frontiers | Frontiers in Audiology and Otology

TYPE Brief Research Report
PUBLISHED 06 January 2026
DOI 10.3389/fauot.2025.1746635

OPEN ACCESS

EDITED BY
Jorge Humberto Ferreira Martins,
Escola Superior de Saúde do Alcoitão, Portugal

REVIEWED BY
João Moura,
Coimbra Hospital and University Center, Portugal
Lars Bramsløw,
Eriksholm Research Center, Denmark

*CORRESPONDENCE
Karolin Marija Beatrix Schulz
✉ s243370@dtu.dk

RECEIVED 14 November 2025
REVISED 29 November 2025
ACCEPTED 03 December 2025
PUBLISHED 06 January 2026

CITATION
Schulz KMB, Wittmann H, Hackenberg L, Jürgens T and Tchorz J (2026) Effects of low-latency deep-learning-based noise reduction on speech intelligibility for normal hearing, hearing-impaired listeners and cochlear implant users.
Front. Audiol. Otol. 3:1746635.
doi: 10.3389/fauot.2025.1746635

Effects of low-latency deep-learning-based noise reduction on speech intelligibility for normal hearing, hearing-impaired listeners and cochlear implant users

Karolin Marija Beatrix Schulz[1,2]*, Hannes Wittmann[1,3], Lea Hackenberg[1,4], Tim Jürgens[1] and Jürgen Tchorz[1]

[1]Institute of Acoustics, University of Applied Sciences Lübeck, Lübeck, Germany, [2]Technical University of Denmark, Lyngby, Denmark, [3]Universität zu Lübeck, Lübeck, Germany, [4]German Institute of Hearing Aids, Lübeck, Germany

Objective: To investigate the effects of low-latency deep-learning-based noise reduction on measured and predicted speech intelligibility in noise in normal-hearing (NH) listeners, hearing-impaired (HI) listeners and cochlear implant (CI) users.
Design: A pre-trained convolutional single-channel time-domain audio separation network with ultra-low latency (1 ms) was used to process mixtures of speech from the German matrix sentence test in speech-shaped noise. Speech reception thresholds (SRTs) were measured with and without algorithm processing adaptively aiming at 50% speech understanding. The short-time objective intelligibility measure (STOI) was used for predictions.
Study sample: Fifty-one adults participated in this study: 20 NH, 19 HI listeners and 12 CI users.
Results: For NH listeners, the noise reduction algorithm significantly decreased speech intelligibility in noise (median SRT deterioration: 0.9 dB). In contrast, the SRT significantly improved for HI listeners and CI users (0.8 dB and 5.7 dB, respectively). The strong correlation between individual unprocessed SRT and SRT benefit was closely qualitatively predicted using STOI. Maximum SRT benefit was predicted at +5 dB signal-to-noise ratio.

KEYWORDS

cochlear implants, deep neural networks, hearing aids, noise reduction, speech intelligibility

1 Introduction

Understanding speech in the presence of competing background noise is one of the greatest challenges for persons with hearing impairment. Hearing aids (HA) and cochlear implants (CI) should support communication in such situations and reduce the interference caused by background noise. In situations where speech is spatially separated from noise, beamforming techniques in modern hearing devices improve speech intelligibility substantially (e.g., Jürgens et al., 2025). In addition, or in situations without usable spatial separation, single-channel noise reduction algorithms can be used. They are available in digital HAs since the mid-1990s (Hamacher et al., 2005) and exploit

temporal differences (e.g., in modulation rate and depth) and spectral differences between speech and noise (e.g., with spectral subtraction, Boll, 1979; Ephraim and Malah, 1984). Studies on the benefits of these 'traditional' single-channel methods have generally not shown any improvement in speech intelligibility in noise for HA users. Instead, only moderate improvements in hearing comfort, acceptable noise level and speech quality have been demonstrated (Lakshmi et al., 2021). For CI users small, but significant speech intelligibility benefits were shown with traditional single-channel noise reduction algorithms (e.g., Büchner et al., 2010).

Recently, machine learning methods have been used for single-channel speech enhancement, where artificial neural networks are trained with noisy speech from a wide variety of speakers, noise types and signal-to-noise ratios (SNR) to output clean speech. In particular, deep neural networks (DNN) have led to dramatic performance improvements in speech enhancement (Ochieng, 2023).

There are two important groups of DNN-based speech enhancement algorithms: spectral domain approaches and time domain approaches.

DNN models that are trained via supervised learning on spectral features derived from Short-Term Fourier Transformation (STFT) usually employ spectral masking or spectral mapping techniques to generate an estimated clean signal from a noisy signal. The role of the DNN-derived mask is hereby to indicate the target source in given bins of the time-frequency representation. The mask is applied to the time-frequency bins to reduce areas dominated by noise prior to inverse STFT which yields an estimate of the clean speech (Wang et al., 2014). Alternatively, or in addition, the phase of the noisy input signal or its complex spectrum can be used as input to the DNN (Erdogan et al., 2015; Williamson and Wang, 2017). Spectral mapping techniques directly estimate the spectrogram of the pure speech signal spectrogram instead of a mask.

In these algorithms based on spectral features, the time-domain signal is divided into short segments for subsequent STFTs. The algorithm latency therefore equals the length of the time segment and thus increases with spectral resolution. To achieve good performance, many algorithms use 32 ms segments or more (Ochieng, 2023). In HAs, however, the total processing latency, which includes additional steps such as AD/DA conversion, beamforming and compression, should not exceed 20–30 ms (Stone and Moore, 2005) for closed earmolds, or 5–6 ms for open fittings (Stone et al., 2008), since direct sound and processed sound interfere and lead to disturbance reducing the sound quality. Thus, the tradeoff between low latency and spectral resolution limits the applicability of spectral domain approaches in HAs. For CI users, tolerable delays are much higher, at 200–250 ms (Hay-McCutcheon et al., 2009), due to the absence of direct sound interference.

In DNN-based time domain algorithms without explicit feature extraction latencies are lower (around 4 ms and less), due to the size of the analysis windows. The basic concept of these algorithms is to replace the discrete Fourier transform (DFT)-based input with a data-driven representation that is jointly learnt during model training. The models therefore accept as input the raw mixed waveform and then generate either the estimated clean sources end-to-end or generate masks applied to the noisy waveform to obtain clean sources (Luo and Mesgarani, 2019; Subakan et al., 2021; Ochieng, 2023). The algorithm hereby needs to consider the temporal structure of speech.

Simple feed-forward DNN architectures are only partially suitable for speech data, as they cannot effectively model long-time dependencies that are present in the speech data. The use of convolutional neural networks (CNN), which were originally developed for the processing and classification of image and video data, is widespread for speech enhancement. CNNs use convolutional layers to recognize local patterns and build them up hierarchically. The U-Net is a special type of CNN originally proposed by Ronneberger et al. (2015) for biomedical image segmentation. Its encoder-decoder structure with skip connections allows efficient learning from limited data. This makes it particularly attractive for domains such as single-channel noise suppression, where labeled data is often scarce.

The performance of DNNs can be determined technically using objective measures such as Short-Time Objective Intelligibility (STOI; Taal et al., 2011) or Perceptual Evaluation of Speech Quality (PESQ; Rix et al., 2001), which predict speech intelligibility and subjectively perceived speech quality, respectively. Behaviourally, DNN-based single-channel noise reduction algorithms showed considerable improvements in speech intelligibility scores for both NH and hearing-impaired (HI) listeners (Healy et al., 2013, 2015) in offline evaluations. For HI listeners, these improvements in the range of up to 50% higher speech intelligibility scores in noise persisted even when further developing the algorithm toward real-time application using talker-independent training (Healy et al., 2021). Although in principle real-time capable with relatively low latency, the computational load for this algorithm was still too high for actual application in head-worn hearing devices.

A DNN-based algorithm (Andersen et al., 2021) that is less computationally expensive and already implemented on HAs, was shown to result in smaller, but still significant improvements for speech reception threshold (SRT) of a speech-in-noise test for HI listeners (<1 dB). Diehl et al. (2023a) proposed another spectral-domain DNN-based algorithm with a 23 ms STFT analysis window. Their result, however, showed the potential for even higher improvements for HI listeners of up to 5 dB lower SRTs.

For CI users, considerable speech intelligibility improvements were shown with neural-network speech enhancement algorithms (Goehring et al., 2017), and in particular with DNNs (Lai et al., 2018; Gajecki and Nogueira, 2024; Borjigin et al., 2024). These algorithms were tested offline, i.e., without implementing them (yet) on the head-worn speech processor either due to latency and computational complexity (Lai et al., 2018; Borjigin et al., 2024), or because current commercial speech processors do not allow for the high necessary wireless data rates (Gajecki and Nogueira, 2024).

Abbreviations: AD/DA, analog-digital/digital-analog; B&K, Brüel&Kjær; CI, cochlear implant; CNN, convolutional neural network; dB, decibel; DNN, deep neural network; DFT, discrete Fourier transform; DNN, deep neural network; HA, hearing aid; HI, hearing-impaired; HL, hearing level; NH, normal-hearing; OLSA, Oldenburg sentence test; PESQ, perceptual evaluation of speech quality; SNR, signal-to-noise ratio; SRT, speech reception threshold; STFT, short-time Fourier transform; STOI, short-time objective intelligibility measure.

The present study uses a DNN noise reduction algorithm working in the time domain with very low latency (1 ms). To the authors' knowledge no study has assessed such a low-latency DNN noise reduction algorithm within all three groups of listeners, NH, HI listeners and CI users and compared them. This has the advantage of covering a large range of speech-in-noise performance values, investigating the potential of the algorithm at different SNRs and across different hearing abilities.

The goal of this study is therefore to predict and assess speech intelligibility in background noise with and without this state-of-the-art low-latency DNN noise reduction algorithm in three groups of listeners: NH listeners, HI listeners and CI users. The null hypothesis was that DNN noise suppression would not affect SRTs and that there would be no differences between the three groups of subjects.

2 Materials and methods

2.1 Participants

Fifty-one adult German native-speaking subjects participated in this study, who were distributed into three groups. Twenty NH listeners (age 21–54, mean 27.8 years, 9 female, 11 male) had pure-tone thresholds of not more than 25 dB hearing level (HL) for all octave frequencies between 500 Hz and 8,000 Hz, as tested with an Affinity 2 (Interacoustics, Middelfart, Denmark) audiometer. Nineteen HI listeners (age 36–78, mean 67.5 years, 3 female, 16 male) had mild to moderate symmetrical hearing impairment with pure-tone threshold differences between left and right ear (between 500 Hz and 4,000 Hz) not exceeding 15 dB. All HI listeners were experienced HA users (>2 years experience). Audiometric thresholds of the tested ears are shown in Figure 1. Twelve postlingually deafened CI users (age 23–82, mean 56.8 years, 2 female, 10 male) participated with using their own speech processors. Two CI participants (#5 and #12) experienced connection problems with their audio streamer and could not finish speech tests in both conditions. No speech test results were taken into account for these two participants in order to maintain a paired within-groups design. Demographic details of the CI users can be found in Table 1. All participants signed written consent before the measurements. HI listeners and CI users received monetary compensation for their participation. Ethical approval was granted by the Ethics committee of the University of Applied Sciences Lübeck (approval 311.012.17 from 11. September 2018).

2.2 Apparatus

Stimuli were generated on a standard PC connected to a Fireface 800 audio interface (RME, Haimhausen, Germany). NH and HI listeners were listening to the stimuli presented diotically via Sennheiser (Wedemark, Germany) HD 555 headphones. Calibration of the headphones was done using an artificial ear (Brüel & Kjær, B&K type 4153 with 1/2″ microphone B&K type 4192, connected to sound level meter B&K type 2250). For CI users the signal was streamed to their own CI speech processor via the Audiolink (MED-EL, Innsbruck, Austria) interface or the Cochlear (Sydney, Australia) TV-streamer.

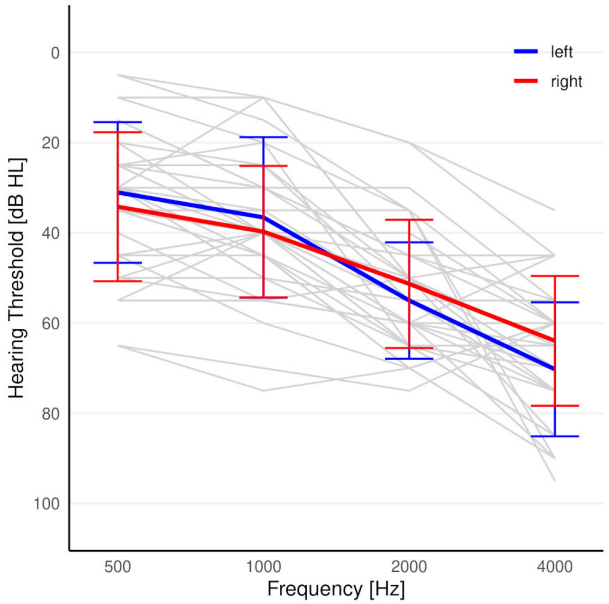

FIGURE 1
Audiometric thresholds of the HI listeners' tested ears. The colored lines and error bars indicate mean and standard deviation.

2.3 Speech-in-noise test

The German matrix sentence test (Oldenburger Satztest, OLSA, Wagener et al., 1999) was used for speech-in-noise testing. This test consists of sentences in the fixed word order subject, verb, number, adjective, object, whereas each of these five words can be realized in 10 different response alternatives. Speech signals were mixed with stationary speech-shaped noise ("olnoise"), with keeping the speech level constant and adaptively varying the noise level (and thus the SNR) over the course of the measurement run, aiming at 50% speech understanding. A start SNR of +10 dB was used. The experimenter marked the words that the participants understood on a touch screen. The entire test was implemented using customized scripts in MATLAB (The Mathworks, Natick, USA). After a briefing about the testing procedure, a training run with 20 sentences (not processed by the DNN speech enhancement) was used to familiarize the listener with the adaptive conduction of the test. The result of this training run was not evaluated for the present study. For CI users the training run also served to set the volume of their own speech processor being used for streaming to most comfortable loudness. NH listeners were tested at a fixed speech level of 65 dB (A) and HI listeners were tested at a fixed speech level of 80 dB (A). Prior to the speech test in noise, a speech test was conducted in quiet at 80 dB speech level in the HI group, using the Freiburger Test. Subjects who did not reach at least 80% correct in this condition were excluded from the study. This was to ensure that the SNR for 50% speech understanding in noise is not affected by insufficient audibility of speech at 80 dB. The adaptive

TABLE 1 Demographic details of the CI users.

Subject	Age (years)	Gender	Implant side	Implant use (years)	Manufacturer	Processor	contra-lateral side
01	23	M	L	15	Cochlear	Nucleus 7	CI
02	65	M	L	4	Cochlear	Nucleus 7	CI
03	74	M	R	3	MED-EL	Sonnet 2	CI
04	82	M	L	8	MED-EL	Sonnet 2	CI
05	26	M	R	22	Cochlear	Nucleus 6	Deaf
06	63	F	L	5	Cochlear	Nucleus 7	Deaf
07	56	M	L	3	Cochlear	Nucleus 7	HA
08	38	M	L	5	Cochlear	Nucleus 7	HA
09	74	M	L	4	Cochlear	Nucleus 7	HA
10	67	M	R	3	Cochlear	Nucleus 7	NH
11	62	M	L	5	Cochlear	Nucleus 7	NH
12	51	F	L	7	Cochlear	Kanso 2	NH
Median	62.5			7			

OLSA was then conducted using randomly chosen 20-sentence-lists with and without single-channel DNN speech enhancement for each subject in randomized order.

2.4 Single-channel DNN speech enhancement

For the DNN condition, the sentences from the German matrix test were mixed with olnoise at SNRs ranging from −16 to +16 dB in 1 dB increments. The noisy test material was processed using Conv-TasNet (Luo and Mesgarani, 2019), a fully convolutional single-channel time-domain audio separation network that is here used for noise suppression. Unlike STFT-based methods, it learns optimal representations directly from waveforms. Full details can be found in Luo and Mesgarani (2019).

In the present study, a pre-trained open-source Conv-TasNet model was used without further modification (Keles, 2022). The encoder had 512 filters with a kernel size of 16 samples operating on overlapping frames with a stride of 8 samples to extract temporal features that implicitly capture spectral information. Our analysis of the algorithm's impulse response confirmed 16-sample (1 ms) look-ahead, i.e. an algorithmic delay of 1 ms. This matches the size of the kernels (16 samples with a sample rate of 16,000). A detailed description of the training parameters can be found in Keles (2022).

The training material used by Keles (2022) consisted of 364 h of speech from 921 speakers, derived from English audiobooks from the LibriVox project (LibriVox, n.d.), to which reverberation was added. The background noise data set (30 h) was recorded in urban environments such as coffee shops, restaurants, bars, office buildings, and parks (Maciejewski et al., 2020). The noise data set was augmented using time stretching (factor 0.8–1.2). The training material had a sampling rate of 16 kHz and the mean SNR was −2 dB with a standard deviation of 3.6 dB. Note that neither the speech, nor the noise that is used in the testing was used within the training. After processing noisy speech with this pre-trained DNN model, the background noise is almost completely removed. The amount of distortion in the speech signal depended on the SNR before processing.

2.5 Objective measurements

In addition to subjective speech tests, the STOI (Taal et al., 2011) was used to predict speech intelligibility. Therefore, 120 matrix test sentences were mixed with olnoise and babble noise at SNRs ranging from −16 to +16 dB in 1 dB increments. Unlike the olnoise, the babble noise was not used to measure speech intelligibility in the subjects, but served as a reference for how predictions would transfer to another non-stationary noise. The source of the babble noise was about 100 people talking in a canteen. It has fast temporal variations and individual voices are audible. The resulting 3,960 mixed sentences (120 sentences × 33 SNRs) for each noise were then processed with the DNN speech enhancement. STOIs were calculated for all sentences, both processed and unprocessed.

2.6 Statistical testing

For sample size estimation, a large detectable effect ($\eta^2 = 0.26$) and a power of 0.8 was specified for differences between groups, which resulted in a minimum sample size of 11 per group. SRT data was tested for normal distribution using the Shapiro–Wilk test. The paired-samples Wilcoxon signed-rank test with Holm-correction for multiple pairwise comparisons was used to investigate SRT-differences with and without the DNN noise reduction algorithm within each participant group. The Kruskal–Wallis rank sum test was used to detect SRT differences and SRT benefit differences between the three participant groups, and the Dunn test for

10.3389/fauot.2025.1746635

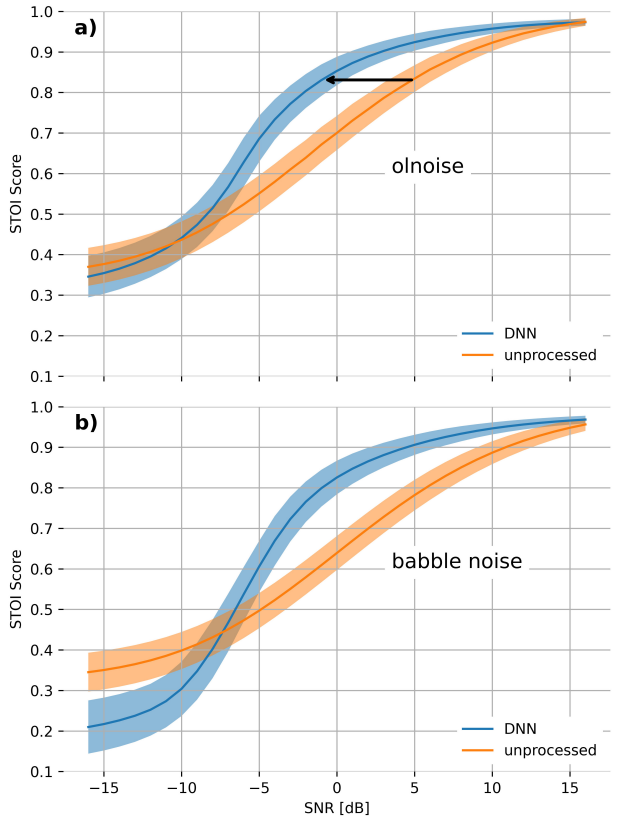

FIGURE 2
Mean STOI values as a function of input SNR for **(a)** olnoise and **(b)** babble noise. Shaded areas indicate standard deviation of STOI values for 120 sentences for each SNR. The length of the black arrow illustrates the estimated SNR improvement in olnoise for an unprocessed input SNR of 5 dB.

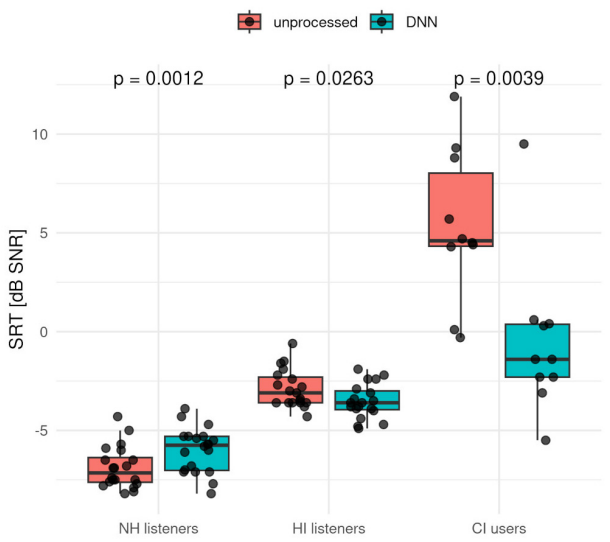

FIGURE 3
Boxplots and individual SRTs for NH listeners, HI listeners and CI users in unprocessed conditions and with using DNN-based speech enhancements. Boxes show interquartile ranges with the median being indicated as horizontal bar. Whiskers extend to 1.5 times the box length.

post-hoc analysis (p-values adjusted with the Holm method). A 5% significance level was chosen for all statistical tests (two-sided, if applicable). All testing was done using R (Version 4.3.1, R Foundation for Statistical Computing) with RStudio (Version 23.06.2, Posit Software, PBC).

3 Results

Figure 2 shows the predicted speech intelligibility (STOI) in olnoise and babble noise for the unprocessed and DNN conditions as a function of SNR. STOI predicts improved speech intelligibility for input SNRs in the range between −8 dB and +13 dB for the olnoise and between −7 dB and +16 dB for the babble noise. The predicted SNR benefit from DNN processing can be calculated from the horizontal distance between the two curves for a given unprocessed input SNR: The algorithm compensates for the reduced input SNR and enables equally good speech intelligibility. The largest predicted benefit from DNN processing in terms of SRT improvement (i.e., the largest horizontal distance between the two curves for a given unprocessed input SNR) is found for input SNRs of around +5 dB in both noise types.

Figure 3 shows boxplots and individual SRTs in olnoise for the three groups of subjects, for unprocessed speech and after DNN speech enhancement. A lower SRT indicates better speech-in-noise performance. Normal distribution of SRT scores could not be assumed in the CI/DNN condition ($W = 0.81$, $p = 0.017$). All other conditions showed non-significant results in the test for normal distribution (NH/unprocessed: $W = 0.91$, $p = 0.08$, NH/DNN: $W = 0.97$, $p = 0.69$, HI/unprocessed: 0.92, $p = 0.11$, HI/DNN: $W = 0.96$, $p = 0.48$, CI/unprocessed: $W = 0.94$, $p = 0.5$).

Without processing, the median SRT in NH listeners is −7.2 dB SNR. DNN processing significantly increased the median SRT to −5.8 dB SNR for NH listeners, i.e., produces poorer speech understanding. In HI subjects, in contrast, DNN processing significantly decreased the median SRT from −3.1 dB to −3.6 dB SNR, i.e., improved speech understanding. CI users had the highest median SRT in the unprocessed condition (4.6 dB), which significantly decreased (i.e., improved) to −1.4 dB. More details on within groups comparisons are listed in Table 2.

In the following, "SRT benefit" refers to individual SNR change with opposite sign, so that positive numbers indicate an improvement in speech intelligibility. The median SRT benefits for NH listeners, HI listeners and CI users are −0.9 dB, 0.8 dB and 5.7 dB, respectively. Thus, speech intelligibility in persons with hearing impairment, in particular in CI users, improves with DNN-based speech enhancement. The Kruskal–Wallis tests showed significant differences between the three groups for unprocessed SRTs, SRTs after DNN processing, and SRT-benefits ($p < 0.001$). The results of subsequent *post-hoc* analysis are summarized in Table 3. All pairwise differences between groups are significant, except the SRTs after DNN processing between HI listeners and CI users. This means that CI users gain more benefit from DNN processing

10.3389/fauot.2025.1746635

TABLE 2 Results of within-group differences.

Group	N	Mean SRT un-process (dB)	SD SRT un-process (dB)	Mean SRT DNN (dB)	SD SRT DNN (dB)	Mean diff (dB)	SD diff (dB)	Median diff (dB)	CI95 (dB)	Rank biserial	Wilcox W	p	p_{adj}
NH	20	−6.9	1.1	−6.0	1.1	0.9	0.9	0.9	[0.4, 1.4]	0.829	14	0.001	0.004
HI	19	−2.9	1.0	−3.5	0.9	−0.7	1.1	−0.8	[−1.1, 0]	−0.589	137	0.026	0.026
CI	10	5.3	3.9	−0.5	4.0	−5.9	3.8	−5.7	[−7.4, −3.7]	−0.964	54	0.004	0.008

The difference (diff) is the individual SRT change after processing.

TABLE 3 Results of *post-hoc* analysis of pairwise between-groups differences using the Dunn-test and Cliff's delta.

Group difference	SRT unprocessed	SRT DNN	SRT benefit
CI – NH	$Z = 6.1$	$Z = 5.3$	$Z = -5.3$
	$p = 8 \times 10^{-10}$	$p = 1 \times 10^{-7}$	$p = 1 \times 10^{-7}$
	$p_{adj} = 2 \times 10^{-9}$	$p_{adj} = 3 \times 10^{-7}$	$p_{adj} = 4 \times 10^{-7}$
	$\Delta = -1$	$\Delta = -0.93$	$\Delta = 0.96$
CI – HI	$Z = 2.6$	$Z = 1.8$	$Z = -2.6$
	$p = 0.009$	$p = 0.071$	$p = 0.009$
	$p_{adj} = 0.009$	$p_{adj} = 0.071$	$p_{adj} = 0.009$
	$\Delta = -1$	$\Delta = -0.7$	$\Delta = 0.84$
NH – HI	$Z = -4.2$	$Z = -4.2$	$Z = 3.2$
	$p = 2 \times 10^{-5}$	$p = 2 \times 10^{-5}$	$p = 0.001$
	$p_{adj} = 4 \times 10^{-5}$	$p_{adj} = 4 \times 10^{-5}$	$p_{adj} = 0.003$
	$\Delta = -0.997$	$\Delta = -0.93$	$\Delta = 0.71$

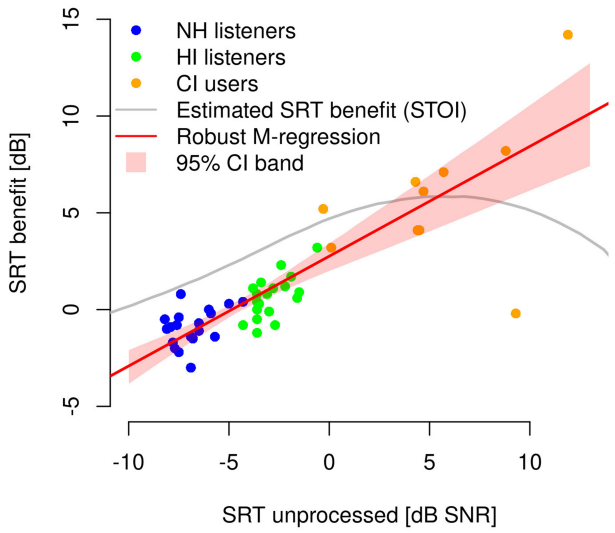

FIGURE 4
Scatterplot of individual SRT benefits as a function of SRT in unprocessed condition for all participating subjects. Negative SRT benefits indicate degradation of speech intelligibility after DNN processing. The robust linear regression is shown as a red line. SRT benefit predicted using STOI is shown as continuous gray line.

than HI listeners and achieve speech intelligibility in noise after DNN processing, which is not significantly different from that of HI listeners.

A more detailed view on the individual data is presented in Figure 4 showing the individual SRT benefit as a function of the individual SRT in the unprocessed condition. There is a clear trend of higher benefit from DNN processing for higher unprocessed SRT. Visual inspection of residual plots confirmed an approximately linear relationship between unprocessed SRT and SRT benefit across the observed range. The Shapiro–Wilk test indicated significant departure from normality in the residuals (W

= 0.82, $p < 0.001$), primarily driven by heavy-tailed distributions and the presence of outliers. The CI group exhibits markedly higher variability compared to NH and HI groups. Formal testing using the Breusch–Pagan test indicated significant heteroscedasticity (BP = 13.1, $p < 0.001$). This violation of homoscedasticity motivated the use of robust *M*-estimation regression. The robust regression analysis revealed a statistically significant linear relationship ($p <$ 0.05 based on bootstrapping, intercept: 2.76, 95% CI for intercept: [1.98, 3.42], slope: 0.57, 95% CI for slope: [0.41, 0.72], standardized slope: 0.89, correlation-based robust *R*-squared: 0.712). In addition, group and interaction models were tested, which led to minor improvements of robust *R*-squared values (0.725 for group and 0.729 for interaction).

The STOI-predicted SRT benefit (horizontal distance between STOI-values in DNN vs. unprocessed condition) is plotted as a thin continuous gray line. For unprocessed SRTs below 0 dB, STOI predicts an increase of SRT benefit of about 0.5 dB for every 1 dB increase in SRT. Note that an SRT benefit of 0 dB is predicted for very low SRTs around −10 dB. The largest SRT benefit is predicted for unprocessed SRTs of around 5 dB. At higher SNRs the curve bends downwards predicting smaller SRT benefits. Note the general offset of SRT benefit predictions vs. SRT benefit measurements: STOI predicts about 2–3 dB higher SRT benefit for NH and HI listeners than is achieved on average.

4 Discussion

The present study evaluated the effectiveness of a state-of-the-art, low-latency DNN-based noise reduction algorithm in improving speech intelligibility in background noise for three listener groups: NH listeners, HI listeners, and CI users. The results demonstrate a significant SRT benefit for HI listeners and CI users, while NH listeners experienced a slight (but statistically significant) degradation in performance. The CI group was small ($N = 10$), but both between-groups and within-groups effects were large enough to be detected. The median benefit in the HI group (0.8 dB) can be regarded as clinically relevant, given the slope of 17% points per dB of the psychometric function of the speech test (Wagener et al., 1999). Pairwise group comparisons revealed substantial effect sizes for SRT measures across listener groups. Cliff's delta values indicated large effects ($|\Delta| \geq 0.7$) between groups in all comparisons.

STOI-based predictions closely replicate the trend of SRT benefits as a function of SNR found among participants.

The unprocessed mean SRT of −6.9 dB for NH listeners aligns with normative values for the German matrix test in olnoise (Wagener et al., 1999), confirming the validity of our experimental setup. Notably, the DNN processing led to a small but significant worsening of speech intelligibility for NH listeners. This may be attributed to the algorithm's relative aggressive noise suppression, which could introduce artifacts or distortions detrimental to NH listeners' performance in those poor SNR conditions. In contrast, HI listeners showed a modest but significant median improvement in SRT (0.8 dB), consistent with previous findings [e.g., 0.5 dB−1 dB found in Andersen et al. (2021)]. The most striking benefit was observed among CI users, with a median SRT improvement of 5.7 dB, bringing their performance close to that of HI listeners. This underscores the large potential of DNN-based noise reduction for CI users, as also reported by Goehring et al. (2017) and Borjigin et al. (2024). The greater benefit for CI users may reflect their higher baseline SRTs, where the algorithm's noise suppression is most effective, as predicted by STOI.

Robust regression indicates 0.57 dB more SRT benefit for every 1 dB higher input SRT. The simple robust regression model was preferred over more complex alternatives because it provided comparable explanatory power while maintaining superior interpretability and parsimony. Although models including group factors and interaction terms showed marginal improvements in R^2 (0.712 vs. 0.729), the minimal gain did not justify the substantial increase in complexity and reduced generalizability. The standardized coefficient of slope (0.89) represents a large effect size according to conventional guidelines.

Our findings share similarities with Diehl et al. (2023a), particularly in the trend of greater benefits for listeners with poorer baseline performance (see their Figure 3c in comparison with Figure 4 here). While Diehl et al. (2023a)'s trend suggests a steeper increase of SRT benefit with increasing SNR (about 0.7 dB/dB) compared to our 0.57 dB/dB, their trend is not based on many data points above +4 dB SNR, which makes their study less robust in this SNR range. Another critical distinction lies in the latency of the algorithm used in this study (1 ms) compared to 23 ms in Diehl et al. (2023a). The ultra-low latency of the present study makes time-domain DNN noise suppression approach suitable for real-time applications in hearing devices, where stringent latency constraints exist (<20 ms for HAs with closed acoustic couplings; Stone and Moore, 2005; <6 ms for HAs with open couplings; Stone et al., 2008). Diehl et al. (2023a) used a laptop for noise suppression and a modified version of their approach has been implemented on a smartphone streaming the de-noised signal to HAs (Diehl et al., 2023b). For the use of DNNs for single-channel noise cancellation in HAs, it is necessary that the sum latency (which does also include AD/DA-conversion and further algorithmic latency) remains short enough. In addition, the complexity of the DNN in terms of parameters and operations per second must match the available capacity of the HA or CI hardware without unacceptable compromises in performance.

Saoji et al. (2024) assessed speech perception in multi-talker babble noise at 5 and 10 dB SNR in ten patients with severe hearing loss (67 dB better-ear pure-tone average). Their DNN-based noise suppression algorithm which is implemented in commercially available hearing aids and which is a modified version of the one used by Diehl et al. improved speech perception by 20 to 32 percentage points, compared to the baseline program. The authors concluded that noise reduction using DNNs might prevent some patients with significant residual hearing from qualifying for cochlear implantation, which could delay its adoption or render it unnecessary. On the other hand, the strong improvements for CI users shown in the present study indicate that cochlear implantation in combination with DNN-based noise reduction may offer for many CI users similar performance as HI listeners have with HAs.

The Conv-TasNet algorithm used here achieves noise suppression with minimal latency by operating directly in the time domain, avoiding the spectral resolution trade-offs of STFT-based methods. While the model's size and computational load (number

of floating point operations) were not a limiting factor in the present study (the processing time was 4 times faster than real time on a standard laptop), implementing such algorithms in hearing devices requires careful optimization to match hardware limitations. Recent advances in time-domain DNNs (e.g., Subakan et al., 2021) suggest a rapidly evolving field, with newer models offering further reductions in complexity and latency. Despite these advancements, widespread adoption of DNN-based speech enhancement in HAs and CIs remains limited due to hardware constraints. Only a few HA models currently incorporate such features, highlighting the need for continued development of efficient, low-power DNN architectures tailored to these devices.

4.1 Limitations and future directions

The use of a pre-trained model "off the shelf" (Keles, 2022) that was trained on English speech and environmental noises, and applying it to a German clinical test scenario with the OLSA and stationary noise matching the long-term speech spectrum, raises questions about whether or not the algorithm was optimally trained for this purpose. On the other hand, this approach raises fewer questions about generalizability than studies with DNN-based noise reduction algorithms so far (e.g., Healy et al., 2013, 2015), because neither the tested language, nor the tested talker, nor the tested noise were used during training. Nevertheless, the generalizability to other languages and noise types remains an open question. Future work should explore fine-tuning of the algorithm with German speech data of many different talkers or other noise scenarios to explore further improvements of using this algorithm within the German market. Additionally, subjective measures of sound quality and listening effort could provide further insights into the algorithm's perceptual effects that may extend beyond the SNR range tested in the present study.

The OLSA speech level was different for NH and HI group to avoid uncomfortable sound levels in the NH group and to ensure audibility in the HI group. Although the SRT was measured (which is an SNR), an effect of different speech levels cannot be excluded. In the CI group, the speech level was individually adjusted to ensure comfortable loudness via streaming devices from two different manufactures. Compression or bandwidth differences might have had an influence on measured SRTs, but these aspects have not been further investigated.

Another limitation of the present study concerns the use of STOI. It was used to predict the effect of noise reduction on SRTs, which means that the part of algorithmic SNR-benefit that is realizable by the listeners in the form of speech intelligibility benefit is predicted. As STOI does not consider additional aspects of audibility, the present study made sure that audibility in quiet condition was sufficient in each subject. However, since hearing loss compensation was not done on an individual basis, perfect audibility was probably not achieved in each HI listener. The overestimation of the SRT-benefit by STOI for HI listeners might therefore partly be due to slight audibility effects. Since STOI only uses temporal envelope features for its prediction of speech intelligibility (Taal et al., 2011), this overestimation may also be related to distortions in the temporal fine-structure created by the relatively aggressive DNN-based algorithm that may reduce speech intelligibility in NH listeners (Xu et al., 2017) and potentially also in HI listeners.

The trend of higher SRT benefit with higher input SNR up to about 5 dB is well captured by this approach. The study presented here contains only limited data for unprocessed SRTs above 5 dB SNR. Therefore, no valid statements can be made about the course of behavioral SRT benefits in this area. The reliability of the SRT benefit predictions at very low (<-5 dB) and very high (>10 dB) SNRs may also be questionable due to ceiling and flooring effects with the raw STOI values in both processed and unprocessed conditions (see Figure 2).

5 Conclusions

This study investigated the effect of a low-latency DNN-based noise reduction for improving SRTs in noise for NH listeners, HI listeners, and CI users and HI listeners. The open source DNN algorithm was taken "off-the shelf", i.e., without specific training on the noise or the speech involved in testing. The following conclusions can be drawn.

- There were significantly poorer SRTs found for NH listeners in the processed condition, i.e., speech intelligibility degraded (SRT benefits: median −0.9 dB, 95% CI: [−1.4, −0.4], $W = 14$). There were small significant SRT benefits for HI listeners (median 0.8 dB, 95% CI: [0, 1.1], $W = 137$) and considerable SRT benefits for CI users found (median 5.7 dB, 95% CI: [4.1, 7.4], $W = 54$).
- Individual SRT benefit was strongly correlated with the individual SRT in the unprocessed condition, meaning that poor-performing subjects benefited the most from the algorithm. Predictions using STOI closely mirrored this trend, but with an offset overestimation of 2–3 dB.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors upon reasonable request.

Ethics statement

The studies involving humans were approved by Ethics committee of the University of Applied Sciences Lübeck. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

KS: Formal analysis, Methodology, Writing – review & editing, Investigation, Software. HW: Formal analysis, Methodology, Writing – review & editing, Investigation,

Software. LH: Writing – review & editing, Investigation, Resources. TJ: Formal analysis, Methodology, Supervision, Project administration, Conceptualization, Writing – original draft. JT: Methodology, Formal analysis, Visualization, Conceptualization, Writing – review & editing, Supervision.

Funding

The author(s) declared that financial support was received for this work and/or its publication. The authors acknowledge financial support for publication fees by Land Schleswig-Holstein within the funding programme Open Access Publikationsfonds.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. Generative AI was used to support statistical programming in R (ChatGPT Code Pilot) and linguistic expression (DeepL).

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

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

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* 27, 113–120. doi: 10.1109/ICASSP.1979.1170696

Borjigin, A., Kokkinakis, K., Bharadwaj, H. M., and Stohl, J. S. (2024). Deep learning restores speech intelligibility in multi-talker interference for cochlear implant users. *Sci. Rep.* 14:13241. doi: 10.1038/s41598-024-63675-8

Büchner, A., Brendel, M., Saalfeld, H., Litvak, L., Frohne-Büchner, C., and Lenarz, T. (2010). Results of a pilot study with a signal enhancement algorithm for HiRes 120 cochlear implant users. *Otol. Neurotol.* 31, 1386–1390. doi: 10.1097/MAO.0b013e3181f1cdc6

Diehl, P. U., Singer, Y., Zilly, H., Schönfeld, U., Meyer-Rachner, P., Berry, M., et al. (2023a). Restoring speech intelligibility for hearing aid users with deep learning. *Sci. Rep.* 13:2719. doi: 10.1038/s41598-023-29871-8

Diehl, P. U., Zilly, H., Sattler, F., Singer, Y., Kepp, K., Berry, M., et al. (2023b). Deep learning-based denoising streamed from mobile phones improves speech-in-noise understanding for hearing aid users. *Front. Med. Eng.* 1:1281904. doi: 10.3389/fmede.2023.1281904

Ephraim, Y., and Malah, D. (1984). Speech enhancement using a minimum mean square error short-time spectral amplitude estimation. *IEEE Trans. Acoust. Speech Signal Process.* 32, 1109–1121. doi: 10.1109/TASSP.1984.1164453

Erdogan, H., Hershey, J. R., Watanabe, S., and Le Roux, J. (2015). "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brisbane, Australia), 708–712. doi: 10.1109/ICASSP.2015.7178061

Gajecki, T., and Nogueira, W. (2024). A fused deep denoising sound coding strategy for bilateral cochlear implants. *IEEE Trans. Biomed. Eng.* 71, 2232–2242. doi: 10.1109/TBME.2024.3367530

Goehring, T., Bolner, F., Monaghan, J. J. M., Van Dijk, B., Zarowski, A., and Bleeck, S. (2017). Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. *Hear. Res.* 344, 183–194. doi: 10.1016/j.heares.2016.11.012

Hamacher, V., Chalupper, J., Eggers, J., Fischer, E., Kornagel, U., Puder, H., et al. (2005). Signal processing in high-end hearing aids: state of the art, challenges, and future trends. *EURASIP J. Appl. Signal Process.* 18, 2915–2929. doi: 10.1155/ASP.2005.2915

Hay-McCutcheon, M. J., Pisoni, D. B., and Hunt, K. K. (2009). Audiovisual asynchrony detection and speech perception in hearing-impaired listeners with cochlear implants: a preliminary analysis. *Int. J. Audiol.* 48, 321–333. doi: 10.1080/14992020802644871

Healy, E. W., Tan, K., Johnson, E. M., and Wang, D. (2021). An effectively causal deep learning algorithm to increase intelligibility in untrained noises for hearing-impaired listeners. *J. Acoust. Soc. Am.* 149, 3943–3953. doi: 10.1121/10.0005089

Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., and Wang, D. (2015). An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type. *J. Acoust. Soc. Am.* 138, 1660–1669. doi: 10.1121/1.4929493

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (2013). An algorithm to improve speech recognition in noise for hearing-impaired listeners. *J. Acoust. Soc. Am.* 134, 3029–3038. doi: 10.1121/1.4820893

Jürgens, T., Ihly, P., Tchorz, J., Nishiyama, T., Tanaka, C., Suzuki, D., et al. (2025). Closedness of acoustic coupling and audiological measures are associated with individual speech-in-noise benefit from noise reduction in hearing aids. *Trends Hear.* 29:23312165251325983. doi: 10.1177/23312165251325983

Keles, M. C. (2022). *ConvTasNet Distribution on Online Platform huggingface.co*. Available online at: https://huggingface.co/cankeles/ConvTasNet_WHAMR_enhsingle_16k (Accessed November 3, 2025).

Lai, Y. -H., Tsao, Y., Lu, X., Chen, F., Su, Y. -T., Chen, K. -C., et al. (2018). Deep learning-based noise reduction approach to improve speech intelligibility for cochlear implant recipients. *Ear Hear.* 39, 795–809. doi: 10.1097/AUD.0000000000000537

Lakshmi, M. S. K., Rout, A., and O'Donoghue, C. R. (2021). A systematic review and meta-analysis of digital noise reduction hearing aids in adults. *Disabil. Rehabit. Assist. Technol.* 16, 120–129. doi: 10.1080/17483107.2019.1642394

LibriVox. (n.d.). *LibriVox|Free public Domain Audiobooks*. Available online at: https://librivox.org (Accessed December 12, 2025).

Luo, Y., and Mesgarani, N. (2019). Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* 27, 1256–1266. doi: 10.1109/TASLP.2019.2915167

Maciejewski, M., Wichern, G., McQuinn, E., and Le Roux, J. (2020). "WHAMR!: noisy and reverberant single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona, Spain), 696–700. doi: 10.1109/ICASSP40776.2020.9053327

Ochieng, P. (2023). Deep neural network techniques for monaural speech enhancement and separation: state of the art analysis. *Artif. Intell. Rev.* 56, 3651–3703. doi: 10.1007/s10462-023-10612-2

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Salt Lake City, USA), 749–752. doi: 10.1109/ICASSP.2001.941023

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, eds N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi (Berlin: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Saoji, A. A., Sheikh, B. A., Bertsch, N. J., Goulson, K. R.;, Graham, M. K., McDonald, E. A., et al. (2024). How does deep neural network-based noise reduction in hearing aids impact cochlear implant candidacy? *Audiol. Res.* 14, 1114–1125. doi: 10.3390/audiolres14060092

Stone, M. A., and Moore, B. C. J. (2005). Tolerable hearing aid delays. IV. Effects on subjective disturbance during speech production by hearing-impaired subjects. *Ear Hear.* 26, 225–235. doi: 10.1097/00003446-200504000-00009

Stone, M. A., Moore, B. C. J., Meisenbacher, K., and Derleth, R. P. (2008). Tolerable hearing aid delays. V. Estimation of limits for open canal fittings. *Ear Hear.* 29, 601–617. doi: 10.1097/AUD.0b013e3181734ef2

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., and Zhong, J. (2021). "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 21–25 June 2021* (Toronto, Canada), 21–25. doi: 10.1109/ICASSP39728.2021.9413901

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Proc.* 19, 2125–2136. doi: 10.1109/TASL.2011.2114881

Wagener, K., Brand, T., and Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests (Development and evaluation of a sentence test for German language III: evaluation of the Oldenburg sentence test). *Zeitschrift für Audiologie/Audiol. Acoust.* 38, 86–95.

Wang, Y., Narayanan, A., and Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP)* 22, 1849–1858. doi: 10.1109/TASLP.2014.2352935

Williamson, D. S., and Wang, D. (2017). Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP)* 25, 1492–1501. doi: 10.1109/TASLP.2017.2696307

Xu, Y., Chen, M., LaFaire, P., Tan, X., and Richter, C. -P. (2017). Distorting temporal fine structure by phase shifting and its effects on speech intelligibility and neural phase locking. *Sci. Rep.* 7:13387. doi: 10.1038/s41598-017-12975-3