A spectro-temporal modulation test for predicting speech reception in hearing-impaired listeners with hearing aids

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

Published in:
Hearing Research

Link to article, DOI:
10.1016/j.heares.2024.108949

Publication date:
2024

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

Hearing Research 443 (2024) 108949

Contents lists available at ScienceDirect

Hearing Research

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

A spectro-temporal modulation test for predicting speech reception in hearing-impaired listeners with hearing aids

Johannes Zaar [a,b,*], Lisbeth Birkelund Simonsen [b,c], 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

Keywords:
Speech intelligibility
Psychoacoustics
Noise
Clinical test
Noise reduction
Subjective preference

ABSTRACT

Spectro-temporal modulation (STM) detection sensitivity has been shown to be associated with speech-in-noise reception in hearing-impaired (HI) individuals. Based on previous research, a recent study [Zaar, Simonsen, Dau, and Laugesen (2023). Hear Res 427:108650] introduced an STM test paradigm with audibility compensation, employing STM stimulus variants using noise and complex tones as carrier signals. The study demonstrated that the test was suitable for the target population of elderly individuals with moderate-to-severe hearing loss and showed promising predictions of speech-reception thresholds (SRTs) measured in a realistic set up with spatially distributed speech and noise maskers and linear audibility compensation. The present study further investigated the suggested STM test with respect to (i) test-retest variability for the most promising STM stimulus variants, (ii) its predictive power with respect to realistic speech-in-noise reception with non-linear hearing-aid amplification, (iii) its connection to effects of directionality and noise reduction (DIR+NR) hearing-aid processing, and (iv) its relation to DIR+NR preference. Thirty elderly HI participants were tested in a combined laboratory and field study, collecting STM thresholds with a complex-tone based and a noise-based STM stimulus design, SRTs with spatially distributed speech and noise maskers using hearing aids with non-linear amplification and two different levels of DIR+NR, as well as subjective reports and preference ratings obtained in two field periods with the two DIR+NR hearing-aid settings. The results indicate that the noise-carrier based STM test variant (i) showed optimal test-retest properties, (ii) yielded a highly significant correlation with SRTs (R^2=0.61) exceeding and complementing the predictive power of the audiogram, (iii) yielded significant correlation (R^2=0.51) with the DIR+NR-induced SRT benefit, and (iv) did not provide significant correlation with subjective preference for DIR+NR settings in the field. Overall, the suggested STM test represents a valuable tool for diagnosing speech-reception problems that remain when hearing-aid amplification has been provided and the resulting need for and benefit from DIR+NR hearing-aid processing.

1. Introduction

Hearing loss is currently defined, in a clinical sense, using the pure-tone audiogram, which describes the sound level needed for detection of pure tones at a range of frequencies (cf. Jerger, 2018). While this definition may lead to interpreting hearing loss as a simple frequency-specific reduction of sound-level sensitivity, it is well-known that reduced audibility is not the only problem that elderly hearing-impaired (HI) individuals face, especially when listening to speech in noise (cf. Plomp, 1986; Lopez-Poveda, 2014). The sound discrimination challenges that remain when audibility has been restored by means of amplification, or even exist in older clinically normal-hearing individuals (cf. Regev et al., 2023a, Regev et al., 2023c), are here referred to as supra-threshold hearing deficits.

Conceptually, Plomp (1986) introduced the idea of a "distortion component" in hearing loss, which manifests itself in an effective reduction in the signal-to-noise ratio (SNR) when listening to speech in noise and can therefore not be mitigated by means of amplifying both speech and noise (leaving the SNR unchanged). A number of psycho-acoustic studies have investigated supra-threshold hearing deficits

* Corresponding author.
E-mail address: jozr@eriksholm.com (J. Zaar).

https://doi.org/10.1016/j.heares.2024.108949
Received 5 August 2023; Received in revised form 15 December 2023; Accepted 3 January 2024
Available online 4 January 2024

focusing mainly on aspects of temporal processing, loss of compression, and reduced frequency selectivity (e.g., Strelcyk and Dau, 2009; Johannesen et al., 2014; Thorup et al., 2016; Regev et al., 2023a, 2023b, 2023c). Recently, a series of studies centering around auditory profiling have suggested a comprehensive battery of auditory tests focusing on various aspects of auditory processing (Sanchez-Lopez et al., 2020; 2021), with one important data-driven finding being that the audiogram may contain more information about supra-threshold hearing deficits than traditionally assumed. Despite the audiogram being fundamentally a measure of pure-tone sensitivity, this indicates that the pure-tone thresholds may nonetheless be connected to other, supra-threshold, aspects of hearing loss.

Hearing-aid (HA) processing is typically adjusted on the basis of the audiogram, where the pure-tone thresholds are translated into a frequency-specific amplification, typically showing a level-dependent (compressive) behavior to account for the abnormal loudness growth resulting from loss of compression in HI individuals (e.g., Keidser et al., 2011). In addition, modern HAs provide a range of advanced sound processing capabilities that go far beyond amplification. Several such advanced HA processing approaches are designed to enhance the SNR, e.g., by selectively amplifying sounds from certain (typically frontal) directions while attenuating other sounds or by applying noise-reduction processing. Recently, directionality and noise reduction (DIR+NR) approaches have increasingly been combined, yielding very powerful SNR enhancement under certain assumptions of directionality (Le Goff et al., 2016; Andersen et al., 2021). Since DIR+NR processing not only improves the SNR but can also limit access to spatial cues and the overall perceived naturalness of the sound scene, it should be provided only to the extent that a given user needs it. However, in contrast to the audiogram-based amplification rationales, there is currently no established connection between a diagnostic measure and the optimal level of such types of advanced HA processing for the individual.

It can be argued that the most reasonable diagnostic to base the prescription/personalization of DIR+NR settings on is performance in a realistic speech-in-noise test, with HA amplification provided (i.e., supra-threshold). However, it appears unrealistic to obtain such a measure as part of a regular clinical workflow due to limited resources. An auditory test that is easy to measure and associated with aided speech-in-noise reception could thus be highly useful as a clinical proxy measure for speech-reception deficits and, by extension, for the need for advanced HA processing. The audiogram has been shown to be limited in predicting speech reception in noise when both speech and noise are audible (Vermiglio et al., 2020; 2022), such that a dedicated supra-threshold auditory test, complementary to the audiogram, would be required for the described purpose.

An excellent candidate for such a test can be found in the family of spectro-temporal modulation (STM) detection tests. First introduced by Chi et al. (1999), the STM stimulus typically consists of a broadband carrier noise that is modulated by a moving-ripple pattern with a spectral modulation along the logarithmic frequency axis, measured in cycles/octave (c/o), and a temporal modulation measured in cycles/-second (Hz). STM sensitivity is typically measured in an adaptive tracking procedure with the modulation depth of the STM pattern serving as the tracking variable. While the initial study by Chi et al. (1999) focused on normal-hearing (NH) listeners, Bernstein et al. (2013) systematically compared NH and HI listeners in terms of their STM detection performance for multiple spectral and temporal modulation frequencies and identified the combination of 2 c/o and 4 Hz as the most promising stimulus variant as it yielded the greatest performance difference between the NH and HI groups. They demonstrated a very strong relationship between STM thresholds and speech-in-noise scores measured with stationary speech-shaped noise at a very high sound pressure level (SPL) of 92 dB. Moreover, the STM thresholds (R^2=0.61) were found to be both superior and complementary to an audiogram-based predictor (R^2=0.4) of speech intelligibility. In a follow-up study, Bernstein et al. (2016) assessed the predictive power of their STM test in a large (n = 154) population of HI listeners, using a band-limited noise carrier (354–2000 Hz) and the 2-c/o and 4-Hz STM pattern. STM thresholds were compared to the average across multiple speech-reception thresholds (SRTs), measured in different noise conditions (speech-shaped noise and 4-talker babble) with individualized amplification and different simulated HA processing conditions (linear, noise reduction, fast compression). In that study, the connection between STM thresholds and SRTs was less pronounced (R^2=0.28) as compared to Bernstein et al. (2013), and the audiogram-based predictor yielded a slightly stronger relationship (R^2=0.31) with SRTs than the STM thresholds, while both predictors remained complementary (R^2=0.44 when combined). An additional analysis revealed that the STM test did not provide significant predictive power for the oldest (>65 years) and most severely hearing-impaired subset of the population. This was potentially related to the difficulty of the STM task, as about a third of the population could not obtain a threshold in the adaptive tracking procedure and their threshold was instead inferred via extrapolation from a percent-correct score obtained in a full-modulation vs. no-modulation procedure.

In our previous study (Zaar et al., 2023), we explored a number of STM stimulus variants, including two novel variants employing a complex-tone carrier, using a modified measurement paradigm and individualized audibility compensation. The main goals were to facilitate the test such that all populations of interest (including elderly and severely HI individuals) could do the task and to investigate to what extent STM sensitivity was related to speech-in-noise performance in realistic settings (including spatial cues, speech interferers, and non-compromising audibility compensation). The results indicate that (i) all tested elderly HI listeners could do the task with the modified measurement paradigm and audibility compensation, (ii) the noise-carrier based STM stimulus used by Bernstein et al. (2016) and a novel complex-tone based stimulus (both with 2 c/o and 4 Hz) were similarly promising SRT predictors, and (iii) SRTs measured in a realistic setting with spatialized speech and noise maskers and linear individualized amplification were significantly correlated with STM thresholds. However, due to its explorative nature and small sample size, the Zaar et al. (2023) study left several questions open as it did not establish which of the STM stimulus variants yielded the best properties with regard to test-retest reliability, and was inconclusive with respect to the significance of the relative contributions of the audiogram and the STM thresholds to predicting SRTs. Furthermore, the study did not address any questions related to HA processing, instead making use of a linear amplification approach optimized for controlled laboratory settings.

The present study was thus designed to further evaluate the suggested STM test with respect to (i) test-retest variability for the noise-carrier and tone-carrier based STM stimulus variants, (ii) its predictive power with respect to realistic speech-in-noise reception with non-linear HA amplification, (iii) its connection to effects of DIR+NR HA processing, and (iv) its relation to DIR+NR preference in the field. The overarching hypotheses were that subjects with poor aided speech reception would (a) show poor STM sensitivity, (b) benefit most from DIR+NR, and therefore (c) prefer strong DIR+NR (and vice versa). The research questions and hypotheses were addressed by means of a combined laboratory and field study with 30 elderly HI participants, collecting STM thresholds with a complex-tone based and a noise-based STM stimulus design, SRTs with spatially distributed speech and noise maskers using HAs with amplification and two different levels of DIR+NR, as well as subjective reports and preference ratings obtained in two field periods with the two DIR+NR HA settings. The data were analyzed using standard test-retest assessment methods, group-level statistical analyses, as well as correlation and regression analyses.

2. Method

2.1. Participants and hearing-aid settings

Thirty HI participants (9 female) aged between 45 and 81 years (mean: 70 years; standard deviation: 9 years) were recruited. Twenty of them were registered in the test-person database of the Hearing Systems Section at the Technical University of Denmark. Ten of the participants were newly recruited after the project had been advertised at Special-center Roskilde Kommunikation, which offers services for HA users who are in need of support and counselling beyond that offered by the regular HA clinics in Denmark. All participants were native speakers of Danish and regular HA users. 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 participants exhibited a range of mild to severe/profound hearing losses, which were largely symmetric: 29 participants showed thresholds within 15 dB between ears for a minimum of 9 out of the 11 test frequencies, one participant for 7 of the 11 test frequencies. In addition to pure-tone sensitivity, working memory capacity was tested using the reversed digit span (RDS) test (Blackburn and Benton, 1957), which measures the ability to repeat sequences of Danish digits in reverse order. The RDS test was administered using binaural presentation of spoken digits via headphones at self-adjusted most comfortable level, as described in Fuglsang et al. (2020). The average RDS score was 11.9 (standard deviation: 3.4), with individual participants reaching scores between 6 and 19 (maximum possible score: 28).

All participants were provided with Oticon Opn HAs for the duration of the study. The HAs were fitted based on the individuals' audiograms using the standard prescription offered by the fitting software (voice-aligned compression – VAC; Le Goff, 2015). Three different DIR+NR settings were defined: "NR_{Off}" (DIR+NR algorithm inactive, HA directivity pattern in omni-directional mode), "NR_{Mild}" (moderate standard parameterization of the DIR+NR algorithm); "NR_{Strong}" (customized strong DIR+NR setting). All participants 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

Spectro-temporal modulation detection sensitivity was measured using an STM test from our previous study (Zaar et al., 2023), which considered six different stimulus variants. Two stimulus variants that showed promising results in Zaar et al. (2023) were considered here: "Noisy-$LP_{2c/o}$" was based on a noise carrier consisting of 2499 components logarithmically spaced between 354 and 2000 Hz (also used by Bernstein et al., 2016), whereas "$Tonal_{2c/o}$" was based on a 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 the linear frequency axis and $1/f$-weighting to obtain a "pink" spectrum. Both stimulus variants had a similar spectral modulation frequency of 2 c/o and temporal modulation rate of 4 Hz (upward moving ripples). An auditory-inspired time-frequency representation of the stimulus variants, obtained using a gammatone filter bank and Hilbert envelope extraction, is provided in Fig. 2.

The STM test was administered as described in Zaar et al. (2023), using adaptive threshold tracking by means of a 3-alternative forced-choice (AFC) procedure with a 1-up/2-down tracking rule (using the AFC framework by Ewert, 2013; for details on the tracking procedure and experimental set up see Zaar et al., 2023). As in Zaar et al. (2023), the nominal broadband level was set at 65 dB SPL at the center-of-head position in a virtual diffuse field and linear amplification was applied to ensure a minimum of 15 dB sensation level in each 3rd-octave frequency band within the stimulus frequency range (cf. Humes, 2007). The amplification was applied in an ear-specific fashion in the present study, whereas it was based on the across-ear average in Zaar et al. (2023). The participants 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 participants 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.

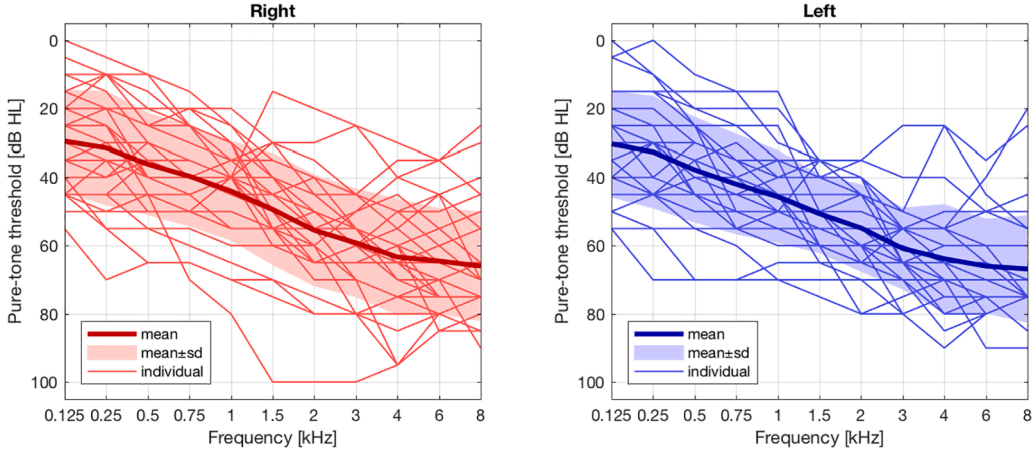

Fig. 1. Pure-tone thresholds obtained for the right and left ear of the 30 individual listeners (thin lines) along with mean and standard deviation across listeners (thick lines and shaded area, respectively).

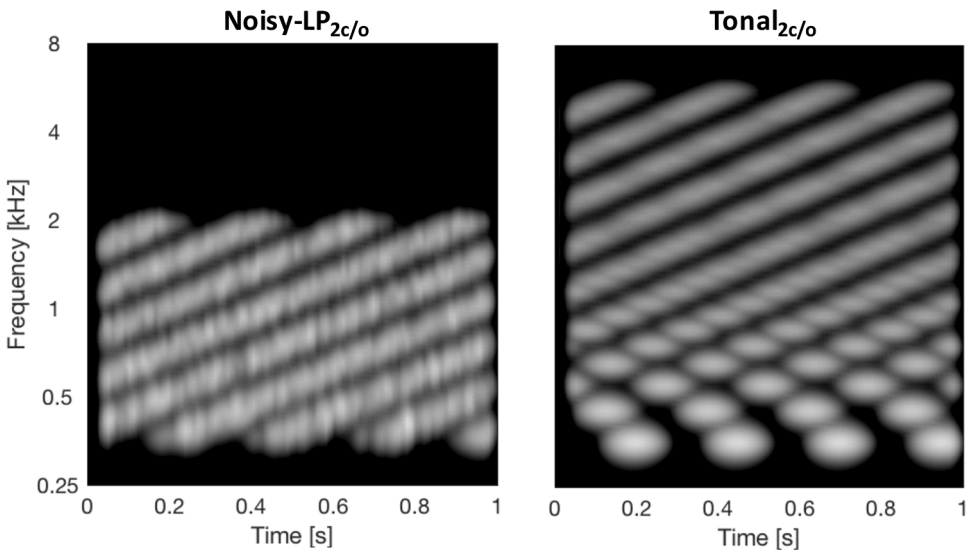

Fig. 2. Auditory spectrograms of the two considered STM stimulus variants with full modulation. Band-limited random-noise carrier (left) and 4-octave wide 100-Hz complex-tone carrier (right), both modulated with 2 c/o and 4 Hz (upward moving).

2.3. Speech-in-noise test

Speech intelligibility was measured using the Danish Hearing In Noise Test (HINT, Nielsen and Dau, 2011) using a spatial loudspeaker set up in a quiet but slightly reverberant room corresponding to a typical living-room environment (reverberation time T_{30} of about 0.5 s; size 7.52 × 4.74 × 2.76 m; designed according to IEC 1985). Three Dynaudio (Skanderborg, Denmark) BM6 loudspeakers were positioned along a circle with a diameter of 2.5 m at azimuth angles of 0° and ±100° (see Fig. 3). The participants wore their HAs and 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. They were instructed to use the headrest to maintain a static head position and 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 2nd 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 processed by an RME (Haimhausen, Germany) Fireface UCX soundcard at a sampling rate of 44.1 kHz and a resolution of 16 bits. The speech-test stimuli were amplified using customized Bang & Olufsen (Struer, Denmark) amplifiers and the talk-back microphone was routed to the soundcard's headphone output such that the audiologist could listen to the responses on Sennheiser (Wedemark, Germany) HDA200 headphones. The target speech, consisting of 5-word meaningful, grammatically correct Danish sentences uttered by a male speaker (Nielsen and Dau, 2011), was presented from the front (see "T" in Fig. 3) at 65 dB SPL(C) relative to the center-of-head position. Running speech maskers, spoken by two different male talkers and mixed with low-level speech-shaped noise (−6 dB relative to the running speech level), were played from the sides (see "M_1" and "M_2" in Fig. 3). SRTs at the 50%-sentences-correct level were tracked by adjusting the masker levels (and thus 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.

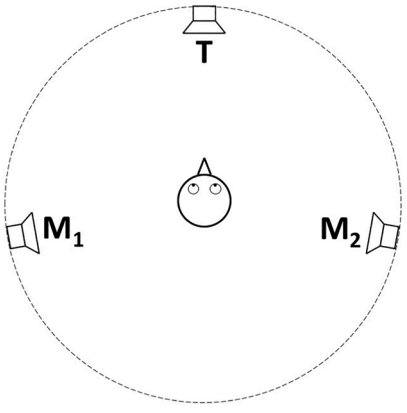

Fig. 3. Sketch of speech-test set up. T refers to the target speech, M_1 and M_2 to the masker signals.

A first training run was conducted with 20 sentences (one training list), starting at an SNR of −4 dB. A second training run was conducted with 40 sentences (two training lists), starting at an SNR 7 dB below the SRT measured in the first training run. In both training runs the HAs were in NR_{Off} setting. Next, the SRT measurement for the NR_{Off} setting was conducted, using 40 sentences and starting at an SNR 7 dB below the SRT measured in the second training run. Next, the SRT for either the NR_{Mild} or NR_{Strong} setting (balanced across participants) was obtained, using 40 sentences and starting at an SNR 12 dB below the SRT_{NRoff}. Lastly, the SRT measurement for the other setting (NR_{Strong} or NR_{Mild}) was conducted, using 40 sentences and starting again at an SNR 12 dB below the SRT_{NRoff}. The SRT measurements for settings NR_{Off}, NR_{Mild}, and NR_{Strong} were obtained using HINT lists 5/6, 7/8, and 9/10 (order balanced across participants). The precision of the SRT estimate, which is ordinarily calculated as the mean across SNRs presented for sentences 5–40, was refined based on the words-correct scores using the procedure described in Rønne et al. (2017) and also used in Zaar et al. (2023).

2.4. Experimental protocol and field testing

The study was structured in three separate sessions of approximately 2 h and two field periods (each 3–5 weeks) in between the three sessions.

In the first session, the participants 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 in either ear 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 (no older than 1 year) was available. Participants were fit with Oticon Opn HAs using the standard fitting rationale (VAC; Le Goff, 2015). The DIR+NR setting of the HAs was set to either NR_{Mild} or NR_{Strong} (balanced across participants) for the first field period. The participants and the audiologist were unaware of the choice of setting. The Danish SSQ12 questionnaire (Noble et al., 2013) was handed out and participants were instructed to expose themselves to the scenarios mentioned in the questionnaire and to fill in the questionnaire directly before the next session. The STM test was run, consisting of a short training followed by three runs with one stimulus variant, a short break, and three runs with the other stimulus variant. The order of the stimulus variants (Noisy-$LP_{2c/o}$ and $Tonal_{2c/o}$) was balanced across participants.

In the second session, the SSQ12 questionnaire for the first field period was collected. The HAs were programmed such that the three DIR+NR settings NR_{Off}, NR_{Mild}, and NR_{Strong} were selectable using a phone app controlled by the experimenter. The speech-in-noise test was run as described above. When the participants left for the second field trial, the DIR+NR setting of the HAs was set to NR_{Strong} for those participants who had had NR_{Mild} in the first field period and vice versa. The participants and the audiologist were again unaware of the choice of setting. The Danish comparative version of the SSQ12 questionnaire (SSQ12-C; Jensen et al., 2009) was handed out and participants were again instructed to expose themselves to the mentioned scenarios and to fill out the questionnaire directly before the next session. The STM retest was conducted using the exact same procedure as in the first session but with the order of the STM stimulus variants (Noisy-$LP_{2c/o}$ and $Tonal_{2c/o}$) flipped.

In the third session, the SSQ12-C questionnaire for the second field period was collected. Next, the participants were asked to rate their final preference comparing between the settings from the two field periods on a 5-point scale (−2, −1, 0, 1, 2), where the extremes indicated strong preference for either of the two settings and the midpoint indicated no preference. In addition, the participants were asked to indicate their level of certainty regarding their preference on a scale from 0 (very uncertain) to 10 (very certain). Another speech test was conducted at pre-selected fixed SNRs to measure listening effort via pupillometry (results not reported here). Lastly, the RDS test was conducted and the participants were de-briefed.

3. Results and analysis

3.1. Spectro-temporal modulation detection

Fig. 4 shows the STM thresholds in dB full scale (dB FS; relative to 1) obtained with the two considered stimulus variants (test and retest separately) in terms of group averages (black lines), standard deviations (gray shaded areas), and individual thresholds (gray dots). All participants were able to complete the task in each of the three runs underlying the plotted thresholds (cf. Section 2.2), for both stimulus variants. The Noisy-$LP_{2c/o}$ stimulus variant yielded higher thresholds (corresponding to greater difficulty) than the $Tonal_{2c/o}$ variant, whereas group results obtained with a given stimulus variant did not differ significantly between test and retest (although a slightly lower average threshold was found in the $Tonal_{2c/o}$ retest as compared to the test). The variability across listeners was similar for the two variants. A two-way ANOVA accordingly revealed a significant effect [$F(1, 116)=33.77$, $p<0.001$] of stimulus variant, whereby $Tonal_{2c/o}$ yielded 4.7 dB lower thresholds than Noisy-$LP_{2c/o}$ on average. No significant effect of test/retest [$F(1, 116)=0.28$, $p = 0.6$] and no interaction between stimulus variant and test/retest [$F(1, 116)=0.1$, $p = 0.75$] was found.

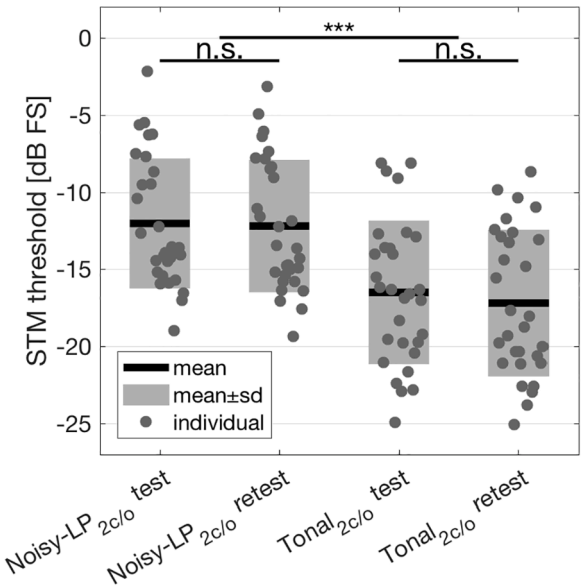

Fig. 4. STM thresholds measured in test and retest for the considered stimulus variants Noisy-$LP_{2c/o}$ (left) and $Tonal_{2c/o}$ (right). The black lines depict the group average, the shaded areas represent ±1 standard deviation around the mean, and the gray dots show the individual results (with a slight horizontal offset for better visibility). The black horizontal lines in the top part indicate the result of a two-way ANOVA (with n.s.: not significant and ***: $p<0.001$).

Fig. 5 provides a detailed overview of the test-retest reliability found for the two stimulus variants. The top panels show the STM thresholds measured in the second session (retest) as a function of the STM thresholds measured in the first session (test) for the Noisy-$LP_{2c/o}$ (left) and $Tonal_{2c/o}$ variants (right). It can be observed that the thresholds obtained with the Noisy-$LP_{2c/o}$ stimulus variant were more aligned between test and retest than those obtained with the $Tonal_{2c/o}$ stimulus variant. This is confirmed by the Intraclass Correlation Coefficients (ICC; Koo and Li, 2016) shown in the respective panels, which were computed between test and retest using a single rating, absolute agreement, one-way random effects model, and indicate "excellent" (ICC>0.9) reliability for Noisy-$LP_{2c/o}$ and "good" (0.75<ICC<0.9) reliability for $Tonal_{2c/o}$. The bottom panels of Fig. 5 show Bland-Altman plots that support this difference in test-retest reliability as the confidence intervals of test-retest differences are almost doubled for $Tonal_{2c/o}$ as compared to Noisy-$LP_{2c/o}$. Furthermore, a slight positive bias can be observed for $Tonal_{2c/o}$, indicating that participants performed slightly better in the retest for this variant (see also Fig. 4), albeit not significantly so according to the ANOVA reported above.

Based on the above observations and analyses, the STM thresholds obtained with the Noisy-$LP_{2c/o}$ stimulus variant were considered to be more reliable and thus selected for further analysis. The average across thresholds measured with Noisy-$LP_{2c/o}$ in test and retest was used to represent STM sensitivity in the remaining analyses reported in this article.

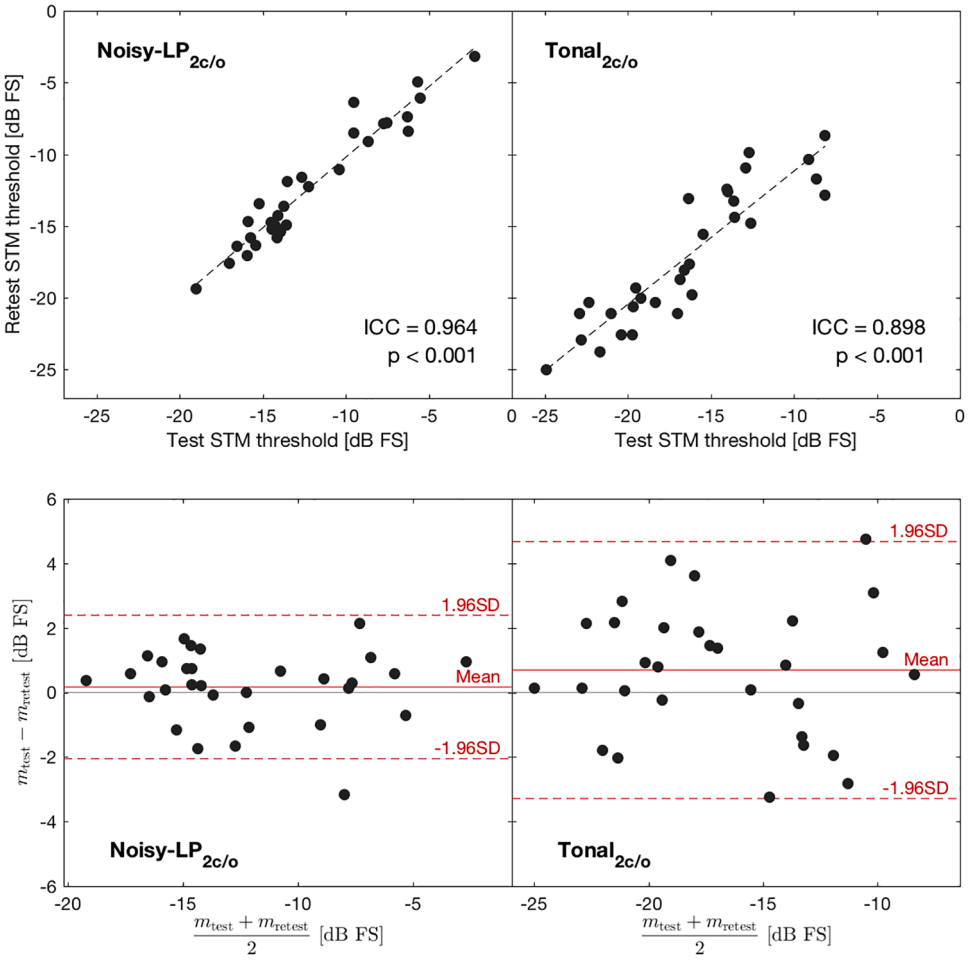

Fig. 5. Top: scatter plots of STM thresholds measured in test vs. retest for Noisy-$LP_{2c/o}$ (left) and $Tonal_{2c/o}$ (right). The dashed lines represent regression lines. The corresponding Intraclass Correlation Coefficient (ICC) is shown in the respective panels. Bottom: Bland-Altman plots of STM thresholds for Noisy-$LP_{2c/o}$ (left) and $Tonal_{2c/o}$ (right). The difference between test and retest is plotted as a function of the mean of test and retest. The red solid and dashed lines show the mean and the confidence intervals, respectively.

3.2. Speech-in-noise reception

Fig. 6 shows the SRTs measured in the speech-in-noise test with the three considered HA settings in terms of group averages (black lines), standard deviations (gray shaded areas), and individual SRTs (gray dots connected by dashed lines for each participant). It can be observed that, on average, the NR_{Off} setting yielded the highest SRT (3.6 dB), the NR_{Mild} setting a 1.8-dB benefit, and the NR_{Strong} setting a 4.4-dB benefit. A large variability across participants was found, especially for NR_{Off} and NR_{Mild}, with some participants exhibiting very poor performance (SRTs above 10 dB). A trend of decreasing SRTs with increasing DIR+NR strength (from left to right in Fig. 6) was present for all participants (see dashed lines), apart from some minor deviations. A two-way ANOVA with HA setting as a fixed effect and participant as a random effect revealed a significant [$F(2, 58)=54.48$, $p<0.001$] effect of HA setting. A Tukey-Kramer post-hoc test indicated significant differences [$p<0.001$] between the SRTs measured for all HA settings.

3.3. Predicting aided speech-in-noise reception performance

The predictive power of the audiogram and the STM thresholds with regard to aided speech-in-noise reception was assessed by means of correlation and linear regression analyses. For that purpose, the 4-frequency better-ear pure-tone average (henceforth referred to as PTA) was obtained by averaging the ear-specific pure-tone thresholds measured for 0.5, 1, 2, and 4 kHz and then selecting the lower value

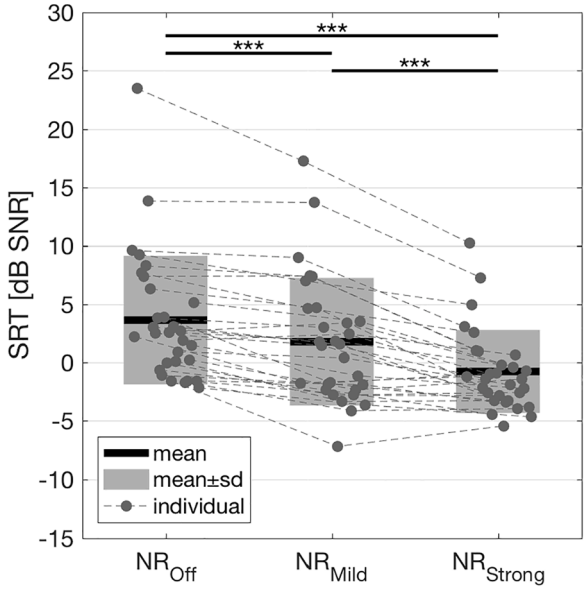

Fig. 6. SRTs measured with the three different HA settings NR_{off}, NR_{Mild}, and NR_{Strong}. The black lines depict the group average, the shaded areas represent ±1 standard deviation around the mean, and the gray dots show the individual results (with a slight horizontal offset for better visibility), which are connected by a dashed line for each participant. The black horizontal lines in the top part indicate the result of a two-way ANOVA (with ***: $p<0.001$).

Table 1
Correlation matrix comparing SRT measured in the NR_{Off} condition (SRT_{NRoff}), 4-frequency better-ear pure-tone average (PTA), STM thresholds, and RDS scores. Correlation coefficients are shown followed by the corresponding p-values in parentheses. The asterisks indicate significant correlation according to Bonferroni-corrected significance levels (*: <0.05; **: <0.01; ***: <0.001).

	PTA	STM	RDS
SRT_{NRoff}	0.713*** ($<10^{-5}$)	0.780*** ($<10^{-6}$)	−0.511* (0.004)
PTA		0.627** ($<10^{-3}$)	−0.233 (0.22)
STM			−0.550* (0.002)

across the two ears. Table 1 shows an initial assessment of the correlation structure between SRTs measured with the NR_{Off} HA setting (SRT_{NRoff}), the PTA, the STM thresholds, and the RDS scores (assumed to represent working memory capacity). Significant positive correlations were found between all pairwise combinations of SRT_{NRoff}, PTA, and STM, where high values indicate poor performance. RDS, where high values indicate good performance, was instead moderately negatively correlated with SRT_{NRoff} and STM, but not with PTA.

For the following considerations on SRT_{NRoff} variance explained it should be noted that the upper bound of predictable SRT_{NRoff} variance was found to be 96% (limited by test-retest variability; calculated by comparing the results of the NR_{Off} measurement run with those of the second training run, which was also conducted with the NR_{Off} HA setting). Fig. 7 shows the SRTs measured with the NR_{Off} HA setting as a function of, respectively, PTA (left), STM (middle), and a linear regression model with PTA and STM as predictors (right). A highly significant positive correlation (see Table 1) was found for the PTA, which accounted for 51% of SRT_{NRoff} variance. An even stronger positive correlation (see Table 1) was found for the STM thresholds, which accounted for 61% of SRT_{NRoff} variance. The linear regression model with PTA and STM as predictors accounted for 69% of the SRT_{NRoff} variance [$p<10^{-6}$], with both PTA [$p = 0.012$] and STM [$p<10^{-3}$] contributing significantly. When adding RDS as a third predictor, PTA and STM remained significant predictors whereas RDS did not contribute significantly [$p = 0.17$]. The PTA added 8% of SRT_{NRoff} variance explained as compared to using STM as the sole predictor. The STM thresholds added 18% of SRT_{NRoff} variance explained as compared to using the PTA as the sole predictor. When factoring out the PTA-based prediction from the SRT_{NRoff}, the residual was still significantly correlated with the STM thresholds [$r = 0.48$, $p = 0.008$], whereas the PTA could not reliably account for the residual when factoring out the STM-based prediction from the SRT_{NRoff} [$r = 0.36$, $p = 0.052$].

As one of the test participants exhibited an extremely high SRT_{NRoff} of 23.5 dB (more than 2.7 times the interquartile range above the third quartile) and since the analysis results may be biased by such extreme outliers, the same analysis as described above was repeated without this participant. The results revealed that removal of this participant slightly reduced the predictive power of the PTA [49% of variance explained; $r = 0.7$, $p<10^{-4}$], whilst substantially increasing the predictive power of STM [68% of variance explained; $r = 0.82$, $p<10^{-7}$] and of the two-predictor regression model [75% of variance explained, $p_{model}<10^{-7}$, $p_{PTA}=0.012$, $p_{STM}<10^{-4}$]. In the regression model, PTA added 7% of variance explained as compared to using STM as the sole predictor and the STM thresholds added 26% of variance explained as compared to using the PTA as the sole predictor.

3.4. Predicting speech-reception performance benefit due to DIR+NR

A second analysis was conducted to determine whether the SRT-benefit induced by strong compared to mild DIR+NR (NR_{Strong} and NR_{Mild}, respectively) was related to PTA and STM. For that purpose, an SRT difference measure was defined as $\Delta SRT = SRT_{NRmild} - SRT_{NRstrong}$, such that the expected benefit (i.e., lowering) in SRT due to NR_{Strong} resulted in positive values and vice versa. The amount of SRT benefit (ΔSRT) was positively correlated with the SRT_{NRoff} [$r = 0.71$, $p<10^{-5}$], indicating that poor speech-in-noise performers benefitted more from NR_{Strong} than good speech-in-noise performers.

Fig. 8 depicts the ΔSRT measure as a function of, respectively, PTA (left), STM (middle), and a linear regression model with PTA and STM as predictors (right), in analogy with the analysis shown in Fig. 7. As expected, ΔSRT was mostly positive, indicating that most participants benefitted from NR_{Strong} relative to NR_{Mild} in terms of SRT (by up to 7 dB; see also Fig. 6) whereas some participants did not benefit from NR_{Strong}, with ΔSRT values around zero and slightly below. A highly significant correlation [$r = 0.73$, $p<10^{-5}$] between ΔSRT and PTA was found, which accounted for 54% of the ΔSRT variance. A slightly smaller positive correlation [$r = 0.71$, $p<10^{-5}$] was found for the STM thresholds, which accounted for 51% of the ΔSRT variance. The linear regression model with PTA and STM as predictors accounted for 64% of the ΔSRT variance [$p<10^{-6}$], with both PTA [$p = 0.004$] and STM [$p = 0.009$] contributing significantly. The PTA added 13% of ΔSRT variance explained as compared to using STM as the sole predictor. The STM thresholds added 10% of ΔSRT variance explained as compared to using the PTA as the sole predictor.

3.5. Subjective evaluation of DIR+NR settings

The subjective preference for the DIR+NR settings NR_{Strong} and NR_{Mild} was assessed by means of the SSQ12-C questionnaire and by

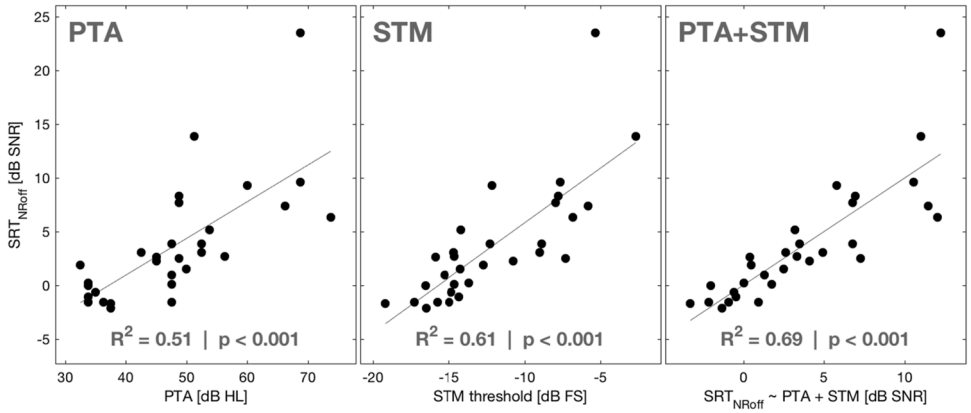

Fig. 7. SRTs measured with NR_{off} as a function of (from left to right): the 4-frequency better-ear pure-tone average (PTA), the STM thresholds, and a two-predictor regression model using PTA and STM as input variables. The dots represent data from individual subjects; the lines are regression lines. The R^2 and corresponding p-values (down to 3 decimals) are indicated at the bottom of the respective panels.

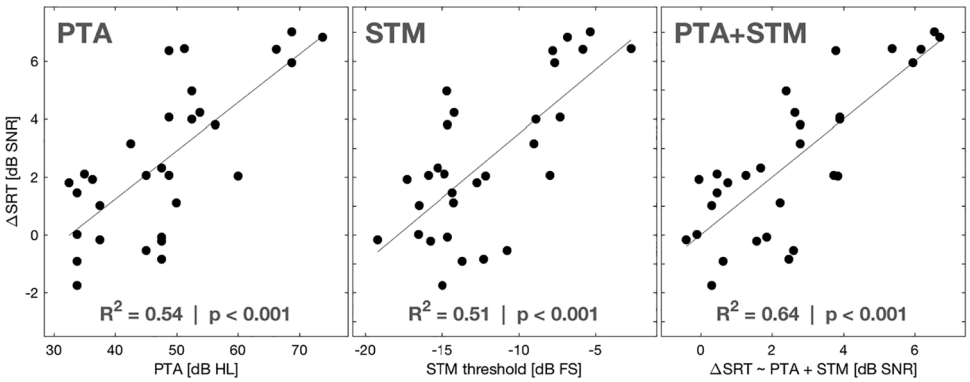

Fig. 8. SRT benefit ΔSRT induced by NR_{Strong} as compared to NR_{Mild} as a function of (from left to right): the 4-frequency better-ear pure-tone average (PTA), the STM thresholds, and a two-predictor regression model using PTA and STM as input variables. The dots represent data from individual subjects; the lines are regression lines. The R^2 and corresponding p-values (down to 3 decimals) are indicated at the bottom of the respective panels.

means of a preference rating. The distribution of preference ratings, shown in Fig. 9, indicates an almost even split of preference for NR_{Mild} (11 participants), no preference (9 participants), and preference for NR_{Strong} (10 participants).

For correlation analyses, the raw preference ratings were weighted by multiplication with the corresponding certainty ratings. The SSQ12-C questionnaire data were considered as an average across all 12 questions ($SSQ12\text{-}C_{ALL}$) and as subscale-specific averages ($SSQ12\text{-}C_{Speech}$, $SSQ12\text{-}C_{Spatial}$, $SSQ12\text{-}C_{Quality}$). Table 2 shows the correlations between the preference ratings and the SSQ12-C scores. The preference ratings were very highly correlated with the SSQ12-C scores, except for the spatial subscale ($SSQ12\text{-}C_{Spatial}$), indicating that the preference ratings were mainly driven by speech understanding and sound quality.

Due to the virtual interchangeability of the SSQ12-C scores and the preference ratings, only the preference ratings were considered for further analyses. Table 3 shows the results of a correlation analysis conducted to establish whether the preference ratings were related to (i) performance in the speech task with no DIR+NR (i.e., SRT_{NRoff}), (ii) SRT-benefit induced by NR_{Strong} relative to NR_{Mild} (i.e., ΔSRT), pure-tone thresholds (PTA), and (iv) STM sensitivity. For the full data set of 30 participants, these comparisons showed correlational trends in the expected positive direction but no significant relationship between the preference ratings and any of the other variables was found (see left column of Table 3). For a subset of 23 participants, obtained by discarding five participants for whom the HA logging data indicated little exposure to non-quiet acoustical scenarios[1] and two participants with technical problems related to the HA fitting, some stronger positive correlations emerged, especially between STM sensitivity and the preference ratings (see right column of Table 3). However, given the number of comparisons, these correlations fell slightly short of reaching statistical significance, considering a Bonferroni correction by division of the most lenient commonly used significance level of 0.05 by the number of variables (0.05/4 = 0.0125).

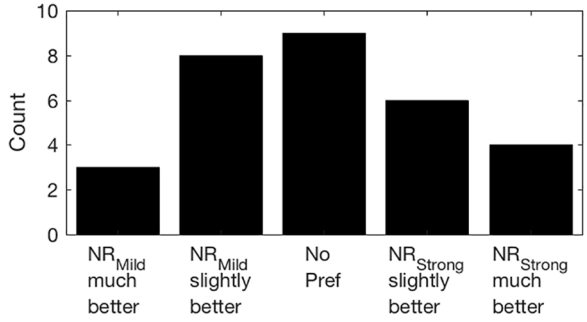

Fig. 9. Histogram of preference ratings for DIR+NR settings provided after the second field period.

Table 2
Correlations between preference ratings (PREF) and SSQ12-C questionnaire results averaged across all questions ($SSQ12\text{-}C_{ALL}$) and across the questions related to the Speech, Spatial, and Quality subscales ($SSQ12\text{-}C_{Speech}$, $SSQ12\text{-}C_{Spatial}$, $SSQ12\text{-}C_{Quality}$). Correlation coefficients are shown followed by the corresponding p-values in parentheses. The asterisks indicate significant correlation according to Bonferroni-corrected significance levels (*: <0.05; **: <0.01; ***: <0.001).

	$SSQ12\text{-}C_{ALL}$	$SSQ12\text{-}C_{Speech}$	$SSQ12\text{-}C_{Spatial}$	$SSQ12\text{-}C_{Quality}$
PREF	0.890*** ($<10^{-10}$)	0.874*** ($<10^{-9}$)	0.483* (0.008)	0.845*** ($<10^{-7}$)

Table 3
Correlations between preference ratings (PREF) and SRT_{NRoff}, ΔSRT, PTA, STM for the full data set of 30 participants (left column) and for a subset of 23 participants (right column). Correlation coefficients are shown followed by the corresponding p-values in parentheses. All p-values are above the 0.05 level when correcting for multiple comparison (i.e., above 0.0125).

	PREF (full data set, $n = 30$)	PREF (subset, $n = 23$)
SRT_{NRoff}	0.204 (0.28)	0.343 (0.11)
ΔSRT	0.219 (0.24)	0.388 (0.07)
PTA	0.267 (0.15)	0.386 (0.07)
STM	0.240 (0.20)	0.483 (0.02)

4. Discussion

4.1. Summary of main findings

The main findings of the study can be grouped in four categories. First, the two most promising STM stimulus variants identified in our previous study (Zaar et al., 2023) were compared with respect to their test-retest reliability. The bandlimited noise-based variant with 2 c/o and 4 Hz (introduced by Bernstein et al., 2016) was found to yield substantially better test-retest reliability than the complex-tone based variant with the same modulation parameters and was thus chosen to represent STM sensitivity for the remaining analyses (Fig. 5). Second, STM thresholds were significantly correlated with realistic SRTs measured with HAs (NR_{Off} condition), accounting for 61% of the SRT variance, while the PTA accounted for 51% and a combined linear

[1] Only participants who had spent more than 30% of the HA use time in sound environments that were not classified as "quiet" and had sound levels above 50 dB SPL according to the HA logging data were retained.

regression model with STM and PTA as inputs accounted for 69% of the SRT variance (Fig. 7). Third, STM thresholds and PTA were significantly correlated with the SRT benefit (ΔSRT) provided by NR_{Strong} relative to NR_{Mild}, accounting for, respectively, 51% and 54% of the ΔSRT variance and for 64% jointly in a linear regression model (Fig. 8). Fourth, the subjective preference for DIR+NR settings was not significantly associated with any of the laboratory measures (SRT, ΔSRT, STM, PTA), indicating that participants did not necessarily prefer what their test performance would suggest (Table 3); however, there were observable trends in the hypothesized direction, indicating that poor performers may prefer a stronger DIR+NR setting than good performers.

The overarching hypotheses were postulated as follows: "Subjects with poor aided speech reception (a) show poor STM sensitivity, (b) benefit most from DIR+NR and therefore (c) prefer strong DIR+NR (and vice versa)". While (a) and (b) were confirmed, (c) could neither be confirmed nor refuted.

4.2. STM test and stimulus variants

The choice of the two STM stimulus variants (Noisy-LP_{2co} and $Tonal_{2co}$) tested was motivated by the results of our previous study (Zaar et al., 2023), where the Noisy-LP_{2co} stimulus (referring to the bandlimited noise-carrier based stimulus with 4 Hz and 2 c/o introduced by Bernstein et al., 2016) and the complex-tone based $Tonal_{1co}$ and $Tonal_{2co}$ stimuli (with 4 Hz and 1 c/o and 2 c/o, respectively) showed significant correlations with SRTs. The $Tonal_{2co}$ variant was selected over the $Tonal_{1co}$ variant as 2 c/o has been established as a meaningful choice for predicting speech intelligibility in various STM studies (Bernstein et al., 2013; Mehraei et al., 2014, 2016). Our binaurally administered STM test procedure was identical to that established in Zaar et al. (2023), with one difference: while identical audibility compensation (according to the sufficiently audible approach suggested by Humes, 2007) based on the across-ear average audiogram was used in Zaar et al. (2023), the present study instead applied ear-specific amplification based on the corresponding audiogram. This was done in analogy to the amplification provided in the respective speech tests, as linear amplification based on the across-ear average audiogram was applied to the speech-test stimuli in Zaar et al. (2023), whereas the current study used ear-specific (non-linear) HA amplification. However, the difference between these two approaches in the tested population was small given the largely symmetric hearing losses considered.

The population of 30 participants tested in the present study showed similar performance trends for the two stimulus variants as observed in Zaar et al. (2023), with Noisy-LP_{2co} yielding higher thresholds than $Tonal_{2co}$. However, the average thresholds in the present study were higher, indicating a more poorly performing population as compared to Zaar et al. (2023). Nonetheless, all participants were able to obtain a threshold in each measurement, confirming the finding from Zaar et al. (2023) that the proposed STM test design is suitable for the target population of elderly listeners with moderate-to-severe hearing loss (as opposed to the STM test paradigm used by Bernstein et al., 2016).

In contrast with the Zaar et al. (2023) study, the present study assessed the test-retest reliability of the STM test (measured across days/sessions). The analysis revealed a clearly superior test-retest reliability of the Noisy-LP_{2co} stimulus variant, as compared to the $Tonal_{2co}$ variant, which showed both a larger spread (lower ICC) and a slight bias due to lower thresholds in the retest (see Figs. 4 and 5). The complex-tone carrier was assumed to provide the advantage of not having low-frequency intrinsic fluctuations (around 4 Hz) that could interfere with the imposed modulation pattern (cf. Dau et al., 1999), which is consistent with the lower thresholds obtained for the $Tonal_{2co}$ variant. Another motivating factor for considering this variant was the fact that speech signals are dominated by periodic carrier/source signals and that such periodic carrier signals may be crucial for the auditory processing and perception of speech (Steinmetzger and Rosen, 2015; Carney et al., 2015; Carney, 2018; Scheidiger et al., 2018; Steinmetzger et al., 2019; Zaar and Carney, 2022). However, the test-retest variation observed with the $Tonal_{2co}$ variant appears problematic and suggests an influence of aspects such as listening strategy and fatigue. Therefore, the Noisy-LP_{2co} variant was selected as the best representative of STM sensitivity, in line with Bernstein et al. (2016).

4.3. STM and speech reception with hearing aids

The present study found a very strong relationship between speech reception with HA amplification and STM sensitivity (61% of shared variance, see Fig. 7), which substantially exceeds the relationship (28% of shared variance) found in the related study by Bernstein et al. (2016) that also used (albeit simulated) HA amplification. Due to multiple methodological differences with regard to the STM- and speech-test designs between the two studies, the contributions of the various design differences to these results cannot be disentangled here. It may be speculated that the present study's STM test design, which allows for convergence of the adaptive measurement procedure even for the poorest performers, and the realistic speech-test design with open-set sentences and spatially distributed speech interferers have contributed to the strong relationship found here. It was demonstrated in Zaar et al. (2023) that a realistic speech-test set up akin to the one used in the present study revealed significantly more variation across subjects than a more artificial set up with a co-located speech-shaped noise masker, and that it also tended to lead to greater STM-SRT correlations. However, it should be mentioned that the present study used a diverse population with large speech-reception performance differences and that the PTA also accounted for a large amount of SRT variance (51%), which was substantially larger than the 31% found by Bernstein et al. (2016) for their high-frequency PTA. Consistent with Bernstein et al. (2016), STM and PTA were complementary in predicting SRTs (69% of variance explained), which furthermore confirmed trends found in Zaar et al. (2023) that did not reach statistical significance due to the small number of participants in that study. Finally, the strong correlations reported here may be considered a conservative account of the data, as a single participant with a very poor SRT of 23.5 dB SNR negatively affected the predictive power of STM and positively affected that of the PTA. Without this participant, STM accounted for 68%, PTA for 49%, and STM and PTA combined for 75% of the SRT variance (with an upper limit of 96% explainable SRT variance based on test-retest variation).

The DIR+NR HA processing conditions used in the speech test showed the expected effects of SRT improvements as a function of the strength of the processing, i.e., more improvement for NR_{Strong} than for NR_{Mild} (see Fig. 6; cf. Le Goff et al., 2016; Andersen et al., 2021). The SRT benefit (ΔSRT) induced by NR_{Strong} relative to NR_{Mild} differed substantially across participants, with some benefitting strongly and others not at all. The amount of ΔSRT was associated with the SRT measured in the NR_{Off} condition and thus also strongly correlated with STM and PTA, which accounted for 51% and 54% of the ΔSRT variance, respectively, and again yielded complementary contributions in a linear regression model that explained 64% of the ΔSRT variance (Fig. 8). The amount of SRT benefit is likely related to multiple factors, including the amount of amplification (related to the audiogram), the acoustic coupling (cf. Cubick et al., 2022), as well as the interaction of the adaptive DIR+NR processing with the speech-test stimuli, which are influenced by the participants' performance levels, especially in terms of SNR. The observed connections between SRT benefit and SRT_{NRoff}, PTA, and STM are well in line with these considerations.

4.4. Preference for hearing-aid settings

The subjective evaluation of the two tested DIR+NR settings NR_{Strong} and NR_{Mild} was conducted by means of questionnaires and preference ratings in the field. The comparison between SSQ12-C scores and preference ratings indicated a very strong alignment between the two, with the preference ratings being mainly driven by aspects represented in the

speech- and quality-related questions of the SSQ12-C questionnaire (see Table 2). Despite the fact that NR_{Strong} was a customized, quite aggressive parameterization of the algorithm behind the Oticon "Open Sound Navigator" (Le Goff et al., 2016), the DIR+NR preferences of the 30 participants were evenly distributed (see Fig. 9), i.e., about a third of the participants appreciated the strong DIR+NR setting. However, the hypothesized connection between poor speech-test performance, strong SRT benefit, high PTA, and poor STM performance (which are partially interchangeable given the above-discussed connections between them) did not emerge in a statistically significant sense. In other words, participants did not necessarily prioritize support with speech intelligibility in difficult scenarios in their preference ratings (assuming that the speech-test results are indicative of the experienced difficulty with speech reception in the field). Correlational trends in the expected direction can be observed (Table 3), especially when excluding 7 participants who did not expose themselves very much to noisy conditions or had technical issues with the HA fitting. However, there were clearly other sources of variation contributing to the individual preferences that were not represented by any other measure considered in the present study. Furthermore, the test-retest reliability of the subjective evaluation is unknown for the considered method and it is thus unclear how reliable the employed measure of preference is. The reasons that participants provided when discussing their choices were in many cases completely unrelated to the expected DIR+NR effects, casting doubt on the precision of the measurement method.

4.5. Limitations of the study

While this study has much to offer in terms of strongly significant and expected effects, there are also a number of limitations. First and foremost, the assessment of DIR+NR preference in the field yielded inconclusive data. While subjective real-life preference is generally difficult to measure reliably and affected by many non-auditory aspects (Jensen et al., 2013; Naylor et al., 2015), methodological improvements are likely possible. For instance, the participants only used one HA setting in a given field period and had to decide after both field periods were completed, comparing not only between settings but also between different field periods. This could be overcome by offering the HA settings simultaneously, such that A/B comparisons can be conducted and preferences can be reported in an on-going fashion. Additionally, advanced tracking of acoustical parameters and program choice over time may help reveal use patterns as well as situational preferences, allowing to focus on, e.g., noisy situations where DIR+NR algorithms are expected to be useful/appreciated (Christensen et al., 2021; Pasta et al., 2022).

Secondly, we observed a substantial correlation between the RDS scores and the SRTs, as well as between the RDS scores and STM sensitivity (Table 1). STM thresholds and PTA provided much stronger correlations with SRTs than RDS and RDS did not explain substantial SRT variance beyond STM. Nonetheless, it appears reasonable to suspect aspects of working memory (supposedly represented by RDS scores) to affect speech reception as well as STM detection when the latter is measured in a 3-AFC procedure. However, the fact that the RDS-test stimuli were administered via the auditory modality (at most comfortable level) casts doubt on whether a poor RDS score comes about as a result of a generally limited working-memory capacity or due to a temporary working-memory limitation induced by the cognitive load resulting from the need to simultaneously decode the acoustic stimuli through an impaired auditory system (cf. Füllgrabe and Öztürk, 2022). The latter interpretation appears equally plausible, as the RDS scores were correlated with the ability to understand speech (SRT), which is connected to STM sensitivity. It remains an open question to what extent working memory and, more generally, cognitive ability affect measures of auditory function such as audiogram and PTA, whereas there is evidence for cognitive ability being associated with speech-reception ability (Lunner and Sundewall-Thorén, 2007; Petersen et al., 2016).

Finally, while the proposed STM test appears to yield powerful predictions of aided speech-reception performance, it is not fit for clinical use in its current form: Firstly, this is due to the duration of the test, which took about 20 min (using a short training and three measurement runs). Secondly, the use of a computer display and interface (computer mouse, keyboard, or similar) to conduct the 3-AFC decision procedure requires substantially more equipment than available in a simple audiometric headphones-and-pushbutton set up in a typical clinic. Further investigations on shortening the test duration and simplifying the test set up, whilst maintaining a high test-retest reliability and predictive power with respect to aided speech reception, are well motivated by the present study.

5. Conclusion

The present study evaluated a previously suggested STM detection test paradigm (Zaar et al., 2023) with respect to (i) test-retest reliability of the two most promising STM stimulus variants (using a noise carrier and a complex-tone carrier, respectively), (ii) its predictive power relative to realistic speech-in-noise reception with non-linear hearing-aid amplification, (iii) its connection to effects of DIR+NR hearing-aid processing, and (iv) its relation to DIR+NR preference in the field. It was hypothesized that subjects with poor aided speech reception would (a) show poor STM sensitivity, (b) benefit most from DIR+NR, and therefore (c) prefer strong DIR+NR (and vice versa).

The results, obtained in a combined laboratory and field study with thirty elderly HI participants allow the following conclusions:

- The noise-carrier based STM stimulus variant (cf. Bernstein et al., 2016) showed far superior test-retest reliability as compared to the complex-tone carrier based STM stimulus.
- Speech-reception thresholds (SRTs) measured in a realistic setting with spatially distributed speech and noise maskers using hearing aids with non-linear amplification were very well predicted by STM sensitivity (R^2=0.61), to a lesser extent by the 4-frequency better-ear pure-tone average (PTA, R^2=0.51), with STM and PTA providing complementary information (R^2=0.69 in a two-predictor regression model). Hypothesis (a) was hereby confirmed.
- SRT benefit induced by strong DIR+NR relative to mild DIR+NR was correlated with the SRTs measured with amplification only (R^2=0.5), as well as with STM (R^2=0.51) and PTA (R^2=0.54), with STM and PTA providing complementary information (R^2=0.64 in a two-predictor regression model). Hypothesis (b) was hereby confirmed.
- Despite trends in the hypothesized direction, subjective preference for DIR+NR settings in the field were not significantly correlated with any of the laboratory measures (SRT, SRT benefit, STM, PTA), suggesting that participants did not necessarily prioritize speech perception in their preference ratings. Hypothesis (c) was thus neither confirmed nor refuted.
- Overall, the suggested STM test represents a valuable tool for diagnosing speech-reception problems with hearing aids and the resulting need for and benefit from DIR+NR hearing-aid processing. Future work should focus on optimizing the test for clinical practice and further investigating the relationship between subjective preference and objective benefit.

CRediT authorship contribution statement

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

Data availability

Most data can be made available on request.

Acknowledgments

We would like to acknowledge the crucial contributions from Thomas Behrens, James Harte, and Torsten Dau in establishing and supporting this project. We thank Thomas Lunner, Raul Sanchez Lopez, Bue Kristensen, Valentina Campagnaro, Fares El-Azm, Jacob Aderholt, Erengül Sabedin, and Sébastien Santurette for their contributions in support of this study.

Funding

This study was funded by the Oticon Foundation [grant number 17–0639].

References

Andersen, A.H., Santurette, S., Pedersen, M.S., Alickovic, E., Fiedler, L., Jensen, J., Behrens, T., 2021. Creating clarity in noisy environments by using deep learning in hearing aids. Semin. Hear. 42 (03), 260–281. https://doi.org/10.1055/s-0041-1735134.

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. https://doi.org/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. https://doi.org/10.1177/2331216516670387.

Blackburn, H.L., Benton, A.L., 1957. Revised administration and scoring of the Digit Span Test. J. Consult. Psychol. 21 (2), 139–143. https://doi.org/10.1037/h0047235.

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1121/1.428100.

Christensen, J.H., Saunders, G.H., Havtorn, L., Pontoppidan, N.H., 2021. Real-world hearing aid usage patterns and smartphone connectivity. Front. Digit. Health 3. https://doi.org/10.3389/fdgth.2021.722186.

Cubick, J., Caporali, S., Lelic, D., Catic, J., Damsgaard, A.V., Steen, R., Terri, I., Schmidt, E., 2022. The acoustics of instant ear tips and their implications for hearing aid fitting. Ear Hear. 43 (6), 1771–1782. https://doi.org/10.1097/AUD.0000000000001239.

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. https://doi.org/10.1121/1.428103.

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.

Füllgrabe, C., Öztürk, O.C., 2022. Immediate effects of (simulated) age-related hearing loss on cognitive processing and performance for the backward-digit-span task. Front. Aging Neurosci. 14. https://doi.org/10.3389/fnagi.2022.912746.

Fuglsang, S.A., Märcher-Rørsted, J., Dau, T., Hjortkjær, J., 2020. Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention. J. Neurosci. 40, 2562–2572. https://doi.org/10.1523/JNEUROSCI.1936-19.2020.

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. https://doi.org/10.3766/jaaa.18.7.6.

IEC, 1985. Sound system equipment part 13: listening tests on loudspeaker. In: International Electrotechnical Commission. Geneva, Switzerland, 13, p. 268.

Jensen, N.S., Akeroyd, M.A., Noble, W., Naylor, G., 2009. The Speech Spatial and Qualities of Hearing scale (SSQ) as a benefit measure. In: Poster presented at 4th NCRAR international conference. Portland, US.

Jensen, N.S., Neher, T., Laugesen, S., Johannesson, R., B., Kragelund, L., 2013. Laboratory and field study of the potential benefits of pinna cue-preserving hearing aids. Trends Amplif. 17 (3), 171–188. https://doi.org/10.1177/10847138135109.

Jerger, J.C., 2018. The evolution of the audiometric pure-tone technique. Hear. Rev. 25 (9), 12–18. https://hearingreview.com/hearing-products/testing-equipment/testing-diagnostics-equipment/evolution-audiometric-pure-tone-technique.

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, 2014. https://doi.org/10.3389/fnins.2014.00214.

Keidser, G., Dillon, H., Flax, M., Ching, T., Brewer, S., 2011. The NAL-NL2 prescription procedure. Audiol. Res. 10 (e24), 88–90. https://doi.org/10.4081/audiores.2011.e24.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15 (2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Le Goff, N. (2015) Amplifying soft sounds – a personal matter. Smørum: Oticon Whitepaper. https://www.oticon.global/-/media/oticon/main/pdf/master/whitepaper/17577uk_wp_amplifying-soft-sounds—a-personal-matter.pdf?la=en&hash=FA6C6B0CC3D48ECA7225FFEAF5DE3ABB2A33C3C6.

Le Goff, N., Jensen, J., Pedersen, M.S., Callaway, S.L. (2016) An introduction to opensound navigator. Smørum: Oticon Whitepaper. https://www.oticon.com/-/media/oticon-us/main/download-center/white-papers/15555-9950—opnsound-navigator.pdf.

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) https://doi.org/10.3389/fnins.2014.00348. Article.

Lunner, T., Sundewall-Thorén, E., 2007. Interactions between cognition, compression, and listening conditions: effects on speech-in-noise performance in a two-channel hearing aid. J. Am. Acad. Audiol. 18 (07), 604–617. https://doi.org/10.3766/jaaa.18.7.7.

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. https://doi.org/10.1121/1.4881918.

Naylor, G., Öberg, M., Wänström, G., Lunner, T., 2015. Exploring the effects oft he narrative embodied in the hearing aid fitting process on treatment outcomes. Ear Hear. 36 (5), 517–526. https://doi.org/10.1097/AUD.0000000000000157.

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

Noble, W., Jensen, N.S., Naylor, G., Bhullar, N., Akeroyd, M.A., 2013. A short form of the Speech, Spatial and Qualities of Hearing scale suitable for clinical use: the SSQ12. Int. J. Audiol. 52, 409–412. https://doi.org/10.3109/14992027.2013.781278.

Pasta, A., Petersen, M.K., Jensen, K.J., Pontoppidan, N.H., Larsen, J.E., Christensen, J.H., 2022. Measuring and modeling context-dependent preferences for hearing aid settings. User Model. User-Adap. Inter. 32, 977–998. https://doi.org/10.1007/s11257-022-09324-z.

Petersen, E.B., Lunner, T., Vestergaard, M., Sundewall-Thorén, E., 2016. Danish reading span data from 283 hearing-aid users, including a sub-group analysis of their relationship to speech-in-noise performance. Int. J. Audiol 55 (4), 254–261. https://doi.org/10.3109/14992027.2015.1125533.

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. https://doi.org/10.1044/jshr.2902.146.

Regev, J., Zaar, J., Relaño-Iborra, H., Dau, T., 2023a. Age-related reduction in amplitude modulation frequency selectivity. J. Acoust. Soc. Am. 153 (4) https://doi.org/10.1121/10.0017835.

Regev, J., Relaño-Iborra, H., Zaar, J., Dau, T., 2023b. Disentangling the effects of hearing loss and age on amplitude modulation frequency selectivity. *bioRxiv* preprint. https://doi.org/10.1101/2023.07.15.549131.

Regev, J., Zaar, J., Relaño-Iborra, H., Dau, T., 2023c. Impact of age versus hearing loss on amplitude modulation frequency selectivity and speech intelligibility. *bioRxiv* preprint. https://doi.org/10.1101/2023.07.20.549841.

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. https://doi.org/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. https://doi.org/10.1177/2331216520973539.

Sanchez-Lopez, R., Nielsen, S.G., El-Haj-Ali, M., Bianchi, F., Fereczkowski, M., Cañete, O., Wu, M., Neher, T., Dau, T., Santurette, S., 2021. Auditory tests for characterizing hearing deficits in listeners with various hearing abilities: the BEAR test battery. Front. Neursci. 15, 724007 https://doi.org/10.3389/fnins.2021.724007.

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1121/1.3097469.

Scheidiger, C., Carney, L.H., Dau, T., Zaar, J., 2018. Predicting speech intelligibility based on across-frequency contrast in simulated auditory-nerve fluctuations. Acta Acust. United Acust. 104 (5), 914–917. https://doi.org/10.3813/aaa.919245.

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).

Vermiglio, A.J., Soli, S.D., Freed, D.J., Fang, X., 2020. The effect of stimulus audibility on the relationship between pure-tone average and speech recognition in noise ability. J. Am. Acad. Audiol. 31 (3), 224–232. https://doi.org/10.3766/jaaa.19031.

Vermiglio, A.J., Fang, X., 2022. The World Health Organization (WHO) hearing impairment guidelines and a speech recognition in noise (SRN) disorder. Int. J. Audiol. 61 (10), 818–825. https://doi.org/10.1080/14992027.2021.1976424.

Zaar, J., Carney, L.H., 2022. Predicting speech intelligibility in hearing-impaired listeners using a physiologically inspired auditory model. Hear. Res. 426, 108553 https://doi.org/10.1016/j.heares.2022.108553.

Zaar, J., Simonsen, L.B., Dau, T., Laugesen, S., 2023. Toward a clinically viable spectro-temporal modulation test for predicting supra-threshold speech reception in hearing-impaired listeners. Hear. Res. 427, 108650 https://doi.org/10.1016/j.heares.2022.108650.