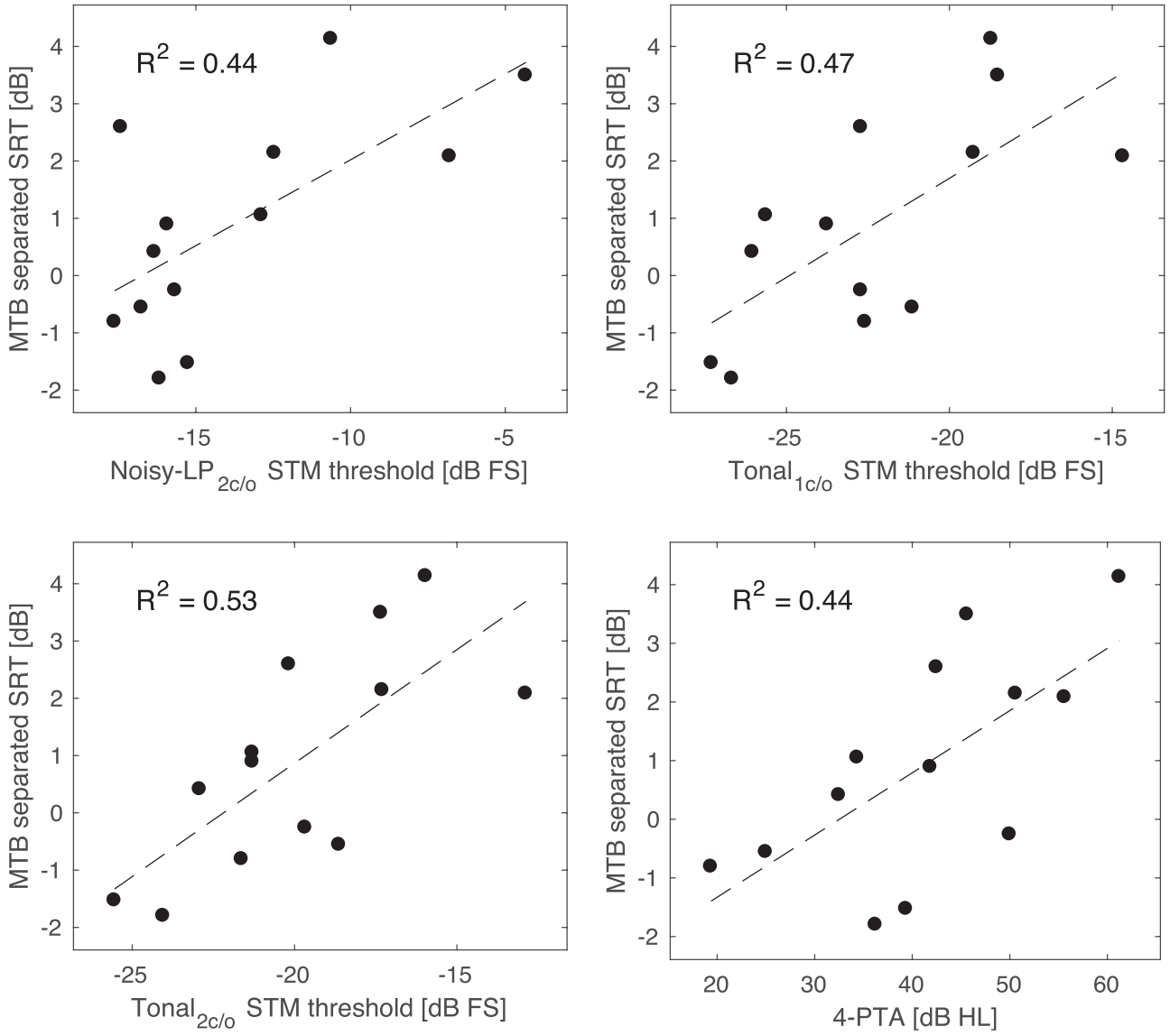

Fig. 7. Scatter plots showing the significant relationships identified in the analysis documented in Fig. 6, i.e., MTB separated SRTs as a function of predictors Noisy-$LP_{2c/o}$ (top left), $Tonal_{1c/o}$ (top right), $Tonal_{2c/o}$ (bottom left), and 4-PTA (bottom right). The data points are represented as black dots, the dashed lines show linear regression fits, and the corresponding R^2 values are shown in the top left corners of the scatter plots.

and the corresponding R^2 values (top left corners of the individual panels).

3.4. Multi-variate linear regression analysis

A multi-variate linear regression analysis was conducted to investigate the (relative) predictive power of the STM thresholds and the 4-PTA measure. Three different models with two predictors were generated, all having MTB separated SRTs as the outcome variable and 4-PTA as one of the predictors. The other predictor was the STM threshold measured with the Noisy-$LP_{2c/o}$, $Tonal_{1c/o}$, or $Tonal_{2c/o}$ STM variant, respectively. For reference, three different models were generated in the same fashion to predict the SSN co-located SRTs. Fig. 8 shows the results of the models in terms of percentage SRT variance explained (effectively R^2 multiplied by 100%) for MTB separated (left panel) and SSN co-located (right panel). In both panels, the dark gray bottom portions of the three bars represent SRT variance explained by the 4-PTA predictor alone, whereas the light gray top portions of the bars reflect the effect of adding the respective STM predictor to the model.

As can be seen in Fig. 8, the three models with 4-PTA and Noisy-$LP_{2c/o}$, $Tonal_{1c/o}$, or $Tonal_{2c/o}$ as predictors accounted for 55%, 59%, and 61% of the MTB separated SRT variance (left panel) and for 41%, 35%, and 35% of the SSN co-located SRT variance (right panel), respectively. The 4-PTA alone accounted for 44% of the MTB separated SRT variance and for 31% of the SSN co-located SRT variance (dark gray portions of the bars in Fig. 8). The two-predictor models thus explained a somewhat (but not significantly) larger percentage of the variance than the 4-PTA. The differences amounted to 11%, 14%, and 17% for Noisy-$LP_{2c/o}$, $Tonal_{1c/o}$, and $Tonal_{2c/o}$, respectively, in the case of the MTB separated SRTs (light gray portions of the bars in the left panel of Fig. 8) and to 10%, 5%, and 5% in the case of the SSN co-located SRTs (light gray portions of the bars in the right panel of Fig. 8). The STM predictors alone accounted for 44% (Noisy-$LP_{2c/o}$), 47% ($Tonal_{1c/o}$), and 53% ($Tonal_{2c/o}$) of the MTB separated SRT variance (not shown

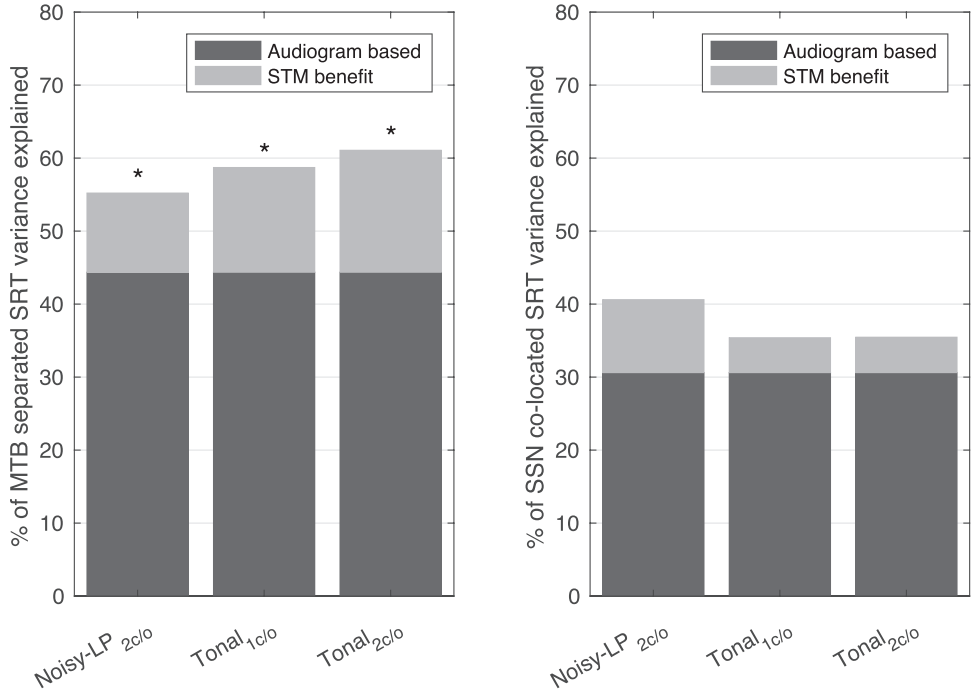

Fig. 8. Results of multi-variate linear regression analyses in terms of explained percentage of SRT variance for MTB separated condition (left panel) and SSN co-located condition (right panel). Both panels show the results obtained with three two-predictor models, each based on 4-PTA and one of the three selected STM predictors (see abscissa). The asterisks indicate the whole-model significance after Holm-Bonferroni correction for multiple comparisons (p_{HB}<0.05). Dark gray portions reflect the % variance explained by 4-PTA alone; the light gray portions show the additional % variance explained when adding the respective STM predictor.

here; cf. Fig. 7) and for 34% (Noisy-LP$_{2c/o}$), 24% (Tonal$_{1c/o}$), and 26% (Tonal$_{2c/o}$) of the SSN co-located SRT variance (not shown here).

While all three models yielded significant predictions of the MTB separated SRTs (p_{HB}<0.05 after Holm-Bonferroni correction for multiple comparisons), none of the individual predictors contributed a significant portion of variance explained in any of the three two-predictor models, indicating some level of collinearity between 4-PTA and each of the Noisy-LP$_{2c/o}$, Tonal$_{1c/o}$, and Tonal$_{2c/o}$ STM measures. In the case of the SSN co-located SRTs, none of the three models yielded significant predictions.

4. Discussion

4.1. Summary of main findings

The data obtained in the present study were analyzed with regard to (i) whether all listeners were able to complete the task, (ii) whether the STM thresholds were predictive of ecologically valid SRTs, and (iii) which STM stimulus variants were predictive of the SRTs. First, the proposed STM test paradigm did not result in any listeners having problems completing any adaptive threshold measurement (thus solving the problems observed by Bernstein et al., 2016), i.e., thresholds could be obtained for each adaptive track, even for the most challenging stimulus variant with a spectral modulation frequency of 4 c/o. Second, all considered STM variants yielded thresholds that were positively correlated with SRTs measured in the MTB separated condition (see Fig. 6), three of them significantly so.[3] This suggests that the previously reported predictive power of STM thresholds with respect to SRTs (Bernstein et al., 2013; Mehraei et al., 2014; Bernstein et al., 2016) generalizes to more realistic speech conditions. In fact, the STM thresholds appeared to be more strongly correlated[2] with the more ecologically valid SRTs measured in the MTB separated condition than with those measured in the more artificial SSN co-located condition. Third, three STM stimulus variants were identified as significant predictors of the MTB separated SRTs, namely Noisy-LP$_{2c/o}$, the low-pass filtered pink-noise based variant with 4 Hz and 2 c/o from Bernstein et al. (2016), as well as Tonal$_{1c/o}$ and Tonal$_{2c/o}$, the two complex-tone based novel variants introduced in the present study, with 4 Hz combined with 1 c/o and 2 c/o, respectively. They accounted for 44% (Noisy-LP$_{2c/o}$), 47% (Tonal$_{1c/o}$), and 53% (Tonal$_{2c/o}$) of the SRT variance, while the 4-PTA accounted for 44% of the SRT variance. The combination of 4-PTA and STM thresholds in a multi-variate linear regression model yielded 55% (Noisy-LP$_{2c/o}$ + 4-PTA), 59% (Tonal$_{1c/o}$ + 4-PTA), and 61% (Tonal$_{2c/o}$ + 4-PTA) of SRT variance explained (see Fig. 8).

4.2. STM detection measurement paradigm

The modifications of the measurement paradigm in relation to previous studies (described in detail in Fig. 2) were implemented with the aim to facilitate the adaptive STM threshold measurement, as Bernstein et al. (2016) reported that about one third of their 154 listeners (particularly the older and strongly HI ones) were not able to complete their adaptive STM threshold measurement. The listeners tested in the present study were, on average, older and had a more pronounced high-frequency average (HFA, 2–6 kHz) hearing loss as compared to Bernstein et al. (2016), who reported major problems with listeners older than 65 years and with an HFA greater than 53 dB HL. Twelve out of the 13 listeners tested here were older than 65 years, most of them substantially older (72.2 years on average). The average HFA was 56.8 dB HL and 10 out of the 13 listeners showed HFAs above 53 dB HL. Thus, most listeners tested in the present study belonged to the same

[3] Note that the predictive power of the STM thresholds beyond the audiogram reported previously could not be statistically proven here given the limited sample size.

age- and HFA-group in which Bernstein et al. (2016) found that the majority of listeners were not able to obtain an STM threshold in their adaptive measurement procedure and in which no significant relationship between STM thresholds and SRTs could be established (potentially because many STM thresholds had to be obtained using a combination of constant-stimuli procedure and extrapolation). The results of the present study suggest that the adaptation of the measurement paradigm was effective in facilitating the STM detection task, as all listeners were able to complete all adaptive threshold measurements, irrespective of the considered STM stimulus variant. For the Noisy-$LP_{2c/o}$ stimulus variant considered by Bernstein et al. (2016), the highest threshold measured in the present study was about −4 dB FS, i.e., well below full modulation (0 dB FS). The overall highest measured threshold was −2.5 dB FS and occurred for the $Noisy_{4c/o}$ stimulus variant. As all changes introduced to the measurement paradigm were done simultaneously, it is impossible to quantify the relative contributions of the different aspects that were modified. Based on informal observations made in the context of the study, it is assumed that the crucial factors were (i) the use of 3-AFC, which allows to detect the odd interval out of three intervals instead of having to detect a learned target modulation pattern in a 2-AFC paradigm (Bernstein et al., 2013; Mehraei et al., 2014; Bernstein et al., 2016), (ii) the individualized linear hearing loss compensation that ensures audibility across the relevant frequency range without stimulating at unnecessarily high levels that may deteriorate STM detection performance (Magits et al., 2019), (iii) the long interval duration of one second, providing twice the observation time window as compared to Bernstein et al. (2016), and (iv) the binaural presentation of the stimuli, allowing listeners to use their better ear as well as to benefit from binaural summation.

Further insights may be gained via comparison with a related recent study. Souza et al. (2020) tested STM detection in 30 HI listeners using a somewhat different approach for stimulus generation with a logarithmic modulation depth definition (Isarangura et al., 2019a) and broadband noise carriers (0.4 – 8 kHz) combined with a 4-interval 2-cue 2-AFC paradigm that also allows for detection of the odd one out. Each interval was 500 ms long and the broadband level was set at 30 dB above the individual carrier-noise detection threshold in the test ear (monaural presentation). The measured STM thresholds were between −18 and −1.7 dB FS, with an average of −7.3 dB FS (after transforming their results to the modulation depth definition used in the present study based on Isarangura et al., 2019a). The basic stimulus design was quite similar to the $Noisy_{2c/o}$ variant considered in the present study, which here resulted in substantially lower STM thresholds between -22.1 and −10.4 dB FS, with an average of −16.2 dB FS (see Fig. 4). With the above-mentioned results of Bernstein et al. (2016) and their experimental design (2-AFC, 500 ms interval duration, band-limited stimulus) in mind, the comparison between the present study and Souza et al. (2020) thus suggests that (i) the ability to use an odd-one-out detection approach and/or a large bandwidth of the stimuli solves the issue of elderly HI listeners not being able to obtain robust STM detection thresholds, (ii) the use of individualized frequency-dependent (instead of broadband) amplification and/or the use of 1-second long (instead of 0.5 s) intervals and/or binaural (instead of monaural) stimulus presentation yield substantially lower STM detection thresholds in elderly HI listeners. It should be noted that a study by Isarangura et al. (2019b) suggested no effects of interval duration and level (in the relevant range) on spectral modulation detection; however, it is unclear how these results relate to spectro-temporal modulation detection, where additional temporal cues are involved that may well be affected by the interval duration.

4.3. Predicting ecologically valid SRTs

The speech test used in the present study was designed with the aim to simulate real-life conditions and supra-threshold speech reception. Therefore, the test was set up in a slightly reverberant room and multiple spatially distributed speech interferers were used (MTB separated condition), as this resembles a “cocktail-party” type scenario that HI listeners are known to especially struggle with. Furthermore, the HINT speech corpus consisting of open-set everyday sentences was used, which yields relatively high SRTs such that listeners operate at SNR levels they may also encounter in everyday life (Smeds et al., 2015), as opposed to closed-set matrix sentence speech tests that typically yield lower SRTs (cf. Nielsen and Dau, 2011 vs. Wagener et al., 2003). The SSN co-located condition was additionally considered as a reference condition containing fewer aspects of real-life speech perception. Linear amplification according to the “sufficiently audible” approach by Humes (2007) was applied to ensure that target-speech audibility was not compromised such that supra-threshold speech reception could indeed be measured. The resulting SRTs suggest that while the group average SRTs were only slightly (yet significantly) elevated for the MTB separated as compared to the SSN co-located condition, the MTB separated condition induced a significantly larger across-listener variability in SRTs than the SSN co-located condition (almost doubling the standard deviation). Note that the larger variability in the MTB separated condition is not ascribable to higher test-retest variability. Indeed, Rønne et al. (2017) found very similar test-retest variability between their HINT co-located SSN condition and an MTB separated condition similar to that used in the present study. Additionally, a test-retest standard deviation tentatively computed for the MTB separated condition using the second training run as “test” amounted to a value of 0.54 dB, which is well below the normative test-retest standard deviation of the Danish HINT measured for SSN (Nielsen and Dau, 2011). The more realistic speech test thus revealed larger and likely more representative and reproducible performance differences across listeners. While the SRTs measured in the two conditions were highly correlated across listeners (70.5% of shared variance, see Fig. 5), this correlation was strongly driven by the four listeners with the poorest performance, whereas the relationship was essentially non-existent for the remaining nine listeners that showed very similar performance levels in SSN co-located (within a range of 1.5 dB) but exhibited substantial performance differences in MTB separated (within a range of 4 dB, see Fig. 5 for details). This is consistent with the SRT relationship induced by artificial vs. more realistic speech material reported by Carlile & Keidser (2020). The STM thresholds measured with all considered stimulus variants showed positive correlations with SRTs and higher correlations with the MTB separated SRTs than with the SSN co-located SRTs, although it should be noted that these differences in correlation were not significant and that the predictive power of the STM thresholds was not significantly different than that of the 4-PTA, likely due to the small number of subjects. The predictive power of STM detection thresholds with regard to speech reception outcomes obtained in a cocktail-party type speech-on-speech scenario with individualized amplification represents an extension of the findings from previous studies: in contrast to the present study, the previous studies presented speech via headphones/earphones using SSN (Bernstein et al., 2013), SSN and modulated noise (Mehraei et al., 2014), or SSN and four-talker babble (Bernstein et al., 2016); furthermore, they used either open-set sentences but unrealistically high broadband presentation levels without individualized amplification (Bernstein et al., 2013; Mehraei et al., 2014) or realistic presentation levels and individualized amplification but closed-set matrix sentences (Bernstein et al., 2016).

4.4. STM variants and their predictive power

The six STM stimulus variants considered in the present study (see Fig. 2) differed in terms of carrier bandwidth and type (noise or complex tone) and regarding the spectral modulation frequency, but all shared the same temporal modulation frequency of 4 Hz (upward moving). This was motivated by the results of Bernstein et al. (2013), who reported the largest differences between NH and HI listeners for 4 Hz, and by the fact that 4 Hz is very close to the syllable rate of most languages and thus has a high relevance in speech reception. The upward moving direction was chosen to further facilitate STM detection, as it has been reported to yield somewhat lower thresholds than downward (e.g., Chi et al., 1999). Regarding the spectral modulation frequencies, the previous studies found 2 c/o to result in the most promising results. However, 1 c/o was also considered here as part of the effort to simplify the test, expecting overall lower thresholds (as demonstrated by Bernstein et al., 2013). 4 c/o was considered because Bernstein et al. (2013) showed a relatively large difference between NH and HI listeners for this variant. The novel complex-tone based stimulus variants represent an attempt to increase the similarity between the STM stimulus and natural speech by mimicking the mostly periodic carrier signal of the voiced parts of speech.

Consistent with earlier studies (e.g., Bernstein et al., 2013), STM thresholds increased with increasing spectral modulation frequency. The complex-tone carrier led to lower thresholds as compared to the noise carrier, facilitating STM detection. This may be explained by the fact that the complex-tone carrier is free of low-frequency intrinsic fluctuations, which have been shown to mask an imposed temporal modulation when using noise carriers (e.g., Dau et al., 1999). The across-listener variability was large for all STM stimulus variants (standard deviation 3–4 dB), with the notable exception of $\text{Noisy}_{1\text{c/o}}$, which showed a significantly smaller variability (standard deviation of 1.4 dB). As mentioned above, STM thresholds appeared to be generally more (albeit not significantly more) highly correlated with the MTB separated SRTs than with the SSN co-located SRTs. The STM variant showing the by far lowest (albeit not significantly different) correlations with the SRTs was $\text{Noisy}_{1\text{c/o}}$. Thus, the small across-listener variability observed in the SSN co-located SRTs and in the $\text{Noisy}_{1\text{c/o}}$ STM thresholds both translated to low correlations. Significant correlations were found between MTB separated SRTs and the three STM variants $\text{Noisy-LP}_{2\text{c/o}}$, $\text{Tonal}_{1\text{c/o}}$, and $\text{Tonal}_{2\text{c/o}}$, with 44%, 47%, and 53% of shared variance, respectively. The fact that the complex-tone based variants, in particular $\text{Tonal}_{2\text{c/o}}$, showed strong predictions of SRTs is well in line with recent research on the importance of periodic carrier signals in the healthy and impaired auditory system (Steinmetzger and Rosen, 2015; Carney et al., 2015; Carney, 2018; Steinmetzger et al., 2019). The SSN co-located SRTs were best predicted by the $\text{Noisy-LP}_{2\text{c/o}}$ STM variant (although not significant), which might indicate that different abilities were required in the two speech test conditions.

Although the speech test conditions as well as the STM test variants differ substantially, the results of the present study and the previous studies are compared here. Bernstein et al. (2013) reported 61% of speech outcome (percent correct at 0 dB SNR) variance explained by their STM measure based on 12 HI listeners. Bernstein et al. (2016) showed that 28% of the SRT variance was explained by STM thresholds, based on 154 HI listeners. The best STM variant considered in the present study accounted for 53% of MTB separated SRT variance based on 13 HI listeners, and the $\text{Noisy-LP2}_{\text{c/o}}$ variant used in Bernstein et al. (2016) still accounted for 44%. While this was not the main focus here, it should still be noted that Bernstein et al. (2013) and Bernstein et al. (2016) found that STM thresholds added significant predictive power in relation to their speech outcomes beyond the audiogram, which could not be replicated in the present study (likely due to the limited number of observations). The results from Bernstein et al. (2016) may be considered the more relevant reference for the present study due to their use of individualized amplification in the speech test. The stronger relationship between SRTs and STM thresholds found in the present study may be related to the fact that all thresholds could be obtained adaptively whereas in Bernstein et al. (2016) about a third of their population was not able to adaptively obtain thresholds and the results were instead obtained in a constant stimulus measurement, potentially at the expense of measurement accuracy. Furthermore, Bernstein et al. (2016) used averaged SRTs measured using a matrix-sentence test in different noise conditions and with different types of simulated hearing-aid processing, whereas the current study focused on more ecologically valid SRTs obtained with open-set sentences, spatialized speech interferers, and non-compromising linear amplification.

4.5. The relationship between SRTs and the audiogram

Despite the respective efforts to provide audible stimuli, an audiogram-based predictor also accounted for a good portion of the SRT variance in the previous studies. In Bernstein et al. (2013), the audiogram-based SII predicted 40% of the variance of percent words correct collected in SSN at 0 dB SNR. However, this number increased dramatically when considering SRTs instead of percent correct at a selected SNR using the same underlying data set (60% for SSN and 72% for modulated noise, see Mehraei et al., 2014). In Bernstein et al. (2016), the high-frequency average pure-tone thresholds accounted for 31% of SRT variance, whereas the low-frequency thresholds did not predict SRTs. Similarly, in the present study, the 4-PTA (i.e., the average across pure-tone thresholds at 0.5, 1, 2, and 4 kHz) accounted for 44% of the MTB separated SRT variance. The 4-PTA was chosen because it is assumed to be related to speech audibility due to the considered frequency range, and it was indeed a substantially stronger SRT predictor than the average across all pure-tone thresholds (not shown here), i.e., the low- and mid-frequency thresholds were most predictive of SRTs, in contrast to Bernstein et al. (2016). The fact that the pure-tone thresholds yield relatively strong correlations with supra-threshold speech reception measures is likely due to their relationship with supra-threshold deficits, such as (but not limited to) reduced frequency resolution and loudness recruitment (see also Sanchez-Lopez et al., 2020). However, the disparity as to which pure-tone threshold-based measures predict SRTs in the previous and current studies emphasizes the added value of the STM test result as an “explicit” supra-threshold measure. Furthermore, the best STM predictor yielded a substantially (yet not significantly) higher correlation with the SRTs as compared to the audiogram-based predictor in the present study.

4.6. Limitations and outlook

The present study was designed as an explorative study to address the issues discussed above. Thus, the selection of the considered stimulus parameters was not balanced across different carrier types (e.g., there was a 4-c/o stimulus variant with a noise carrier but not with a complex-tone carrier), which is suboptimal for a group-level analysis of the relevant factors such as carrier type and spectral modulation frequency. Furthermore, due to the fact that the speech-test design required a high level of audiogram symmetry, the number of listeners was somewhat limited (13), such that the results in terms of percent of SRT variance explained are added and discussed with caution here. Nevertheless, the STM-vs-SRT correlation analyses indicated significant results, but it is likely that a larger population would result in

more STM variants yielding significant predictions of SRTs. Furthermore, the question whether the STM thresholds yield a significant amount of SRT variance explained in addition to the audiogram-based predictor is still open and likely requires a larger population. Lastly, it is unclear how the results collected with the described ear-independent sufficiently audible (linear) amplification approach generalize to speech-test performance with hearing aids, which act as ear-specific non-linear amplifiers.

An interesting next step would therefore be a study with a substantially larger number of listeners representing a clinically more robust sample (including asymmetric hearing losses) that considers a speech test with hearing aids, allowing for the effect of realistic ear-specific non-linear amplification and potentially other hearing-aid processing algorithms to be tested. This design would also imply that the amplification provided in the STM test should not be based on the average hearing loss across the two ears, but rather on the ear-specific hearing loss.

5. Conclusion

The present study investigated the suitability of a spectro-temporal modulation (STM) detection measurement paradigm with individualized audibility compensation, focusing on its clinical viability and relevance as a real-life supra-threshold speech reception predictor. Based on data measured in 13 elderly hearing-impaired listeners, two novel complex-tone based and four previously established noise-based STM stimulus variants were considered and compared to ecologically valid speech reception thresholds (SRTs) measured in a speech-on-speech spatialized set-up with linear amplification based on the individual's audiogram. The study had three main objectives: (i) to adapt the STM test such that all listeners in an elderly hearing-impaired population can complete the test, (ii) to investigate whether the previously established predictive power of STM thresholds with regard to SRTs can be generalized to more ecologically valid SRTs, and (iii) to determine which STM stimulus variants were predictive of ecologically valid SRTs.

First, the results of the study demonstrated that all listeners were able to complete all STM threshold measurements, indicating that the proposed STM detection measurement paradigm indeed facilitated the task. Second, the correlations between STM thresholds and the ecologically valid SRTs reached statistical significance whereas those obtained for standard SRTs collected in stationary co-located noise did not. Third, three STM stimulus variants (one noise-carrier based and two complex-tone based) yielded significant predictions of SRTs, accounting for up to 53% of the SRT variance. The results of this study may motivate and inform future research toward a clinically applicable STM test, which may be useful for quantifying supra-threshold speech reception deficits in aided hearing-impaired listeners - without requiring time-consuming and language-specific speech testing - and thus enable a more individualized prescription of hearing-aid processing in everyday clinical practice.

Data Availability

Data will be made available on request.

CRediT authorship contribution statement

Johannes Zaar: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Lisbeth Birkelund Simonsen:** Investigation, Writing – review & editing. **Torsten Dau:** Resources, Writing – review & editing. **Søren Laugesen:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

We would like to acknowledge the crucial contributions from Thomas Behrens and James Harte in establishing and supporting the project. We thank Laurel Carney, Golbarg Mehraei, Thomas Lunner, Elaine Ng, Alejandro Lopez Valdes, Gary Jones, Nicolas Le Goff, and Raul Sanchez-Lopez for their scientific contributions in support of this study.

Funding

This study was funded by the Oticon Foundation [grant number 17-0639] and by the Swedish Research Council [grant number 2017-06092].

References

Bernstein, J.G.W., Mehraei, G., Shamma, S., Gallun, F.J., Theodoroff, S.M., Leek, M.R., 2013. Spectrotemporal modulation sensitivity as a predictor of speech intelligibility for hearing-impaired listeners. J. Am. Acad. Audiol. 124 (4), 293–306. doi:10.3766/jaaa.24.4.5.

Bernstein, J.G.W., Danielsson, H., Hällgren, M., Stenfelt, S., Rönnberg, J., Lunner, T., 2016. Spectrotemporal modulation sensitivity as a predictor of speech-reception performance in noise with hearing aids. Trends Hear. 20, 1–17. doi:10.1177/2331216516670387.

Carlile, S., Keidser, G., 2020. Conversational interaction is the brain in action: implications for the evaluation of hearing and hearing interventions. Ear Hear. 41, 56S–67S. doi:10.1097/AUD.0000000000000939.

Carney, L.H., Li, T., McDonough, J.M., 2015. Speech coding in the brain: representation of vowel formants by midbrain neurons tuned to sound fluctuations. eNeuro 2 (4), 2–12. doi:10.1523/ENEURO.0004-15.2015.

Carney, L.H., 2018. Supra-threshold hearing and fluctuation profiles: implications for sensorineural and hidden hearing loss. J. Assoc. Res. Otolaryngol. 19 (4), 331–352. doi:10.1007/s10162-018-0669-5.

Chi, T., Gao, Y., Guyton, M.C., Ru, P., Shamma, S., 1999. Spectro-temporal modulation transfer functions and speech intelligibility. J. Acoust. Soc. Am. 106 (5), 2719–2732. doi:10.1121/1.428100.

Dau, T., Verhey, J.L., Kohlrausch, A., 1999. Intrinsic envelope fluctuations and modulation-detection thresholds for narrowband noise carriers. J. Acoust. Soc. Am. 106 (5), 2752–2760. doi:10.1121/1.428103.

Elhilali, M., Chi, T., Shamma, S., 2003. A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. Speech Commun. 41 (2–3), 331–348. doi:10.1016/S0167-6393(02)00134-6.

Ewert, S.D., 2013. AFC—a modular framework for running psychoacoustic experiments and computational perception models. In: Proceedings of the International Conference on Acoustics AIA-DAGA 2013, Merano, Italy, pp. 1326–1329.

Humes, L.E., 2007. The contributions of audibility and cognitive factors to the benefit provided by amplified speech to older adults. J. Am. Acad. Audiol. 18, 590–603. doi:10.3766/jaaa.18.7.6.

IEC, 1985. 268-13, Sound System Equipment Part 13: Listening Tests on Loudspeaker. International Electrotechnical Commission, Geneva, Switzerland.

Isarangura, S., Palandrani, K.N., Stavropoulos, T., Seitz, A., Hoover, E.C., Gallun, F.J., Eddins, D.A., 2019a. Methods for expressing spectral modulation depth and the effects of modulator shape on spectral modulation detection thresholds. Proc. Meet. Acoust. 36, 050003. doi:10.1121/2.0001032, 2019.

Isarangura, S., Eddins, A.C., Ozmeral, E.J., Eddins, D.A., 2019b. The effects of duration and level on spectral modulation perception. J. Speech Lang. Hear. Res. 62 (10), 3876–3886. doi:10.1044/2019_JSLHR-H-18-0449.

Jerger, J.C., 2018. The evolution of the audiometric pure-tone technique. Hear. Rev. 25 (9), 12–18.

Johannesen, P.T., Pérez-González, P., Lopez-Poveda, E.A., 2014. Across-frequency behavioral estimates of the contribution of inner and outer hair cell dysfunction to individualized audiometric loss. Front. Neurosci. 8. doi:10.3389/fnins.2014.00214, 2014.

Keidser, G., Naylor, G., Brungart, D.S., Caduff, A., Campos, J., Carlile, S., Carpenter, M.G., Grimm, G., Hohmann, V., Holube, I., Launer, S., Lunner, T., Mehra, R., Rapport, F., Slaney, M., Smeds, K., 2020. The quest for ecological validity in hearing science: what it is, why it matters, and how to advance it. Ear Hear. 41, 5S–19S. doi:10.1097/AUD.0000000000000944.

Levitt, H., 1971. Transformed up-down methods in psychoacoustics. J. Acoust. Soc. Am. 49 (2B), 467–477. doi:10.1121/1.1912375.

Lopez-Poveda, E.A., 2014. Why do I hear but not understand? Stochastic undersampling as a model of degraded neural encoding of speech. Front. Neurosci. 8, 348. doi:10.3389/fnins.2014.00348, Article.

Magits, S., Moncada-Torres, A., Van Deun, L., Wouters, J., van Wieringen, A., Francart, T., 2019. The effect of presentation level on spectrotemporal modulation detection. Hear. Res. 371, 11–18. doi:10.1016/j.heares.2018.10.017.

Mehraei, G., Gallun, F.J., Leek, M.R., Bernstein, J.G.W., 2014. Spectro-temporal modulation sensitivity for hearing-impaired listeners: dependence on carrier center frequency and the relationship to speech intelligibility. J. Acoust. Soc. Am. 136 (1), 301–316. doi:10.1121/1.4881918.

Moore, B.C., Stone, M.A., Füllgrabe, C., Glasberg, B.R., Puria, S., 2008. Spectro-temporal characteristics of speech at high frequencies, and the potential for restoration of audibility to people with mild-to-moderate hearing loss. Ear Hear. 29 (6), 907–922. doi:10.1097/AUD.0b013e31818246f6.

Nielsen, J.B., Dau, T., 2011. The Danish hearing in noise test. Int. J. Audiol. 50 (3), 202–208. doi:10.3109/14992027.2010.524254.

Pavlovic, Chaslav V., 1987. Derivation of primary parameters and procedures for use in speech intelligibility predictions. J. Acoust. Soc. Am. 82. doi:10.1121/1.395442.

Plomp, R., 1986. A signal-to-noise ratio model for the speech reception threshold of the hearing impaired. J. Speech Hear. Res. 29 (2), 146–154. doi:10.1044/jshr.2902.146.

Rønne, F.M., Laugesen, S., Jensen, N.S., 2017. Selection of test-setup parameters to target specific signal-to-noise regions in speech-on-speech intelligibility testing. Int. J. Audiol. 56 (8), 559–567. doi:10.1080/14992027.2017.1300349.

Sanchez-Lopez, R., Fereczkowski, M., Neher, T., Santurette, S., Dau, T., 2020. Robust data-driven auditory profiling towards precision audiology. Trends Hear. 24, 1–19. doi:10.1177/2331216520973539.

Smeds, K., Wolters, F., Rung, M., 2015. Estimation of signal-to-noise ratios in realistic sound scenarios. J. Am. Acad. Audiol. 26, 183–196. doi:10.3766/jaaa.26.2.7.

Souza, P., Gallun, F.J., Wright, R., 2020. Contributions to speech-cue weighting in older adults with impaired hearing. J. Speech Lang. Hear. Res. 63 (1), 334–344. doi:10.1044/2019_JSLHR-19-00176.

Steinmetzger, K., Rosen, S., 2015. The role of periodicity in perceiving speech in quiet and in background noise. J. Acoust. Soc. Am. 138, 3586–3599. doi:10.1121/1.4936945.

Steinmetzger, K., Zaar, J., Relaño-Iborra, H., Rosen, S., Dau, T., 2019. Predicting the effects of periodicity on the intelligibility of masked speech: an evaluation of different modelling approaches and their limitations. J. Acoust. Soc. Am. 146, 2562–2576. doi:10.1121/1.5129050.

Strelcyk, O., Dau, T., 2009. Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. J. Acoust. Soc. Am. 125 (5), 3328–3345. doi:10.1121/1.3097469.

Thorup, N., Santurette, S., Jørgensen, S., Kjærbøl, E., Dau, T., Friis, M., 2016. Auditory profiling and hearing-aid satisfaction in hearing-aid candidates. Dan. Med. J. 63 (10), 27697129.

Wagener, K., Josvassen, J.L., Ardenkjær, R., 2003. Design, optimization and evaluation of a Danish sentence test in noise: diseño, optimización y evaluación de la prueba Danesa de frases en ruido. Int. J. Audiol. 42 (1), 10–17. doi:10.3109/14992020309056080.